

Expanding Energy Efficiency Open-Source Measurement Methods to Incorporate Demand Response for Grid Stability

Joe Glass, Recurve
Stephen Suffian, Watt Carbon
Carmen Best, Recurve
Adam Scheer, Recurve

Abstract

A modern demand-side portfolio must help balance and decarbonize the grid at a time of increasing heat waves, peak supply constraints, and a rapidly evolving generation mix. The simultaneous and interrelated objectives of energy efficiency, electrification, and demand response present a challenge that traditional program and measurement approaches are not able to meet. Developing integrated efficiency and load management programs capable of providing genuine demand flexibility and optimal carbon reductions requires transparent, meter-based measurement applied quickly and equivalently across all resources.

In this paper, we present a new set of open-source comparison group methods called FLEXmeter, which can be applied equivalently to meter energy efficiency and demand response. These methods have been developed and tested for the US Department of Energy, National Renewable Energy Laboratory, and the California Independent System Operator (CAISO) with the support of numerous program administrators and implementers. In the FLEXmeter methods, site-based matching on categorical customer characteristics and load shape produces quality comparison groups. Using CalTRACK and the OpenEEmeter, hourly calculations are performed to generate long-term and “event” baselines, and a normalized difference of differences computation yields adjusted savings that can be attributed to energy efficiency or demand response impacts. These methods and open-source code base are already being deployed to measure integrated programs in California.

Introduction

Evolving Needs of Efficiency and Load Management Programs

As many jurisdictions achieve higher penetration of renewable resources, the electric grid is experiencing increased price volatility and more frequent resiliency challenges. At the same time, a changing climate is driving more severe heat waves which exacerbate stress on the grid. Many states’ climate plans call not only for continued growth of renewables, but also for extensive building and transportation electrification. All these factors create an environment where demand flexibility—the ability to deploy responsive load as a resource—must play a much more prominent role in balancing supply and demand.

While long term load reductions offered by energy efficiency and “instant” load reductions delivered by demand response can both be elements of a larger solution, today’s grid and climate challenges need more holistic and coordinated efforts. Demand flexibility in the form of integrated efficiency and load management interventions is now critical for balancing customer demand with supply side constraints. However, despite overlapping objectives, demand side programs continue to operate and be measured independent of one another. In some cases,

the methods by which certain resources are measured disincentivize inclusion of other resources. Without coordinated programs and measurement, the sum of demand side resources will always be less than the sum of its parts, unable to provide the stability and reliability needed to cultivate a resilient, low-carbon grid.

With the right policy drivers, incentive designs, and advanced meter-based measurement approaches, these challenges can be met through coordinated energy efficiency, electrification, and demand response programs. If programs are used coherently to reduce demand in order to meet supply constraints, excess load from electrification can become a resource instead of a liability. The FLEXmeter methods we present here were developed for the California Independent System Operator (CAISO) and provide a unified measurement standard that can be applied currently to integrated energy efficiency, electrification, and with further advancement can be extended to all demand side resources. Standard measurement is the first step to a fully flexible demand side paradigm.

Limitations of Current Program Approaches

Common methods to baseline demand response programs inherently reward suboptimal consumption patterns in the days and weeks leading to a demand response event. The more energy a customer uses during the peak hours throughout the baseline period, the more they appear to shed load due to the event. This directly disincentivizes efficiency and routine (non-event) load management by only crediting short-term load reductions without considering or incentivizing the value of sustained load shed. The situation is no better in energy efficiency measurement where prescribed energy savings and load shape impacts are assumed and rewarded with no direct link to meter-based performance. Taken together, these measurement barriers force energy efficiency and demand response to operate separately, representing a massive missed opportunity for participating customers, ratepayers, and the grid. Before detailing the FLEXmeter methods, in the next section we describe several additional issues with common demand response measurement that have prevented this resource from being valued appropriately. Remedying these issues will serve as the basis for FLEXmeter and the integration of energy efficiency.

Comments on Existing Demand Response Measurement Methods

The most common demand response measurement methods in circulation are “X of Y” baseline methods (Goldberg 2013) which rely on average observed hourly consumption during the days leading up to an event. The 10 of 10 method takes the average hourly consumption of the 10 most recent non-event days and uses that hourly value as the counterfactual. The 5 of 10 is the same method using only the 5 hottest days. These baselines are often adjusted based on the degree to which event day usage is higher or lower than the baseline during certain non-event hours. This “same day adjustment” is an attempt to capture and adjust for the difference between the baseline and event day consumption. Further, these same day adjustments are often capped to avoid anomalous results from being taken at face value. It is well documented that capped same day adjustments can be too restrictive to capture the magnitude of increased consumption on very hot days. For this reason, California regulators have temporarily lifted the adjustment caps. Unfortunately, any flavor of “X of Y” methods (unadjusted and adjusted, capped and uncapped) are prone to major sources of error and their use has led to controversies and ultimately a lack of confidence in the resource. Among these issues:

Demand response events impact usage during non-event hours as well as event hours. “X of Y” methods take into account all but the two hours directly surrounding an event. When adjusting baselines with hours experiencing a significant rebound effect the event period savings will be biased high. In contrast, programs driving savings during these non-event adjustment hours will be penalized. This is particularly important for integrated EE/DR programs.

The changing mix of end usage throughout the day will introduce errors. A same-day adjustment is likely to be based on midday to early evening pre-event hours and, in certain variations of “X of Y” methods, late evening post-event hours. A typical residential load consists of air conditioning, refrigeration, and baseload in the midday and early evening hours. In contrast, event hours often occur during periods where lighting, cooking, television, dishwashing, and other variable loads are in operation in addition to air conditioning, refrigeration, and other baseloads. Scaling event period baselines based on non-event hours will result in predicting a residential customer’s variable loads based on a fundamentally different (midday) usage mix. The opposite will often be true for the commercial sector.

Peak loads are likely to occur earlier on hot event days than cooler baseline days. During the August heatwave, sweltering high temperatures often exceeded 100 °F and usually occurred between 1 - 5 pm. In response, peak usage in comparison group (non-participant) data occurred earlier than in the cooler baseline periods – and therefore predicted by “X of Y” baseline methods. This effect will tend to bias same-day adjustments high, especially among later event hours (6 - 9 pm).

“X of Y” methods disincentivize efficiency and ongoing load management. As discussed above, with “X of Y” baselines generated using non-event days, demand response providers and participating customers are disincentivized from adopting more permanent load management strategies that can continually deliver value to the grid. In “X of Y” methods, efficiency or routine load shifting in the days leading to an event would lower the event period baseline and, consequently, the savings.

The challenges facing “X of Y” methods will only get worse. Extreme weather variation is likely to become more common, and methods will need to accurately capture the response of intentional demand response interventions in relation to the grid. In addition, as new technologies such as battery storage take root and opportunities arise for more holistic load management via integrated programs, it will be essential that methods for assessing demand response do not disincentivize the more continual load shaping needed to create stability on the California grid.

The FLEXmeter methods are not prone to these significant sources of error or limitations. By using temperature-sensitive models and matched comparison groups, robust results can be obtained without the need for controversial multipliers or restricting customer participation in other programs.

Study Objectives

In 2021, the California Independent System Operator (CAISO) engaged Recurve¹ to conduct a statewide analysis of demand response impacts, refine the FLEXmeter methods, and

¹ <https://www.recurve.com/>

standardize a comparison group approach that can be used as a settlement method for supply-side demand response resources participating in the CAISO market. CAISO intended to gauge if a matched control group and weather matching baseline method could provide a robust assessment of the performance of Proxy Demand Resources and Reliability Demand Response Resources (PDR/RDRR) in relation to similar actors on the grid. In addition, Recurve wanted to ensure that such methods were compatible with integrated energy efficiency and more permanent load shifting measurement. The study, released in 2022, detailed how the FLEXmeter produced quality comparison groups and successfully measured impacts even during extreme events. The measurement methods, open-source codebase, and streamlined process are significant given the critical role demand response can play in a clean energy future and the deep discussions on integrating its flexibility into the grid.

FLEXmeter methods provide hourly impact measurements for demand response that can also be applied to energy efficiency. This paper details the FLEXmeter methods and lays the groundwork for resource agnostic measurement and valuation.

Methodological Foundations

FLEXmeter's core approach consists of the CalTRACK 2.0² methods for hourly methods and the GRIDmeter³ methods. These methodologies, utilized for whole building modeling and hourly comparison group load impact adjustments, were developed in open stakeholder processes and are publicly available. The codebases that implement these methods are also open-source.

To safeguard customer data, the FLEXmeter methods utilize Energy Differential Privacy approaches developed in partnership with the National Renewable Energy Lab. The Energy Differential Privacy⁴ applications have been detailed in several publications and are also supported by an open-source codebase. Differential privacy is an emerging best practice in consumer privacy protection and has been utilized recently by the US Census and Google, among others. All components of the FLEXmeter are described in greater detail below, along with references to their fully specified public sources.

CalTRACK 2.0 Hourly Methods for Baseline and Counterfactual Modeling

The CalTRACK 2.0 hourly methods implemented through the OpenEEmeter Python codebase generate unadjusted hourly baseline models and event-day predictions (counterfactuals) at an individual meter level. The CalTRACK hourly model is a Time-Of-Week

²The CalTRACK methods are based on industry guidelines established by The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE Guideline 14) and the Uniform Methods Project (Chapter 8 - Whole Building Methods). The CalTRACK methods meet all International Performance Measurement and Verification Protocol (IPMVP Option C) requirements. CalTRACK represents the most detailed public specification of IPMVP Option C and includes rigorous steps for data cleaning and organization, weather station selection and weather normalization, and selection of specific model parameters for best fit to the raw consumption data. <http://docs.caltrack.org/en/latest/methods.html>. See www.caltrack.org for more information.

³ <https://grid.recurve.com/>

⁴ <https://edp.recurve.com/>

and Temperature (TOWT) model that is normalized for weather and occupancy.⁵ For each hour in the baseline dataset usage is assigned across up to seven temperature bins.⁶ The model is piecewise linear across the bins.

The CalTRACK methods were developed in partnership with Pacific Gas and Electric Company (PG&E), the California Public Utilities Commission (CPUC), and the California Energy Commission (CEC). A group of industry subject matter experts and stakeholders worked in collaboration throughout the development process. The CalTRACK methods⁷ process is described in full detail at www.caltrack.org (website). The OpenEEmeter Python codebase is publicly available under the stewardship of the Linux Foundation Energy⁸ as an open-source project using an Apache 2 license.⁹ For optimal performance in the measurement of demand response impacts, certain CalTRACK requirements are modified slightly in ways that are discussed in detail in the following sections of this paper.

GRIDmeter Comparison Group Selection and Savings Adjustment Methods

The GRIDmeter method (Recurve, 2020) and codebase¹⁰ generate highly accurate comparison groups using a variety of algorithms. Recurve developed the GRIDmeter methods in partnership with the United States Department of Energy (DOE) and MCE.¹¹ The GRIDmeter methods consist of advanced stratified sampling and site-based matching, and provide recommendations on which procedure should be applied given the data available and precision requirements of a measurement. GRIDmeter takes in meter level loadshapes (predicted or observed aggregated energy usage) and outputs a matched comparison group. Along with sampling methods, the GRIDmeter report also specifies the steps to apply comparison group adjustments to the measurement of demand-side program impacts.

The full description of the GRIDmeter methods is available in the report, Comparison Groups for the COVID Era and Beyond¹². Initially, GRIDmeter methods have been utilized to adjust energy efficiency program savings for the impacts of the COVID-19 pandemic. However, since the sampling algorithms and code were tested and optimized using hourly data, the transition to demand response applications is straightforward. Recurve has successfully applied

⁵ See sections 3.8 and 3.9 of CalTRACK Technical documentation for full detail. Occupancy is handled in the TOWT model by segmenting the times-of-week into periods of high load and low load (also referred to as occupied/unoccupied, although the states may not necessarily correspond to occupancy changes). The segmentation is accomplished using the residuals of a HDD-CDD model.

⁶ Bins are (degrees F): <30, 30-45, 45-55, 55-65, 65-75, 75-90, >90. Bins with fewer than 20 hours are combined with the next closest bin by dropping the larger bin endpoint, except for the largest bin, where the lower endpoint is demand dropped. The bin endpoints are then used to develop the binned temperature features.

⁷ The CalTRACK methods are based on industry guidelines established by The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE Guideline 14) and the Uniform Methods Project (Chapter 8 - Whole Building Methods). The CalTRACK methods meet all International Performance Measurement and Verification Protocol (IPMVP Option C) requirements. CalTRACK represents the most detailed public specification of IPMVP Option C and includes rigorous steps for data cleaning and organization, weather station selection and weather normalization, and selection of specific model parameters for best fit to the raw consumption data.

⁸ <https://www.lfenergy.org/>

⁹ <https://github.com/openeemeter/eemeter/blob/master/LICENSE>

¹⁰ https://github.com/recurve-methods/comparison_groups

¹¹ <https://www.mcecleanenergy.org>

the GRIDmeter and CalTRACK methods and code to measure and adjust meter-based savings for numerous demand-side programs across the United States.

Differential Privacy to Protect Customer Privacy

A great deal of sensitive data was required for the FLEXmeter study. Multiple demand response providers provided participant data, several CCAs supplied non-participant data, and an investor-owned utility provided both participant and non-participant data pools. Data sharing took place under strict third-party non-disclosure agreements. To provide confidentiality to these partners, all results are only presented anonymously.

To minimize the risk of the reidentification of non-participants within the data, Recurve applied differential privacy (Young 2020) algorithms to usage and savings values as well as to figures, utilizing methods developed by the Energy Differential Privacy (EDP) project, to enable the secure sharing of energy data.¹² As part of the EDP Recurve has released an open-source library, eeprivacy,¹³ and uses the corresponding methods throughout this work.

Results Summary

A wide variety of demand response interventions are investigated in the study. They range from residential behavioral interventions to direct scheduling and load control of large commercial customer usage. The FLEXmeter methods were able to reliably assess impacts across market segments providing the CAISO a common view of performance regardless of the program design. Summary results for event period savings, aggregated by anonymized demand response provider and sector, are given in Table 1.

Table 1. Event Load Impacts Summary

DRP	Sector	DRP/LSE/Date Combinations	Unique Participants	Unique CA Climate Zones	Total Participant Event Hours	Avg. Event Hours	Avg. kW Savings	% Savings
A	Res	12	13,496	10	144,890	3.7	0.20	9.5%
B	Res	9	2,771	4	27,652	2.1	0.80	26.9%
D	Res	6	5,077	8	146,094	7.0	0.79	26.2%
C	Non-Res	5	137	6	2,311	5.0	37	28%
D	Non-Res	6	2,758	8	77,062	7.0	1.30	6.9%

The FLEXmeter methods enable load impact measurements across all hours of the day. Because FLEXmeter provides a load impact measurement across all hours, total electricity savings on account of demand response events can be assessed simultaneously with the event-based impacts. Table 2 summarizes the total savings observed for the demand response provider/sector combinations. Assessment at this level demonstrates how seamlessly one could integrate demand response measurement into meter-based energy efficiency programs. An energy efficiency program savings can be measured relative to a long term baseline while during event periods the demand response impact is measured relative to that baseline.

¹² <https://www.energy.gov/eere/buildings/energy-data-vault>

¹³ <https://github.com/recurve-inc/eeprivacy>

Table 2. Total Load Impacts Summary

DRP	Sector	Total Event Savings (MWh)	Total Non-Event Savings (MWh)	Total Savings (MWh)	Avg. Hourly Savings (kWh)	% Savings
A	Res	28.7	-13.1	15.6	0.0	0.9%
B	Res	22.2	-7.6	14.6	0.1	2.9%
D	Res	115.2	-22.6	92.6	0.1	6.4%
C	Non-Res	87	17.3	104	6.8	5%
D	Non-Res	99.9	-129.9	-30.0	-0.1	-0.4%

Negligible to significant positive savings are observed across the programs, which is the expected result given that the programs were primarily targeting savings during the event periods. Non-event negative savings indicates a load shift outside of event windows. This could be a result of measures such as precooling. One program delivered savings in all hours, an indication of a more constant reduction strategy, such as facility shutdown. This level of granularity and insight is only possible through a FLEXmeter approach. Hourly analysis allows for an expansion of integrated programs and targeting for grid health. Disaggregation between energy efficiency and demand response programs is as simple as comparing the corresponding FLEXmeter models.

Anatomy of a FLEXmeter Load Impact Calculation

The FLEXmeter load impacts calculation consists of four primary parts:

1. Comparison group sampling
2. Treatment and comparison group hourly load impact calculations
3. Final load impact calculations via the percent difference of differences adjustments
4. Application of differential privacy to protect non-participant data

Baseline Specifications

The concept of a “baseline period” is important to both comparison group selection and the load impacts calculation steps. In this paper, a “baseline period” refers to the timeframe under consideration in the sampling of comparison groups or the computation of baseline models for the purpose of generating parameters used in that sampling process. To keep computational costs reasonable, Recurve allowed the baseline period for comparison group selection to differ slightly from the baseline period for the model generation used directly in savings calculations. The baseline period under reference is qualified as either the “sampling period” or the “baseline period,” with the former being used exclusively to refer to comparison group selection and the latter being used in the context of savings calculations.

The FLEXmeter methods utilize a sampling period that consists of a 45-day pre-event timeframe in which all days with a demand response event are eliminated from the analysis or “blackened out.” The 45-day window provides enough time that the 20 non-event day threshold is not expected to pose a data sufficiency issue while not so much time as to include data that is less relevant to gauging event day impacts, such as early spring consumption patterns or data that may be more prone to changing consumption trends, including those associated with COVID. As is the case in most traditional demand response baseline methods, no post-event data are used for comparison group selection. Including post-event data can help ensure weather patterns are captured that best represent event day weather and is a key benefit of this approach. The baseline

period for savings calculations utilizes a 45-day pre-event and 15-day post-event window, again with event days blacked out.

The definition of baseline periods is a dynamic component of the FLEXmeter codebase. Extending a baseline period to a year results in a CalTRACK compliant energy efficiency model. A year-long baseline model can also be used to measure regular load shifting. The underlying methods are agnostic to the intervention being measured. Different programs do not require different codebases, only different baselines and corresponding sufficiency requirements. In the case of combined energy efficiency and demand response projects, both a DR counterfactual and an EE counterfactual are generated. The EE savings can then be attributed as the EE counterfactual less the DR counterfactual with the DR impacts being the DR counterfactual less the observed consumption. This will be discussed further below.

GRIDmeter Comparison Group Sampling

GRIDmeter comparison group sampling starts with a pool of non-participating customers (the “comparison pool”). From the comparison pool, the objective is to sample a group of customers (the “comparison group”) that best represents the program participants (the “participant group” or “treatment group”).

In the FLEXmeter methods, comparison groups are selected utilizing the GRIDmeter individual site-level matching approach. The site matching is conducted on the basis of modeled average weekly load shape to select the comparison pool customers that are the closest representation of each individual treatment meter. By using average weekly load shape as the basis for site matching, the FLEXmeter captures and accounts for both the weekday and weekend usage patterns displayed by individual customers.

A modified CalTRACK 2.0 hourly model is used to produce the average modeled weekly load shape. Using a modeled average weekly load shape ensures that all hours of the week have the same influence over site-based matching regardless of imbalances in the number of weekday days vs. weekend days that may be present in the baseline data or any days of the week that may have lower representation due to event-day blackouts.

The site level matching takes place on a stratified sample of the comparison pool. Flexmeter stratifies by sector (residential or non-residential), solar PV status and climate zone. These sampling requirements are designed to ensure that residential and business customers are appropriately separated, and that treatment and associated comparison customers experience similar grid conditions and weather patterns and react similarly to those patterns. This ensures sites with similar load profiles due to lurking variables will not be matched across key categories.

For each stratified group, GRIDmeter calculates the sum of squares of the differences in the modeled weekly load shape for every possible combination of treatment and comparison pool meters. Each treatment meter is simultaneously assigned the comparison group meter with the lowest sum of squares value. Duplicate meters are dropped from the comparison group, and the comparison group meters assigned are removed from the remaining comparison pool ensuring that a comparison meter can only be assigned to a single treatment meter.

Figure 1 shows the results of the site-based meter matching for the example event. This treatment group consisted of about 1,500 customers and was matched to about 5,300 non-participants. Though the matching is conducted based on average weekly load shape, this figure represents the aggregated average daily load shape in order to limit privacy expense. (With more data points, the privacy cost of a weekly load shape is greater than a daily load shape.) The

left panel shows weekday load shapes and the right panel shows weekend load shapes. The solid green line is the average baseline modeled load shape of the treatment meters. The dotted orange and dashed red lines are the analogous traces for the comparison pool and selected comparison group, respectively. The site-based sampling clearly yields a much better match on an average customer basis than would be expected from random sampling with improvement evident across both weekdays and weekends.

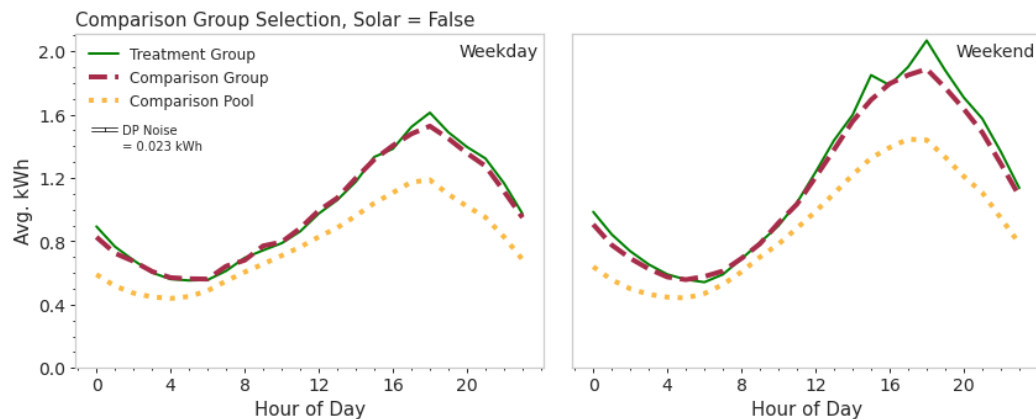


Figure 1. The average weekday (left) and weekend (right) load shape of a meter in the treatment group (solid green), comparison pool (dotted orange), and comparison group (dashed red).

Quality matches have been achieved across the entire spectrum of treatment customers. Figure 2 shows the average modeled daily load shape for treatment (solid green) and comparison pool (dotted orange) customers grouped into quartiles of max to min daily consumption ratio. .

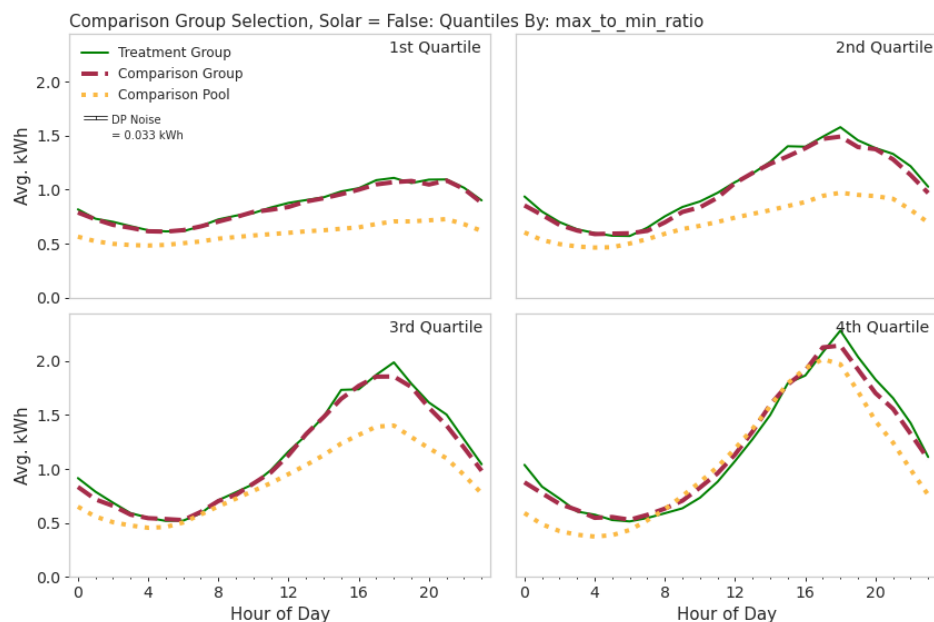


Figure 2. The average load shape of a meter in the treatment group (solid green), and comparison pool (dotted orange) broken out by baseline usage quartiles. Matched comparison group average load shapes are also shown (dashed red). The lowest quartile is in the top left panel and the highest quartile is shown in the bottom right panel.

The dashed red lines are the average load shapes of the comparison group customers that have been matched to the treatment customers represented in each quartile. A good match has been produced across each of these groups.

Treatment and Comparison Group Load Impacts Calculation

With the treatment and comparison groups in place, a baseline model is calculated for each meter. As a reminder, this baseline period for savings calculations is the 60-day window (45 pre, 15 post) surrounding the day of the event in question with all demand response event days blacked out. Using the baseline model, a prediction of energy consumption for the event day is produced based on outside temperature and hour-of-week variables. This prediction is known as the counterfactual. This step is conducted on an individual meter basis and the individual-meter counterfactuals are aggregated within the treatment and comparison groups.

The dashed blue line in Figure 3 shows the average treatment meter counterfactual for the example event. (At this point comparison group adjustments have not been applied.) The open and filled circles are the average treatment meter actual observed usage during non-event and event hours respectively.

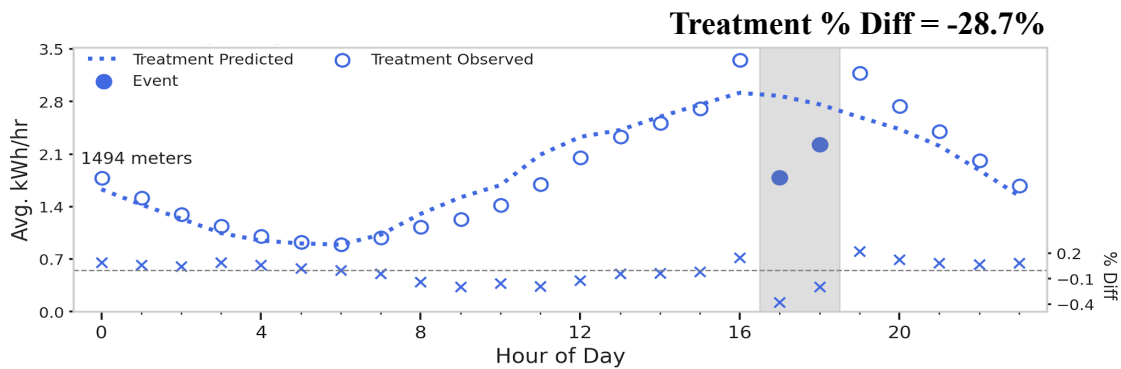


Figure 3. The dotted trace gives the average counterfactual (model-predicted) usage of a participating customer. The circles are the average observed meter readings (open = non-event hour, filled = event hour). The fractional difference between the two traces is shown as X's and refer to the right-hand axis.

Finally, the X's (right-hand axis) show the *fractional* difference between observed and counterfactual usage, which is calculated as follows:

$$\%Diff_{Treatment,i} = (Observed_{Treatment,i} - Counterfactual_{Treatment,i}) / Counterfactual_{Treatment,i} \quad (1)$$

Where the subscript i refers to the hour of day. The dashed horizontal line indicates zero difference between counterfactual and observed usage.

Several observations in Figure 6 can serve as good barometers for the comparison group insights and adjustments that will follow. First, the counterfactual is close to actual usage for most non-event hours, but there are some exceptions. For instance, in the late morning (hours 8 - 12), observed usage is consistently lower than counterfactual usage. Also, in the hours immediately surrounding the event, observed usage is significantly higher than the counterfactual. These differences may be on account of altered customer behavior due to the

event or may be model error. Without additional information, it would be very difficult to appropriately attribute these discrepancies.

Second, during the event period, a substantial demand response in consumption is observed and this demand response is clearly not evident in the counterfactual. Therefore, upon first glance, it would appear that the event had a significant impact. However, without a comparison group, it may not be possible to definitively assign these event savings to the demand response program. After all, during this week there had been rolling blackouts and public messaging asking customers to reduce evening load. It could be possible that these non-program factors altered customer behavior. An analysis of the comparison group can provide more concrete answers and associated results are given in Figure 4.

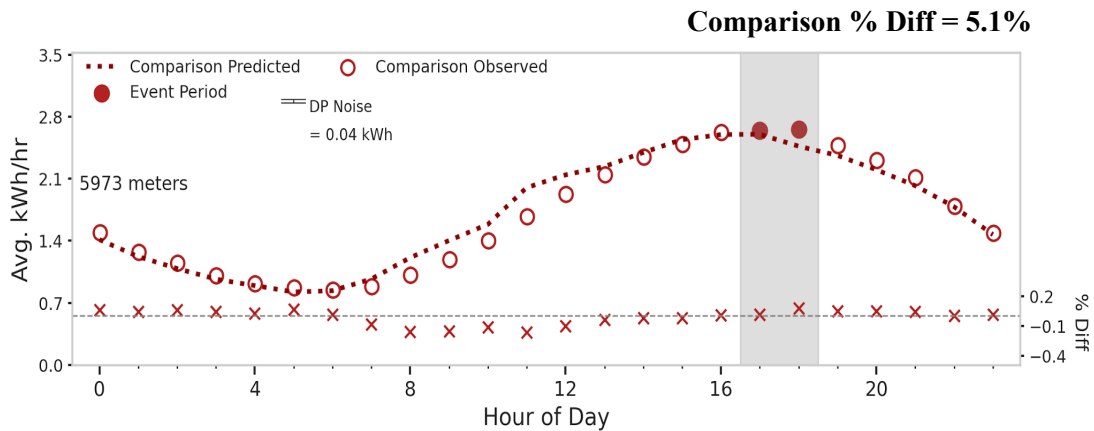


Figure 4. The dotted trace gives the average counterfactual (model-predicted) usage of a comparison group customer. The circles are the average observed meter readings (open = non-event hour, filled = event hour). The fractional difference between the two traces is given as X's and refers to the right-hand axis.

Looking first at the late morning hours in the comparison group, we see that the same pattern is observed as was seen in the treatment group: actual consumption consistently falls below the counterfactual. Therefore, this discrepancy is attributable to model error and, in step 3 (comparison group load impact adjustments) will be effectively removed.

However, in the hours immediately surrounding the event, the comparison group exhibits little difference between the observed and counterfactual usage. Thus, the increased consumption seen in the treatment group during these hours is attributable to the event. This phenomenon is known as the “rebound effect” and has been well documented for many demand-side programs. We observed rebound for many of the events studied here. For grid planners and LSE's, it is important to understand and account for the rebound effect when calling demand response events given they may have other unintended effects on reliability.

Finally, during the event hours, only a slight difference between the comparison group observed and counterfactual usage is seen, with the latter being marginally higher. Therefore, we can conclude with confidence that the demand response in consumption during event hours in the treatment group was indeed due to the intervention. As with the treatment group, the fractional difference between observed and counterfactual usage (Figure 4 red X's) is calculated by hour as follows:

$$\%Diff_{Comparison,i} = (Observed_{Comparison,i} - Counterfactual_{Comparison,i}) / Counterfactual_{Comparison,i} \quad (2)$$

Adjusted Load Impact Calculations Via % Difference of Differences

Comparison group adjustments are made to the treatment group savings via equations 3 and 4. Equation 3 adjusts the fractional difference between observed and counterfactual hourly usage in the treatment group by the fractional difference of the analogous quantities in the comparison group.

$$\%Diff\ of\ Diff_i = \%Diff_{Treatment,i} - \%Diff_{Comparison,i} \quad (3)$$

$$Load\ Impact_i = \%Diff\ of\ Diff_i \times Counterfactual_{Treatment,i} \quad (4)$$

Both model error and exogenous factors are estimated and removed in the final load impact calculation. Final load impacts are determined by multiplying the result of Equation 3 by the treatment group's hourly counterfactual (Equation 4). For the example event, the results of Equations 3 and 4 are shown in Figures 5 and 6.

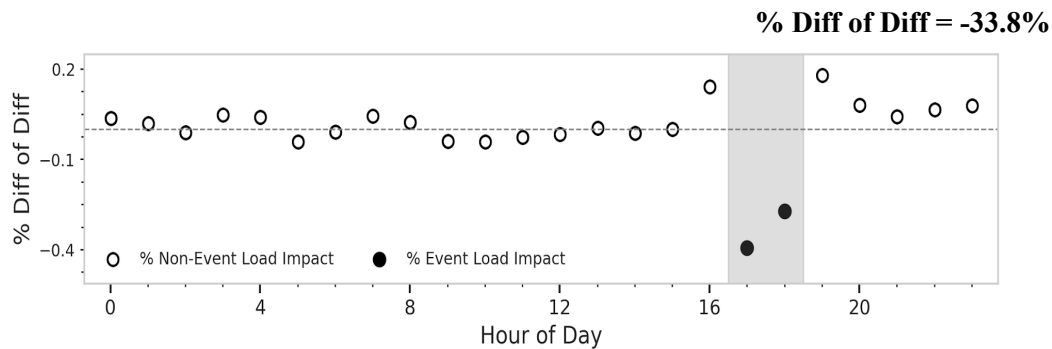


Figure 5. The fractional difference of differences (Equation 3) between treatment and comparison group. Open and filled circles represent non-event and event hours, respectively. This plot represents the comparison group adjusted fractional load impact of the participants.

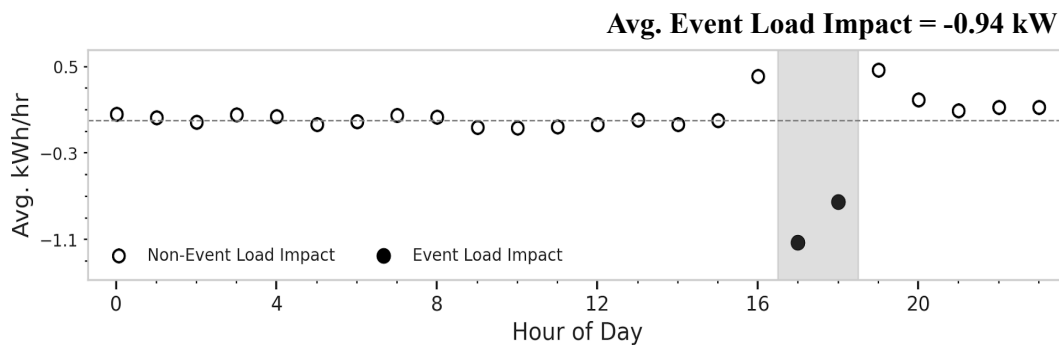


Figure 6. The final average participant comparison group adjusted hourly load impacts (Equation 4). Open and filled circles represent non-event and event hours, respectively.

The components of Figures 3 - 6 can be effectively condensed into a summary plot (Fig. 7) that compares observed consumption to an adjusted counterfactual. The adjusted counterfactual represents the treatment group counterfactual modified by the trends observed in

the comparison group (Equation 5). This adjusted counterfactual should be considered the FLEXmeter prediction of hourly consumption absent the intervention.

$$\text{Adjusted Counterfactual} = \text{Counterfactual}_{\text{Treatment},i} \times \text{Observed}_{\text{Comparison},i} / \text{Counterfactual}_{\text{Comparison},i} \quad (5)$$

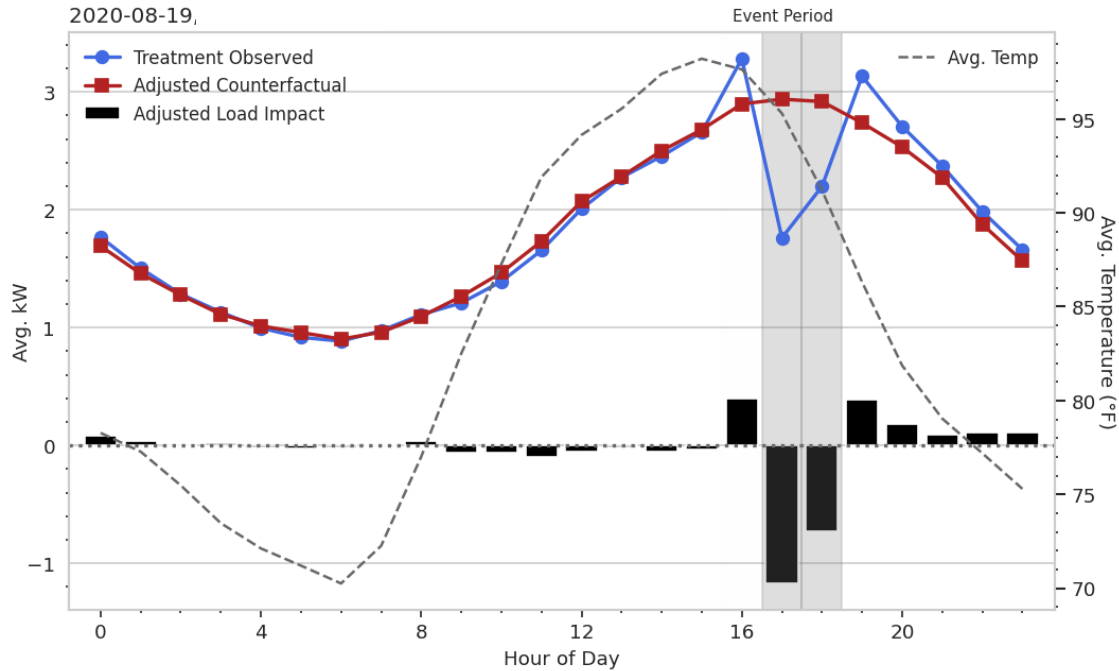


Figure 7. Summary figure comparing average participant event day usage (blue circles) and adjusted counterfactual (Eq. 5; red squares) along with hourly load impacts (black bars). The adjusted counterfactual in this figure has been privatized with $\epsilon = 10$ as in Figure 7.

Also included in Figure 7 is the average temperature participants experienced on the event day. As is the case for many of the August 2020 heatwave days, temperatures peak near or above 100 °F in the early afternoon, before the event period. While temperatures remain high there is a rapid drop from 5 - 9 pm.

Generating models and producing counterfactuals is essential for reliable comparative analysis. Despite the reasonable comparison group match shown in Figures 3 - 5, the observed usage in hour 0 of the event day differs by about 0.25 kWh, or 16%, between the treatment and comparison group. If the assessment was only made on the basis of observed usage, we would conclude that the program was responsible for this difference. But with the model prediction, we have a critical gauge of why this difference exists and whether it should be assigned to the intentional intervention rather than typical consumer behavior. There is an *expected* difference between the treatment and comparison counterfactuals of about 0.17 kWh (12%). Therefore, only a small portion of the difference (about 0.09 kWh) should be associated with the program response. Without the modeling step, there would be enormous pressure to produce a near-exact match to the treatment group in the comparison group sampling step, a prospect that may not even be possible given the limitations of most datasets.

Load impacts are calculated on a percentage (normalized) basis and then applied to the counterfactual usage to produce a final load impact. When treatment customers are unique and/or comparison pools are small, it may be impossible to find a sufficient number of highly similar comparison matches for all treatment meters. In such cases, the hourly usage of average treatment and comparison meters can significantly differ. If calculating the comparison group adjustment without the normalization of the *% Difference of Differences* steps, the adjustment will directly reflect the difference in the magnitude of consumption between treatment and comparison customers. A schematic example of this effect is given in chapter 4 of the report *Comparison Groups for the COVID Era and Beyond*. With the normalization step, differences in the relative sizes of treatment and comparison customers do not manifest directly as under- or over-counted savings.

Compatible Measurement for Demand Response and Energy Efficiency

The FLEXmeter methods provide a novel and robust way to assess the impacts of demand side programs. CalTRACK models paired with a comparison group correction factor deliver accurate, stable, and replicable energy models. This modeling approach is not limited to demand response counterfactuals. When measuring demand response, the blackout dates are selected as the days a resource was called to participate in an event. The model is then built excluding any event usage data. This ensures that the resource response does not impact the counterfactual prediction. This is perfectly analogous to energy efficiency, where the blackout dates are defined as every day that encompasses an efficiency intervention. A CalTRACK energy efficiency model is built using baseline data of a building prior to an intervention, thus keeping the impacts of the efficiency project out of the baseline model.

The baseline period duration is also critical to defining the character of a CalTRACK counterfactual. For demand response, a short duration baseline of 45 days is used, and can be supplemented with 15 additional days post event. Short duration models capture the immediate behavior of a building without introducing bias from long term usage patterns. Conversely, for energy efficiency a year long baseline is used to build the model. This allows for the usage patterns of the entire year to be reflected in the counterfactual, which is often used for measurement for multiple years. For each program type the baseline duration directly reflects the program duration. For combined efficiency and demand projects, both efficiency and demand response counterfactuals can be independently developed and savings attributed. This is done as shown schematically in Figure 8.

FLEXmeter produces consistent outputs regardless of which input parameters are selected. A demand response baseline model will produce an identical savings dataset as an energy efficiency model in terms of column names and interpretability. Both produce datetime indexed counterfactual models that can be joined into a single dataset.

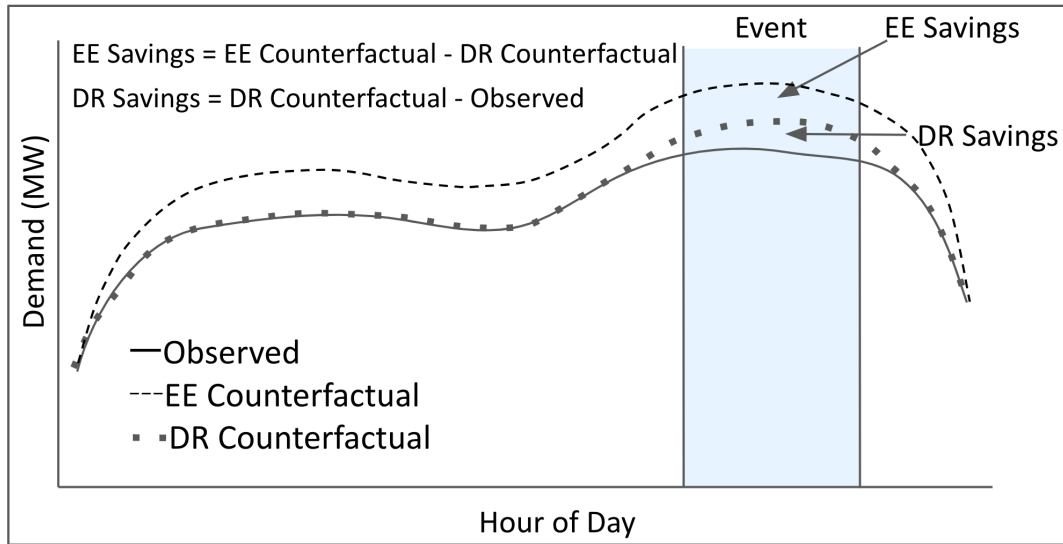


Figure 8. Savings measurement for integrated program participants. During an event the demand response savings are measured relative to the observed usage. The energy efficiency savings are measured relative to the demand response counterfactual during an event and relative to the observed otherwise.

Data Access and Recommended Future Pathways

FLEXmeter presents a potential breakthrough for integrated program design, but depends on comparison group adjustments to achieve the best possible results. While non-participant energy consumption data is a primary component of the required data, other information regarding non-participant customers, including their sector and location is required for population stratification. In California, load serving entities may provide access to covered data without consent for activities classified as a "primary purpose," which includes activities that enable grid operations or in any case authorized explicitly by order of the California Public Utilities Commission.¹⁴ Data protections are provided through non-disclosure agreements with third parties¹⁵ to ensure security provisions are in place and can be contractually enforced.

But these provisions apply only to load-serving entities. Non-load serving entities such as state agencies or third-party demand response providers are dependent on the legislature (for example, California Energy Commission's successful petition for data access under AB802 (2015)) or rely on Commission order (like the Technology and Equipment for Clean Heating (TECH) Initiative authorization for third party data access via [D.20-03-027](#)). The data for

¹⁴ In Decision [11-07-056](#) July 28, 2011 the Commission established the following:

Finding of Fact:

12. It is reasonable to define as "a primary purpose" the use of information to:

- (1) provide or bill for electrical power;
- (2) fulfill other operational needs of the electrical system or grid,
- (3) provide services as required by state or federal law or specifically authorized by an order of the Commission, or
- (4) implement demand response, energy management, or energy efficiency programs under contract with an electrical corporation, under contract with the Commission, or as part of a Commission authorized program conducted by a governmental entity under supervision of the Commission.

¹⁵ See Finding of Fact 46 in Decision [11-07-056](#) for the specific rules defining use and disclosure limitations.

comparison group analysis is extremely challenging to obtain in practice and can impede the ability to actually execute comparison group methods.

Comparison group analysis is one of the most robust means of assessing demand response impacts and the provision of non-participant data for this type of analysis should be seamless and straightforward. The California Public Utilities Commission has already classified implementation and evaluation of demand response performance analysis as a "primary purpose," clarifying that load-serving entities can provide non-participant data as it directly supports grid operations. This should extend beyond demand response to include energy efficiency programs. Comparison groups consistently demonstrate significant value to modeling and are a crucial component of FLEXmeter. It is imperative that safe and secure access to non participant data be expanded.

Conclusions

FLEXmeter provides standard performance-based methods for any demand-side intervention and can help reconcile the view of performance across state agencies, currently siloed demand-side programs, portfolios and regulatory silos. Comparison groups not only remove the exogenous model error of extreme weather, they also address sudden trend changes such as the impacts of COVID-19. It is critical to expand the availability of data for comparison groups to allow for accurate counterfactual construction.

By incorporating both a comparison group and weather-normalized hourly modeling applied equivalently to treatment and comparison customers, the FLEXmeter approach provides a robust and reliable M&V methodology. The model is sensitive to weather and the comparison group provides adjustments for exogenous effects and directional model error. Most importantly the FLEXmeter provides an integrated approach to both efficiency and load management programs. FLEXmeter methods are transparent, meter-based, and can be applied quickly and equivalently across all resources.

References

A. Aktay, S. Bavadekar, G. Cossoul et al. [*Google COVID-19 Community Mobility Reports: Anonymization Process Description \(version 1.0\)*](#), (2020).

California Independent System Operator, California Public Utilities Commission, and California Energy Commission, *Root Cause Analysis: Mid-August 2020 Extreme Heat Wave*, January, 2021

Facebook, [*New privacy-protected Facebook data for independent research on social media's impact on democracy*](#), February, 2020.

J. Glass, S. Suffian, A. Scheer, C. Best Recurve, [*Demand-Response Advanced Measurement Methodology*](#), February 2022

M. L. Goldberg and G. K. Agnew, DNV KEMA, [*Measurement and Verification for Demand Response*](#), Energy and Sustainability, 2013.

McGee Young (Recurve), Marc Paré (Recurve), and Harry Bergmann (DOE), [*Differential Privacy for Expanding Access to Building Energy Data*](#), ACEEE Summer Study 2020.

Recurve, [*Comparison Groups for the COVID Era and Beyond*](#), Submitted to the United States Department of Energy, September, 2020.