Marginal Emissions Modeling:

WattTime’s approach to modeling and validation

EXECUTIVE SUMMARY

The energy industry has long been aware of the potential for changing the timing of electricity use to reduce costs. But what effect do such changes have on emissions?

WattTime is a nonprofit founded by former UC Berkeley researchers to continuously gather and raise awareness of the latest research on this question, help advance that research, and provide technical assistance to anyone seeking to deliberately time load to reduce emissions. We refer to this practice as **Automated Emissions Reduction (AER)**.

AER works by first quantifying the amount of environmental harm–most often measured in greenhouse gas emissions per megawatt-hour–that would be caused by using a given amount of additional electricity at different times and places. This is referred to as the **marginal operating emissions rate (MOER)** for that time and place.[1] Then, the electric load can be deliberately shifted to times of lower MOERs, to cause less environmental harm.

For the last decade, the WattTime team has been continuously gathering, evaluating, and improving the accuracy and efficacy of MOER models for electricity grids around the world.

Several factors influence the efficacy of a MOER model in successfully reducing emissions. For example, a model that cannot produce real-time data is not particularly helpful in informing real-time decisions of when to use load. But the most important consideration is of course the accuracy of the model used to measure or estimate MOERs.

An accurate MOER correctly measures the difference in emissions per megawatt-hour between the case where load is changed, versus a hypothetical counterfactual in which it *wasn’t* changed. But this raises a fundamental challenge: one can’t directly measure a counterfactual. How then to know which is more accurate when two MOER estimation methods disagree?

It can be tempting to throw up one’s hands and assume that there is no clear answer, and the best one can do is either check multiple models for consistency, or use data designed to measure an

[1] Note this can sometimes be different than the long-term effects of permanently shifting load. The Greenhouse Gas Protocol calls these long-term effects the Build Margin. As of 2022, there has been much less research on Build Margins so uncertainty is far higher. However, the US National Renewable Energy Laboratory produces some estimates.

unrelated concept as a proxy (for example, marginal prices or total emissions per megawatt-hour). Note that of course neither of these approaches actually does ensure accuracy.

But it is not in fact true that without a directly measured set of ground-truth data, the causal effect of load on emissions is unknowable. The extensive scientific literature on causal inference provides a clear consensus on how to approach this problem.[2]

The universally acknowledged "gold standard" in causal inference is to conduct randomized controlled trials (RCT). Just as with medical trials, all the assumptions and theory in the world are no substitute for actually taking an action in the real world and directly measuring its effect.

WattTime is working with a number of organizations that control large amounts of load to do just that–conduct controlled experiments randomizing the timing of their load and studying its effects.

The problem, of course, is that it takes randomizing a lot of load and repeating the experiments many times to produce reliable, robust results. In the meantime, over the years WattTime has used many other rigorous data-driven techniques designed for causal inference, principally examining natural experiments.

This document:

1. Lays out the criteria that influence the efficacy of a MOER model in reducing emissions (accuracy/unbiasedness, availability at the time a decision must be made, variability, correct rank ordering of lowest-MOER time, and applicability);
2. Reviews five common classes of MOER models and the research record of these approaches in meeting these criteria for successful emissions reduction;
3. Lays out the most successful approach WattTime has yet found (a hybrid model that combines several of the common approaches); and
4. Describes WattTime's plans for how we think further experimentation can even more definitively compare and improve MOER models.

WattTime is solely focused on whatever it takes to ensure that anyone seeking to reduce emissions through AER has the most accurate and impactful data possible. As we continuously explore, evaluate, and iterate on new approaches, we welcome new techniques, collaborations, and better ideas. We look forward to sharing notes with anyone else who seeks the same.

[2] See, e.g. the foundational causal inference paper: 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.

Background

WattTime is a nonprofit tech startup founded to raise awareness of, advance research in, and provide technical assistance for the practice of shifting electricity use to times that cause less environmental harm. We call this practice "Automated Emissions Reduction," or AER.

AER works by first quantifying the amount of environmental harm–most often measured in greenhouse gas emissions–that would be caused by using a given amount of additional electricity at different times and places. This is referred to as the marginal operating emissions rate (MOER) for that time and place. Then, the electric load can be deliberately shifted to times of lower MOERs, to cause less environmental harm.

To this end, WattTime has for many years maintained a real-time API that provides sub-hourly estimates of MOERs (including forecasts), 24/7 year-round, for all local electricity grids in the continental United States and currently 27 other countries (as of mid-2022).

Naturally, the better one's MOER estimates, the more effective shifting load can be at successfully reducing real-world environmental harm. To support this, a large fraction of our team has been conducting research continuously since 2014 in developing and advancing the most accurate and effective MOER algorithms possible.

Evaluating Marginal Emissions Data

A MOER estimate is a measure of the change in emissions (in pounds of CO2 or other pollutants) that is *caused by* a change in electrical load (per megawatt-hour). Over the years, our team has researched and developed many different algorithms to estimate MOERs. A central challenge in this field is evaluating which algorithms are most accurate and effective.

The fundamental challenge behind evaluating MOER accuracy

Marginal emissions models are a classic causal inference problem: they seek to compute the causal effect of taking an action (e.g., increasing electric load).

In theory, one could obtain a perfect measurement of a causal effect by comparing the results of taking the action, to the results of the counterfactual world in which one didn't take it. The problem, of course, is that one can't both take an action and not take it at the same time. And since MOERs are constantly changing based on fluctuating grid conditions, one can't simply take the action at one time and compare it to the results of not taking it at a different time.

In statistics this is known as the Fundamental Problem of Causal Inference. Its consequence is that for causal effects that change under different conditions, there isn't (and can't ever be) a single source of measurable and perfectly accurate ground truth. And without a source of measured ground truth, one can't quantify the accuracy of MOER algorithms by simply measuring their correlations with the ground truth data.

How, then, can the accuracy of a MOER model be evaluated? Conventional wisdom has suggested that there is no clear answer, and the best one can do is either check correlation with a different model, or abandon the concept of marginal emissions entirely and instead use a proxy. Note that neither of these approaches necessarily improves accuracy.

How WattTime gauges MOER algorithm accuracy

This lack of a simple universal dataset does not in fact mean that the causal effect of load on emissions is somehow unknowable. The extensive scientific literature on causal inference provides clear guidance on how to approach such a problem.[3]

Because the central problem is that MOERs fluctuate based on fluctuating grid conditions, one thing that can be done is to compare different times that have almost identical grid conditions. In the language of the literature, a MOER estimate of multiple times with similar grid conditions is an estimate of an "average treatment effect." This measures the quantity of emissions per megawatt-hour that would be caused, on average, by increasing load under these particular conditions.

A model's bias is the difference between this *estimated* effect and the *true* average of the effect under those same conditions. 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.

The ideal way to gauge MOER model bias would be to compare the modeled results to the universally acknowledged "gold standard" for accurately measuring causal effects: large-scale randomized controlled trials (RCTs).[4] If a MOER model's estimates of the average effect consistently matched the true measured average emissions changes for those same grid conditions during multiple large-scale RCTs, it would essentially be proven unbiased. Each event

[3] 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.

[4] 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 .

in the RCT experiment would consist of a randomly timed call to a large number of devices to synchronously increase or decrease load, which will produce a large increase or decrease in the aggregate load, and corresponding measurable change in emissions.

WattTime has many partners well-positioned to run such RCT experiments using devices in their control and has begun collaborating on the design of upcoming RCTs. But conducting this type of trial requires access to a large number of devices randomized a number of times under similar grid conditions to precisely measure the true average effect of variable electricity load on emissions. It will take time before results from RCTs are available.

In the meantime, a second option is leveraging natural experiments that produce as-good-as-random, or "quasi-random," variation in load or demand. The goal of quasi-random experiments is to find a situation where the variation of an independent variable (in this case: electric load) produces an effect that one is trying to measure, but not in a way that could be plausibly driven by or correlated with changes in grid conditions (known as confounding variables).

A classic example is halftime breaks during UK soccer matches. A large number of viewers simultaneously use this time to boil water for tea. The precise timing of a large spike in electricity is therefore based on the game clock, not underlying power grid conditions such as electricity prices. The halftime break occurs independent of any variables that would otherwise confound the effect of demand on energy production, making it quasi-random for the purposes of measuring electricity emissions.

When properly designed, sufficiently large quasi-random experiments also, like RCTs, approach zero bias[5] and have been a key tool for validating MOER models. The fundamental problem with properly quasi-random natural experiments and RCTs for use in MOER research is just finding enough of them.

WattTime's basic strategy for validating MOER algorithm accuracy is to examine the predictions of these models–particularly their average treatment effects–and compare them with as many experiments as possible–particularly RCTs and quasi-random natural experiments of high quality.

To date, WattTime has already identified or analyzed many experiments. This has included quasi-random natural experiments, e.g. differences in changes in weather conditions in neighboring ISOs; the start and end of regulatory ozone season; errors in ISO load forecasts; etc.

[5] See. e.g. Angrist, J. D., and Jorn-Steffen Pischke. *Mostly Harmless Econometrics*. 2008.

But it has also included other simple common-sense tests, such as whether the power plants a given model estimated to be marginal even *ever* changed their behavior; whether the predictions of a given model were even possible under the laws of physics; and so on.

And we are always looking for more experiments. While any individual experiment can have limited statistical power, as we continue to gather more and more, the picture of marginal emissions is becoming increasingly clear and can be extremely informative in validating (or rejecting) various MOER models and modeling assumptions.

Gauging MOER algorithm efficacy

One of WattTime's key learnings in evaluating MOER models has been that when models are applied in real-world operation, bias/accuracy is not the only metric that heavily influences their effectiveness in reducing emissions.

Some other important considerations are obvious, such as *availability* in real-time to actually influence operational decision-making. Often more subtle considerations can make the difference between a MOER model that is highly effective and one that drives little real-world emissions reductions. The key factors in real-world efficacy we've observed are:

- **Real-time availability.** Given that the fundamental mechanism behind AER is to decide to use energy at a particular time based on how current MOERs compare to a forecast of future MOERs within the window in which load could be shifted, one needs to have both estimates in time to make each decision.
- **Rank ordering of the extreme lowest values.** A MOER model that correctly identifies minimum MOER values, for example times of renewable energy curtailment, will drive substantial emissions reductions. By contrast, even a MOER model that is fairly accurate overall can still fail to drive much real-world impact–or even accidentally *increase* emissions–if it fails to correctly capture which times are the lowest values. This is because in actual operation, nearly every AER user deliberately shifts load to the lowest MOER value within some set of constraints and ignores all other values. (E.g., they may shift load to the cleanest times during which an electric vehicle is plugged in each night.)
- **Model applicability**. More subtly, the effect on emissions from load-shifting can also depend on factors decided by the user, not just grid conditions. For example, the true marginal effect of shifting very large loads may not be the same as that of shifting very small loads. Whether the shift is known in advance to the grid operator, or how fast the load ramps, can also affect results. To best reduce emissions, a MOER model must be applicable for the actual type of load shift the user is considering.

- **Granularity (particularly temporal).** Unlike most causal inference settings which focus on a single average treatment effect, MOER models drive results by distinguishing between different treatment effects at different times and places. However, individual treatment effects can be slightly above or below the average for that type of condition. For example, suppose a model finds the average treatment effect of every day at 2-3pm. It can still be the case that within that hour, 2-2:30pm is slightly above that average and 2:30-3pm is slightly below it. The more a MOER model can not just capture average treatment effects, but precisely capture fine-grained variations within those averages, the greater its impact. This is particularly true for temporal granularity, since most AER shifts load between different times, not different places.[6]

A small caveat is in order on granularity. Adding granularity to a MOER model effectively introduces an additional choice: what is the relative ranking of times and places with a given larger area or time period?

This introduces an additional opportunity for bias to interfere with emissions reductions. For this reason, it is very easy to design a model that is more granular, but less accurate. Thus, additional granularity increases emissions reduction efficacy if–but only if–it does not increase bias.

Review of Marginal Emissions Model Types

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 main classes of models that WattTime has explored. As of today (October 2022), we have concluded that the best model yet available anywhere is a hybrid of several of the below approaches.

1. *Difference-based models*

One basic MOER model is to simply divide the change in emissions between two consecutive time steps by the corresponding change in load across those times.

Though highly granular, a simple difference model is highly subject to bias, and thus inaccuracy, since changing grid conditions frequently confound the effect of varying load. For example, load often increases at the same time as solar output does, as seen in California's well-known spring day "duck curve". But the increased load did not *cause* the sun to rise. Worse, in

[6] This is for AER. By contrast, in renewable energy siting or what WattTime calls "emissionality", most generation is optimized between different places, not different times. So in that use case, temporal granularity can help but geographic granularity is more important.

difference-based model this bias is magnified any time load is changing slowly, resulting in big errors in emissions rates any time grid demand is fairly flat. (This is because when calculating change-in-emissions per change-in-load, if the denominator is small it makes a large bias multiplier). Thus, such models can surprisingly often lead to extremely high or low values based on pure coincidence, particularly when load is changing slowly. Such noise often seriously degrades the rank ordering of the lowest emissions times.

WattTime completely ceased using difference-based models in 2014.

2. *Conventional binning-based regression models*

Binning-based regression models instead fit with a linear model between emissions and load, giving a slope, or marginal emissions relationship. Using multiple "bins" of historical data for each hour, independent regression analyses produce different MOERS for different grid conditions.

These models represent a substantial improvement in accuracy from pure differencing models in two ways. First, by using regression, they are not subject to the problem of magnified bias during times of small load changes. Second, by cross-comparing times when grid conditions are similar, they better isolate the causal effect of only the change that is being evaluated–a change in load–and are much less subject to bias from confounding variables.

While not completely perfect, evidence suggests that these models are typically fairly accurate. These models have quickly become the most widespread technique among marginal emissions experts in academia.[7] Very similar models have also been used outside of academia, e.g., by ISONE and the EPA.

The primary mechanism by which a binning-based regression model can still be inaccurate is if grid conditions within a bin are not similar enough along a crucial confounding dimension. (This is known as covariate imbalance.) WattTime has experimented with addressing this by comparing whether multiple models binned on different grid conditions produce consistent results, and found this can be quite effective in detecting confounding variables.

[7] 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

Probably the most common such confounding bias is in measuring marginal renewable energy. This can be both because most binning-based regression models today do not include renewable generation in the dataset (since it is not available in EPA's CAMPD database).[8] It is also because renewable energy can be highly correlated with load even within a single regression or bin. As renewable energy grows more common, this bias in detecting the low MOER times is beginning to seriously reduce the efficacy of this class of models.

Separate from accuracy, these models also have a substantial advantage in applicability. By adjusting the bins and the independent variable in the regression, they can be easily tuned to be applicable to particular types of load shifts. For example, the EPA AVERT model effectively "bins" on the size of a load shift, yielding different MOER estimates for different size load shifts. Similarly, WattTime has generated different MOERs for different types of load shifts, e.g., those known in advance to the grid operator versus those that were abrupt changes.

However, these models have other disadvantages in emissions-reducing efficacy. They tend to have low variability (because they can distinguish only so many different grid conditions without losing statistical power) and often have low spatial granularity (depending on how multiple plants are aggregated together).

In 2017, due in part to rising U.S. renewable energy penetration, WattTime changed its approach and began deviating from exclusively binning-based regression models.

3. *Heat rate models*

Another family of models uses Locational Marginal Pricing (LMP) and fuel prices to back out what heat rate the LMP should correspond to for any given fuel type, and use this to directly calculate a resulting marginal emissions rate. Key advantages of heat rate models are that they can be highly variable and highly granular.

In some grids, like CAISO for example, this approach can be fairly effective because there is almost always only one emitting fuel type that could be marginal (typically gas). However, these models don't perform as well in any region where two different emitting fuel types–typically gas and coal–are intermixed in the bid stack. In 2012, this began to occur frequently in many US grid regions, making heat rate models less accurate than they had been previously. This inaccuracy is

[8] Regression-based models, and to some extent all empirical- rather than assumption-driven model types, are made possible by comprehensive systems for direct measurement of emissions like the US EPA's Continuous Emissions Monitoring System (CEMS). The US EPA requires all fossil-fueled power plants larger than 25 MW to measure and report their emissions of carbon dioxide and other pollutants at an hourly frequency through CEMS and the data is made publicly available via the EPA's CAMPD (updated quarterly).

only aggravated any time that power plant dispatch deviates even slightly from what short-term heat rates would suggest, as studies have found often occurs with both coal-fired[9] and gas-fired[10] plants.

WattTime uses a heat rate model in CAISO to provide the MOER data for the California Self-Generation Incentive Program (SGIP). In 2017, WattTime explored a project to expand the same approach to every other grid region in the US, but found they had substantially lower success in predicting real-world power plant behavior in many regions, and therefore stopped the effort.

4. *Experiment-based models*

Another class of models uses strictly experiments (whether RCTs or quasi-random natural experiments) to generate MOERs. The strength of these models is that given enough experiments they have extremely low or even zero bias, making them best-in-class for accuracy.

The challenge is that they have low statistical power. This leads to models that provide a good unbiased estimate of the average treatment effect, but have very low variability and granularity because they have trouble producing a range of treatment effects under various grid conditions.

To be not just accurate but also effective, such models require very large amounts of high quality data. This leaves MOER model makers using fully experiment-based models with two choices: either (1) only use them in specific times and places where very large amounts of data happen to already have been gathered and nowhere else; or (2) deliberately gather large volumes of high quality data. In the long term, WattTime hopes and expects to be able to use the second strategy.

WattTime developed some fully quasi-random experiment-based models in 2018 and had some initial success in Sweden where we had access to an unusually rich non-public dataset very suitable for quasi-random natural experiments.[11] But we were unable to access similarly rich datasets for every grid, and found that without them the statistical precision was very low, leading to very low temporal variability.

[9] See Fisher, Jeremy, Al Armendariz, Matthew Miller, Brendan Pierpont, Casey Roberts, Josh Smith, Greg Wannier. *Playing With Other People's Money: How Non-Economic Coal Operations Distort Energy Markets* Sierra Club, 2019.

[10] See Massie, Ashtin, Joe Daniel. *The Shaky Economics of Gas-Fired Power* Union of Concerned Scientists 2022.

[11] Both because the Nordpool grid operator offers unusually detailed public grid data on key sources of quasi-random variation, and because a local utility shared additional non-public data with WattTime.

5. *Emissions of the grid operator's marginal unit*

Another family of models works by estimating the MOER as equal to the annual average emissions rate of the current "marginal" unit in an economic sense, i.e., the unit the grid operator uses to set the marginal price. 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. Due to the shared use of the word "marginal," it is tempting to conclude that the "marginal emissions" of a shift of N megawatts must surely be equal to the emissions rate of the "marginal unit" times N. However, this is only true in a very special case where the effect of shifting a small or large amount of load is precisely the same.

Most grid operators set the marginal unit using a highly deterministic set of equations, such as a mixed integer linear program[12], whose output is specifically designed to apply to some fixed, typically very small, unit of load. Applying linear extrapolations (naively) on the marginal unit is a well-known faux-pas in Operations Research, known to give widely incorrect results.

Consider the example of X megawatts of wind curtailment occurring, during which time increasing load by up to X megawatts will cause no emissions, for a marginal emissions rate of zero. But clearly increasing load by 10X megawatts cannot cause the wind farm to produce at ten times its capacity. For load changes much larger than X, the marginal emissions rate is not equal to the emissions rate of the marginal generator.

Even today, with AER still in its relative infancy, WattTime already runs into situations where its partners are shifting more load than the size of the marginal unit. As more and more load is being shifted using MOER signals, this spread is only growing, greatly reducing the practical utility of this class of models.

WattTime explored the use of grid operator dispatch models in 2018 and again in 2022. But the models exhibited considerable extreme behavior, with MOER results that were sometimes negative and sometimes far higher than any unit on the relevant grid. In interviews with grid operators, ISO staff expressed the opinion that these extreme values were likely technically true, but were driven by special cases that applied only to extremely small amounts of load and should

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

not be linearly extrapolated to estimate the marginal emissions rate of shifting larger amounts of load. Thus, WattTime has stopped using models based on the emissions of the grid operator's marginal unit.

Of note, many grid operators *do* have the technical capabilities and sufficient data required to generate MOER estimates that cover larger amounts of load. This would involve modeling two different load scenarios, in which grid conditions are identical except that demand is different by a specific amount. Dividing the difference in the total grid emissions of these two scenarios (including emissions from imports) by the difference in load would yield a very accurate estimate of the marginal emissions for those grid conditions. WattTime has had some discussions with grid operators about performing such modeling and believes that in the long term this will produce one of the most accurate possible measurements of marginal emissions. However, as of October 2022, WattTime is not aware of any datasets sufficient to enable this alternative approach yet being made available external to grid operators.

Developing hybrid models

Each of the above five approaches has considerable advantages for use in AER, but also carries significant drawbacks. Today, WattTime focuses on developing hybrid models that combine the strengths of multiple approaches.

When designing our models, our objective is to combine the low bias and high applicability of econometric (experiment-based) models with the variability and temporal granularity of models like difference and heat rate models.

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

This approach has proven very fruitful. Over the years, WattTime has tried many MOER models (both generated by others or in-house) which clearly did not hold up to such empirical evidence. As just one example, in 2018 we attempted to combine heat rate models and grid operator dispatch models. We used the marginal fuel and power plant attributes from ISO websites, extrapolated heat rates for each fuel at that time from the ISO's posted LMP, and inferred 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 it predicted real-world power plant behavior very poorly. Whether due to transmission constraints, imports, cold start costs, or some other confounding variable, the model was not accurate in practice.

Our current primary model, which is the best combined approach we have found to date, has many strengths: evidence of low bias, very accurate rank ordering of crucial lowest-emissions times, good applicability, and high temporal granularity. Its biggest weakness is low geographical granularity. We are cautious to increase 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 are actually accurate and applicable rather than merely statistical artifacts.

For WattTime's focus on AER, geographic granularity is often not as important since most load shifting happens in time. However, increasing granularity always has some potential to squeeze out even more real-world savings. Thus, WattTime's next research priority is to explore further model combinations and means of validation to see if it is possible to produce a model that has all of the above strengths and also nodal-level geographic granularity.

The Current WattTime Model

Recapping our view of MOER quality, what we prioritize in a MOER model is:

- It is *accurate*—combining the low statistical bias of a conventional binning-based regression model, better handling of renewable energy and net imports, plus additional experimental validation;
- It is available *in time* to inform real-time operational decision-making, including a forecast when necessary;
- It is *variable*—the practice of generating multiple bins and intelligently combining them greatly increases variability compared with conventional binning-based regression models;
- It is *applicable* to the actual type (size, speed, predictability) of load shift occurring;
- It *correctly identifies the lowest MOER times* (typically times of curtailment).

WattTime's current MOER approach is the best solution we have observed to accomplish these goals by using a hybrid, empirical, data-driven model. It leverages data from both (a) historical hourly power plant-specific generation and emissions data sources such as the US EPA's CEMS in CAMPD[13], and (b) real-time APIs[14] for grid conditions, interchange, and weather.

[13] https://campd.epa.gov/

[14] Such as this API from the US EIA: https://www.eia.gov/opendata/

For each balancing authority in the United States, WattTime applies a multi-stage model to generate MOERs: we separately estimate (1) which non-renewable sources are marginal (both fossil fuels and imports), (2) carbon intensities for each non-renewable fuel source, and (3) marginal curtailment of renewables.

For (1), to identify each fuel type that is marginal at a particular time, we use linear regression on many different subsets of data, binned by grid conditions, along with data science techniques such as deseasonalizing to learn much higher variability patterns in historical data. Where an exclusively data science-based approach would seek to maximize predictive performance, we instead leverage domain knowledge to preselect compelling variables to bin, or parameterize, a grid. Within bins, we run linear regressions on load and fuel types to determine the proportion of load changes being met by each fuel type. Renewables are excluded because the correlation between load and renewables generation is frequently non-causal and biased.

Then, (2) 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 (based on the subset of individual plants changing their output in response to load which varies by hour) 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 depend on the MOER at the time when the plant would otherwise have discharged its energy).

An additional model (3) uses historical renewables curtailment data to learn under what grid conditions and 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.

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

We have also prepared a supplemental document explaining our current MOER model methodology in even more detail, which we can share upon request.

Future Model Upgrades

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

1. **More validation with natural experiments**. WattTime will continue over time to examine ever more natural experiments on observational data to further check accuracy. Examples include both properly identified quasi-random natural experiments (e.g., using regression discontinuity based around both power plant closures, difference-in-difference by comparing the relative spread of different covid waves, etc.) and more straightforward common sense natural experiments (e.g., test the correlation of the times that various models predicted marginal renewable energy and the times when renewable farms were actually curtailed, etc.).
2. **Increasingly, validation with RCTs.** As the gold standard for 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. We are currently in discussions with a number of companies who have adopted AER to integrate an RCT component into their use of AER in future. We are on track for enough signups to have the opportunity to conduct such RCTs as early as this year (2022), though it may be longer to find the staff time required to begin operating any ongoing RCTs.
3. **Grid Data Access.** The ideal MOER model would integrate the vast detailed non-public data that typically only grid operators can access. Many 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 model 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.
4. **Applicability**. WattTime intends to begin providing different MOER models tuned to focus on different types of load shifts. Most immediately, we'll focus on load size. Many models, including our own, 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. In 2022, WattTime currently uses this load size by default because it is a reasonably close match to the total amount of load we estimate is actually 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-sized load shifts.

WattTime is strongly committed to constantly evaluating and updating our approach over time as new data and techniques become available. We welcome criticism and collaboration, in service to our mission of providing data that will create as much positive impact as possible.