How WattTime Gauges and Iterates on MOER Algorithm Quality

By Joel Cofield, Sam Koebrich, and Gavin McCormick

Background

WattTime is a not-for-profit tech company founded to advocate for, advance research in, and provide technical assistance to the field we refer to as "Automated Emissions Reduction", or AER. The objective of AER is to reduce emissions by deliberately shifting electricity use to times of lower marginal operating emissions rates (MOERs).

To this end, WattTime has for many years maintained a real-time API that provides sub-hourly MOER data (including forecasts) for over 25 countries, including every location in the continental United States, 24/7 year-round.

In order to maximize the efficacy of AER in reducing real-world emissions, a large fraction of our team has been conducting research continuously since 2014 in developing and advancing the most accurate, granular, and otherwise effective MOER algorithms possible.

Recently, H.R. 3684, the "Infrastructure Investment and Jobs Act," directed the U.S. Energy Information Administration to add to its existing public real-time U.S. electricity dashboards the exact same information (albeit at lower temporarily granularity)—the estimated marginal greenhouse gas emissions per megawatt hour of electricity generated for different U.S. balancing authorities and nodes.

Leading environmental not-for-profit think tank Resources for the Future is assembling recommendations to EIA about how they might go about sourcing such data, including gathering information on the accuracy, granularity, and other aspects of different MOER algorithms.

Of course WattTime would be delighted to share its MOER data with EIA. But even if EIA finds another solution, WattTime nonetheless has a vested interest in helping ensure that whatever MOER data are provided to the U.S. public are accurate and effective in reducing emissions.

Therefore, we are now sharing this never-before published information about how have learned to evaluate and compare the quality of different MOER algorithms; the best algorithm we have yet come across to date; and how even better algorithms could be developed and evaluated. Additional more detailed information is also being provided directly to RFF.

Over the years, our research team has found or developed numerous different algorithms to estimate MOERs. But a fundamental challenge of this field is how to evaluate which algorithms are most accurate. Marginal emissions measure the change in emissions per megawatt-hour *caused by* increasing or decreasing electrical load. But cause and effect is not something one can stick a physical sensor on. Even correlation does not guarantee causation. How then any time two MOER models disagree, to know which is more accurate?

Conventional wisdom suggests that there is no clear answer, and the best one can do is either check multiple models for consistency, or to extrapolate from data that actually were designed to measure an unrelated concept (for example, average emissions rates, or electricity prices). Note that of course neither of these approaches actually does guarantee accuracy.

However, it is not in fact true that the causal effect of load on emissions is somehow unknowable. An extensive scientific literature on causal inference provides clear guidance on how to approach such a problem.[1]

In the terms of this literature, each MOER value produced by an algorithm is an estimate of an "average treatment effect estimate": the amount of emissions per megawatt-hour that would be caused, on average, by increasing load under whatever conditions that data point applies to.

The causal inference literature refers to the difference between this estimated effect and the true effect (i.e., the average amount of emissions shifting this load would actually cause) as the model's bias. Thus, while an unbiased model can still have some drawbacks (see below), for most purposes a MOER model's bias is essentially the definition of its (in)accuracy.

Typically, it is not possible to precisely quantify a model's bias. However, after years of careful experimentation, WattTime has concluded that it is very possible to learn enough about two models' bias to compare their relative accuracy in a slightly less quantitative manner. Not with any one single measurement, but rather the same way one would verify the accuracy of any scientific model: by developing specific testable hypotheses, running multiple empirical experiments to test these hypotheses different ways, and updating hypotheses based on how well different models reflect actual real-world marginal emissions behavior.

[1] See, e.g. See. e.g. Rubin, Donald (1978). "Bayesian Inference for Causal Effects: The Role of Randomization". *The Annals of Statistics*. **6** (1): 34–58. doi:10.1214/aos/1176344064. Or Princeton University Press, 2008 or Card, David. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review*, vol. 43, no. 2, 1990, pp. 245–57, https://doi.org/10.2307/2523702. Accessed 18 May 2022.

The best such experiment to gauge model bias would be to compare the models to the results of the nearly universally acknowledged "gold standard" in statistics: large-scale randomized controlled trials or RCTs.[2] If a MOER model consistently matched the results of multiple large-scale RCTs, it would essentially be proven unbiased/accurate.

The problem, of course, is that conducting large-scale RCTs for MOERs would take quite a large number of devices intentionally randomizing their electricity draw simultaneously in a measured way. So, a second option is natural experiments that produce as-good-as-random, or "quasirandom", variation in load. When properly designed, sufficiently large quasirandom experiments also approach zero bias.[3]

The goal of quasirandom experiments is to find a situation where the independent variable whose effect one is trying to measure (in this case load) varied, but not in a way that could be plausibly driven by or correlated with any confounding variables. A classic example: when a goal is scored in the World Cup when the UK is playing, many people simultaneously get up and go make a cup of tea. The exact timing of the resulting spike in electricity is clearly based on the randomness of the game, not on power grid dynamics plausibly related to MOERs, and so is a quasirandom source of variation in load.

As with RCTs, the fundamental problem with natural experiments in MOERs is finding enough of them. Finding and exploiting such experiments–particularly quasirandom experiments with significant statistical power–can take luck, creativity, access to unusual datasets, and the capability to process very many data sets.

Over the years, WattTime has gone through many natural experiments that can be used to validate MOER models. Examples include: covid waves; ISO load forecast error (after de-biasing for ISOs that bias their forecasts); triple differences with different load forecast errors in neighboring BAs; the start and end of official ozone season in a neighboring state; and wind speed forecast error. We are always looking for more. While some of these experiments have limited statistical power, they are very robust to bias and over the years have been extremely informative in validating (or rejecting) various MOER models and modeling assumptions.

[2] See, e.g. Sackett D.L. Rosenberg W.M. Gray J.A. Haynes R.B. and Richardson W.S . (1996) Evidence based medicine: what it is and what it isn't . *BMJ (Clinical Research ed.)* , 312 , 71 – 72 .
[3] See. e.g. Angrist, J. D., and Jorn-Steffen Pischke. *Mostly Harmless Econometrics.* 2008.

Gauging MOER algorithm efficacy

In addition, a key WattTime learning in evaluating MOER models is that bias is not the only metric that heavily influences their effectiveness in driving actual emissions reductions.

Some other features, such as the ability to compute MOERs in time for them to actually influence decision making (i.e. in real time or forecast) are obvious. But often more subtle considerations can in practice make the difference between a MOER model that is highly effective in reducing emissions, and one which in actual use drives little real-world emissions reductions.

Other key factors in real-world efficacy we have found are:

- The *variability* of a model. Unlike in most causal inference settings, MOER models drive results not by merely by accurately measuring one average treatment effect, but by distinguishing between different treatment effects at different times. Thus, the more precisely a MOER model can distinguish between situations of different average treatment effects–in short the more variable and less flat it is–the more emissions it can reduce.
- The correct *rank ordering of the extreme lowest values*. In actual operation, what nearly every AER user in any field does is deliberately shift load to the lowest MOER values a model gives within some set of available times (e.g. the cleanest times each night while an electric vehicle is plugged in). So even a flawed MOER model that correctly identifies when the genuinely lowest MOER values (typically periods of 0 emissions when renewables are marginal) are happening will drive substantial emissions reductions. By contrast, a seemingly minor error in a MOER model whereby it occasionally produces erroneous extremely low values can lead to a dramatic reduction in its emissions reductions efficacy–or even cause it to accidentally increase emissions.
- The *applicability* of a model. The real-world effect on emissions of shifting load can depend on many factors, including factors about how the load is shifted. One example is the size of load shift. Because any given power plant can only be marginal for at most an amount of load equal to its own capacity, MOERs are not the same for different size shifts in load. So to successfully measure and reduce emissions, a MOER model must not only be accurate, it must be applicable for the actual type of load shift under consideration.

Of note, WattTime has also found that one seemingly important feature of marginal emissions rates is in practice *not* as relevant for MOER emissions reduction efficacy as we initially

assumed: nodal-level geographic granularity. This is because while it is common to deliberately site new renewables in optimal locations using *build* margin rates, there are extremely few use cases that involve dynamically shifting load in place on an hour-by-hour basis.

Geographic granularity affects emissions reduction efficacy if and only if two nearby sites have different rank ordering of MOER values. So, WattTime's own comparative geographic approach is therefore to not always provide a signal with maximum granularity (since that can reduce statistical power and variability), but instead to increase granularity only when doing so actually increases variability. Specifically, each time a balancing authority is not congested and behavior is similar across all nodes, we only calculate one MOER for the entire balancing authority. When there is congestion, we provide different answers for different groups of pricing nodes that exhibit similar behavior. We consider the implications of this later in a later section.

Models considered

WattTime routinely conducts extensive theoretical and empirical research to explore or develop new models and evaluate them according to the above criteria. Below are five of the most common classes of models that WattTime has explored before adopting our current approach. As of today (May 2022), we have concluded that the most accurate and effective model currently available is essentially a hybrid of several of the below approaches.

1. *Difference-based models*

One intuitive estimate of MOERs is simply the observed change in emissions between two points in time, divided by the observed change in load between the same times. This simple and highly granular approach is still used by some grid modelers today. However, it has several drawbacks. Most notably, the model is highly subject to bias because many confounding variables in power grids change over time (e.g. solar output covaries with load). Thus, estimates can be inaccurate. Also, when load change between two periods is relatively small, even slight measurement errors or confounding variables can drive the model to extreme negative values, often making the crucial rank ordering of the lowest emissions times particularly inaccurate.

WattTime stopped using difference-based models in 2014.

2. *Conventional binning-based regression models*

Binning-based regression models represent a significant improvement from pure differencing models. These models involve regressing changes in emissions against changes in load, usually using hourly U.S. EPA CEMS data. Different regressions are run–wholly independently of each

other–within different "bins" of historical hours. The bins group together hours with similar conditions, whether those be time of day, season, grid and weather conditions, or type or size of load shift.

These models have quickly become the most widespread technique among marginal emissions experts in academia.[4] This is partly due to a slightly non-obvious effect of binning, which is that it mitigates many potential sources of model bias. (Unlike using continuous covariates, which are much more likely to introduce confounding variables into a regression.) Thus, while not perfect, these models are typically fairly accurate.

They also can be tuned to be applicable to specific types of load shifts. For example, the EPA AVERT model effectively "bins" on the size of a load shift, and thereby explicitly offers different MOER models that are applicable to different size load shifts, boosting their emissions reduction potential. WattTime has used similar techniques to further increase applicability, such as using different MOER models for sudden, unexpected load shifts (which more often affect plants that bid in the real-time market) versus similarly-sized load shifts that were known in advance to the grid operator (which more affects the day-ahead market).

However, some disadvantages of such models are that they have low variability (because they can only distinguish so many different grid conditions without losing statistical power) and often low spatial granularity (depending on how multiple plants are aggregated together).

Further, most such models exhibit high bias in measuring marginal renewable energy. In some cases, this is because the regressions are run entirely on hourly EPA CEMS data, which contains no data on renewable energy. In other cases, it is because the regressions are run entirely on generation by fuel mix. And renewable energy potential is highly correlated with load, generated a confounding variable that *does* apply even within a single regression or bin. As renewable energy grows more common, this is increasingly affecting the estimated rank ordering of the lowest-MOER times, and beginning to seriously reduce the efficacy of this class of models.

In 2017, due in part to rising U.S. renewable energy penetration, WattTime stopped using conventional binning-based regression models anywhere in the country.

---

[4] See, e.g. See, e.g. Callaway, D., Fowlie, M, and 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. Val 5, No. 1; Siler-Evans, K., Azevedo, I. L., and Morgan, M. G. (2012). Marginal emissions factors for the US electricity system. Environmental science & technology, 46(9):4742{4748; or Kaffine, D. T., McBee, B. J., Lieskovsky, J., et al. (2013). Emissions savings from wind power generation in Texas. Energy Journal, 34(1):155{175

3. *Heat rate models*

Another family of models uses LMP and fuel prices to back out the implied heat rate of the marginal generator. The advantages of a heat rate model are that they're simple to calculate at a nodal level, give high temporal granularity, and are well suited for empirical validation. In CAISO and to a lesser extent a few other grids, the model matches observed power plant behavior reasonably well.

However, a key disadvantage of these models is that they do not generalize well to most U.S. grids. This is particularly the case for grids in which coal and gas are interspersed in the bid stack. In such grids, WattTime experimentation found that even the slightest deviation from theoretically optimal dispatch (which happens often) caused highly unstable behavior in heat rate models, which then did not match actual observed power plant behavior well.

WattTime explored adopting heat rate models for the entire U.S. in 2017, but stopped when it became clear that the models only performed well in a limited few balancing authorities.

4. *Experiment-based models*

Another class of models uses only RCTs and quasirandom natural experiments. While these models have excellent accuracy and nearly zero bias, the disadvantage of course is their very low statistical power. Except in extremely rare circumstances, there are not enough experiments to reliably detect, for a given grid, slight variations in the marginal emissions rates under every possible grid condition.

WattTime developed some fully quasirandom experiment-based models in 2018, but stopped using them when the precision was too low.

5. *Grid operator dispatch models*

Another family of models works by estimating the MOER as equal to the annual average emissions rate of the current marginal unit. This highly intuitive approach has very substantial advantages, most notably that grid operators are typically privy to much more detailed data about their own grids than is available to the general public. This means that these models can often be quite accurate, with very low bias. (Though they can be sensitive to modeling assumptions, particularly about the behavior of grids other than their own.) These models also are highly variable and highly granular.

Their biggest drawback is their applicability. Most grid operators determine the marginal unit using some form of constrained optimization model, such as mixed integer linear programs.[5] These models typically calculate the causal effect only of an arbitrarily small shift in load. They were never intended to be linearly extrapolated to apply to much larger shifts in load. While such extrapolation can often provide a reasonable estimate when estimating costs, the same is frequently not true for emissions.

For example even today, with AER still in its infancy, WattTime already routinely runs into the situation where it is dispatching hundred of megawatts simultaneously in a single balancing authority, while the local grid operator is curtailing tens of megawatts of renewable energy. While it is true that the emissions intensity of the marginal unit is 0, it is clearly false to say that shifting ten times that much load to this hour is causing 0 emissions. As WattTime and other groups continue to experience rapid exponential growth in the amount of load responding to MOERs, this means a MOER signal that only measures arbitrarily small changes will rapidly become even less applicable, greatly reducing the practical utility of such models.

In conversations with grid operators, WattTime has also heard practical challenges such as a belief that the models' most extreme low values–the precise values which AER users often shift their load around–are very difficult to forecast and are in fact the most likely to only apply to very small changes in load.

WattTime explored the use of grid operator dispatch models in 2018 and again in 2022, but stopped using them when the linear scaling became a clear concern.

Developing hybrid models

Each of the above five approaches has its own advantages for use in AER, but also carries drawbacks. Thus, over the years, WattTime has increasingly focused on developing hybrid models that combine the strengths of multiple approaches.

Specifically, our objective is models that combine the low bias and high applicability of econometric models (or even better, experiment-based models), but with the greater variability and temporal granularity of models like difference and heat rate models.

One key insight was in moving away from fully experiment-based models (with their low statistical power) to instead using experiments to inform the creation, validation, and adoption of more detailed assumption-driven ("structural") models that are capable of greater variability.

[5] E.g. this class of model: https://gocompetition.energy.gov/

Over the years WattTime tried many MOER models (generated both in-house and externally) that generated results, but which did not hold up to such empirical evidence. For example, in 2018 we attempted to combine heat rate models and grid operator dispatch models by pulling the posted marginal fuel and power plant attributes from ISO websites, extrapolating heat rates for that fuel at that time from the ISO's posted LMP, and inferring a marginal emissions rate. The method successfully generated a detailed, nodal-level hourly MOER dataset. However, when we then sought to validate that model by examining the actual behavior of the power plant in question, we found evidence of very high bias. Whether due to transmission constraints, imports, cold start costs, or some other confounding variable, the model was not accurate in practice.

Thus, even though WattTime is capable of producing nodal-level geographic granularity, as of today, we do not do so. We are so cautious on geographic granularity because while it is trivial to generate numbers for nodal-level data, it is to date very unclear how to even validate whether any such numbers yet generated are actually accurate and applicable, rather than merely statistical artifacts. (Though we anticipate it will be possible in future; see below.) While the Infrastructure Bill is very clear in specifying nodal-level MOER data "where available", it may be worth considering whether the EIA may want to consider at least some basic level of accuracy or efficacy nodal-level datasets must meet in order for them to be considered useable.

The current WattTime model

WattTime's current main MOER model is a hybrid, data-driven model that leverages both historical hourly CEMS power plant-specific generation and emissions and real time APIs[6] for grid conditions, interchange, and weather.

For each balancing authority in the United States, WattTime applies a multistage model to generate MOERs: we separately estimate (1) marginal non-renewable sources (both fossil fuels and imports); (2) marginal curtailment of renewables; and (3) carbon intensities for each marginal fuel source. For (1), we use regression approaches binned on grid condition data, mixed with machine learning techniques such as deseasonalizing and training/test splits data to learn much higher variability patterns in historical data. Where an exclusively data science approach would seek to maximize predictive performance, we instead leverage domain knowledge to preselect compelling grid variables to bin on, and within bins run linear regressions against load. Renewables are excluded because correlation between load and renewables generation is frequently non-causal.

[6] Such as this API https://www.eia.gov/opendata/

Instead, a (2) separate model uses historical curtailment data to learn under what grid conditions circumstances renewable energy is curtailed (and thus is responsive to marginal demand). These models are tailored to the ISO, usually via supervised learning models fit to nodal congestion data. Then, (3) WattTime uses a different variation on bin-based regression models, trained on CEMS data and binned on grid conditions data, that calculates a carbon intensity associated with each fuel type on the margin at each five minute time step. Several adjustments are made for complex edge cases, most notably reservoir hydro (where the marginal emissions caused by running at one time depends on the MOER at the time when the plant would otherwise have discharged its energy).

Separately, WattTime also forecasts MOERs using a purely supervised learning model fit to the historical patterns of our MOER model.

Future model upgrades

As noted above, WattTime is continually exploring potential upgrades to our models. We are currently planning the following particularly notable upgrades:

First, on applicability. Difference models and conventional binning-based regression models are most applicable to a total (not per-user) load shift size that is equal to the average variation in their training data set. By default, this is typically the average variation in load size per hourly time step in a region. While this varies by balancing authority, for a sense of scale this is often around 500 MW in a large ISO. Today in 2022, WattTime uses the default because this is a reasonably close match to the total amount of load we estimate actually is using AER in a given balancing authority today. But as load under AER management grows, WattTime anticipates needing to adjust its models to intentionally optimize for different size load shifts.

Second, WattTime will continue over time to examine more natural experiments to further our checks on accuracy. Examples involve looking at both regression discontinuity of planned and unplanned power plant closures, difference in difference by comparing the spread of covid waves, and so on.

Third, supplementing natural experiments with intentional randomized controlled trials. As the gold standard in causal inference, RCTs can further support refining the quality of MOER models, and further support public third-party validation. The issue, of course, is that it requires a very large number of devices engaging in AER in order to generate a large enough signal to be readily detectable in emissions data. As more companies adopt AER, we are in discussions with a number of them to integrate an RCT component into their use of AER in future. We are on

track for enough signups to begin conducting such RCTs some time in 2023, though it may be longer to find the staff time required to begin operating any ongoing RCTs.

Lastly, the ideal MOER model would also integrate the vast detailed non-public data that typically only grid operators can access. Many, though not all, U.S. grid operators have an extremely advanced understanding of the behavior of their grids. If these operators were to release a new hybrid type of models that can predict MOERs relevant to actual quantities of load being shifted through AER (by running dispatch scenarios and comparing them), those models would potentially be extremely high quality. WattTime has spoken with multiple grid operators and is optimistic that given enough time, some ultimately will develop or co-develop such models.

Therefore, WattTime suggests that RFF recommend to EIA not merely the adoption of any one MOER algorithm, but also specifically mandate that the provider of any MOER data should periodically update their approach over time as new data become available.

The ideal model, of course, would be a hybrid model that was able to combine the strengths of all of the above approaches. This would however require extensive data sharing between multiple parties (e.g. companies that run AER who can operate RCTs, grid operators who have advanced data, and researchers and data providers like WattTime). Therefore, we also suggest that RFF recommend that to the maximum extent reasonable, the provider of any MOER algorithm to EIA should also provide an open data set of detailed grid and emissions data that can be used for model training and validation purposes.

WattTime has experience with this approach as a key convener of, and secretariat for, the Climate TRACE coalition. This collaboration involves coordinating the sharing of detailed open emissions data to facilitate emissions research covering over 100 countries between 12 different data science teams at universities, tech companies, and nonprofits.

Put together these advancements represent a considerable investment of private sector resources to research and improve MOER quality. Again, WattTime would be delighted to provide its marginal emissions data to the general public via EIA's portal. But even if not selected for this purpose, it is strongly in line with WattTime's core mission to support EIA with the best available research on how to evaluate and improve on MOER data quality. We would be interested in direct engagement, support on model validation, and sharing future open data resources—whatever is needed to help ensure the U.S. public gets access to accurate MOER data that is effective in actually reducing emissions through AER.