The California ISO commissioned this report to address concerns about the accuracy of the ISO's current methodology for measuring supply side demand response resource performance. The goal of this analysis was to identify and test an option for measuring demand response resource impacts on grid operations, particularly during extreme stress on the power system. Specifically, Recurve tested and refined open source comparison group methods to evaluate demand response impacts, focusing analysis on the August 2020 heatwave events. The report also explores ways to further improve the methodology for correctly assigning real time performance value.

The attached report details Recurve's methodology, analysis, and results for several demand response providers participating in the study.

Based on the findings on this report, the ISO is able to confirm that the FLEXmeter methodology is a tariff-compliant option that demand response providers can use to settle ISO market dispatches. The report also provides recommendations for strengthening its control group business practice specifications. Appendix A to the attached report provides detail on Recurve's FLEXmeter methodologies and the current ISO Tariff control group methodology.

The report's updated baselining methodology together with the comparison group selection and adjustment process offer reliable and consistent demand response resource performance assessment, particularly in extreme weather situations like August 2020.

As the independent market operator, ISO seeks to settle demand response transactions transparently, appropriately, and fairly for their supply-side performance value. The FLEXmeter methodologies described in this report would allow the ISO to fulfill that objective and consistent with the ISO Tariff. The study results support consideration of its use in other performance-based assessments of demand response.

The ISO is committed to working with the California state agencies to overcome the barriers of data access and handling that enable development of quality comparison groups if it is found to be preferred. We look forward to supporting the State agencies and stakeholders in considering the use of a clear and transparent performance methodology not only for correctly assessing real time performance for demand response, but one that can be consistency applied across all agencies.

Anna McKenna

Anna McKenna
Vice President Market Policy and Performance

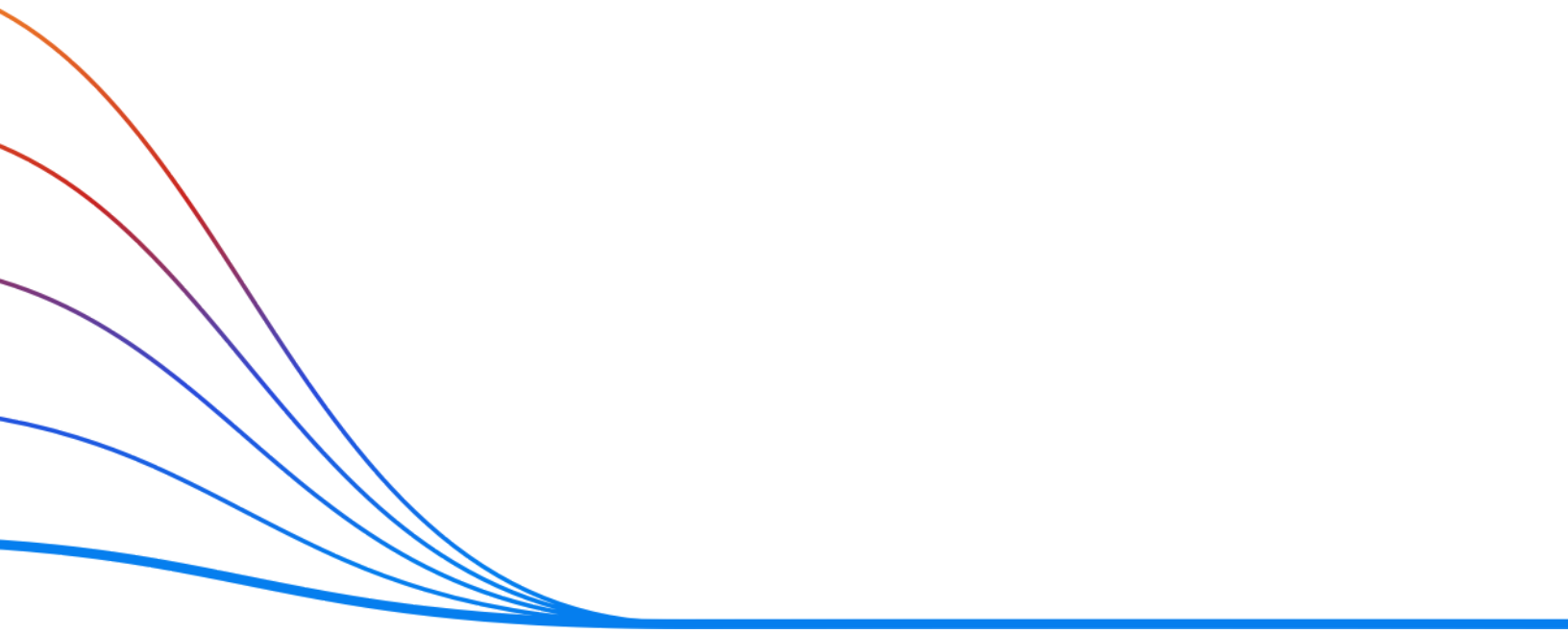
Demand Response Advanced Measurement Methodology

Analysis of Open-Source Baseline and Comparison Group Methods to Enable CAISO Demand Response Resource Performance Evaluation

Prepared By: Joe Glass, Stephen Suffian, Adam Scheer, and Carmen Best

Prepared for:

Market Infrastructure and Policy Development
California Independent System Operator (CAISO)



February 2022

Table of Contents

Executive Summary

Demand response is a critical resource that must scale effectively to balance the grid at a time of increasing heat waves, peak supply constraints, and a rapidly evolving generation mix. However, questions around measurement accuracy and best settlement practices can cloud performance data and negatively affect future market design and policy.

California's climate change-induced rotating power outages of August 2020 exposed shortcomings of common demand response measurement methods. Certain methodological restrictions have since been eased, but accuracy and reliability remain concerns.[1] In addition, the high degree of variation in current demand response methods has created significant uncertainty for demand response providers, utilities, forecasters, and grid operators.

The FLEXmeter demand response measurement methods can offer a more robust path forward to enable demand flexibility resources as a grid resource. FLEXmeter comparison group methods are designed to be revenue-grade,[2] utilizing open-source code and verifiable implementation to enable auditable settlement. These methods were first developed and tested for the US Department of Energy and National Renewable Energy Laboratory with the support of MCE and OhmConnect.[3] After reviewing the results of this study, the California Independent System Operator (CAISO) engaged Recurve to conduct a similar analysis statewide, refine the FLEXmeter methods, and standardize an approach that can be deployed at scale to enable its use as a performance measurement for supply-side demand response resources participation in the CAISO market.

This study demonstrates the effectiveness of the FLEXmeter methods for assessing DR performance during California's mid-August 2020 heatwave and resulting rotating power outages. Recurve analyzed 38 demand response events spanning Investor-Owned Utility (IOU) and Community Choice Aggregator (CCA) territories across 11 climate zones, multiple demand response providers (DRPs), and more than 24,000 participating residential and commercial customers.

The FLEXmeter approach produced quality comparison groups and successfully measured impacts even during extreme events. The measurement methods, open-source framework, and streamlined process are significant and timely given the critical role DR can play in a clean energy future and the deep discussions on integrating its flexibility into the future grid.

---

[1] The same day adjustment cap employed in "X of Y" methods, including 10 of 10, was at times insufficient to capture the degree of increased consumption among DR participants during extreme heat events.

[2] Beyond using approved methods, here we use the term "revenue-grade" to refer to measurements that utilize open-source code, verifiable data, and whose execution can be fully audited.

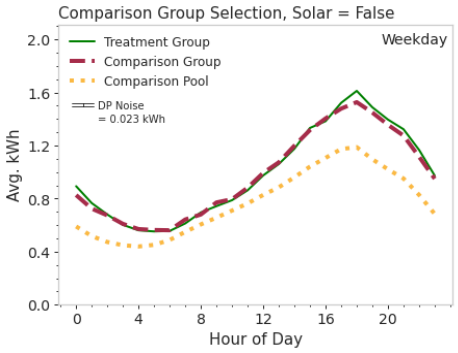
[3] *Applying Energy Differential Privacy To Enable Measurement of the OhmConnect Virtual Power Plant*, Marc Paré, Mariano Teehan, Stephen Suffian, Joe Glass, Adam Scheer, McGee Young, Matt Golden, December, 2020

A. Summary of Methods

This report includes detailed descriptions of the FLEXmeter methods and measurements of the August 2020 demand response events resulting from supply-side demand response resource dispatches during the extreme heatwave. Readers will find three core components of the analysis: comparison group matching, measurement of event impacts, and the application of differential privacy.

The FLEXmeter methods can be summarized as follows: Event day usage of participating demand response customers is compared to a modeled prediction of usage in the absence of the event. An identical calculation is conducted for a group of non-participating customers of similar types and usage patterns. Hourly load impacts are taken as those calculated for the participating customers, adjusted for impacts measured in the non-participant sample.

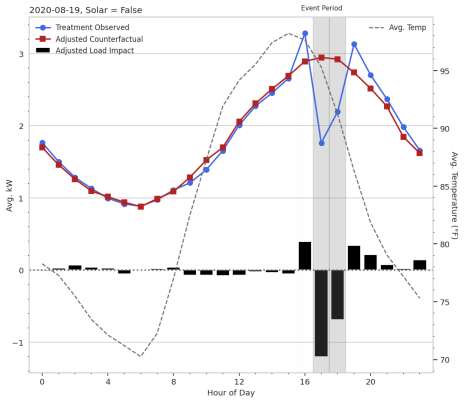
1) GRIDmeter Comparison Group Matching



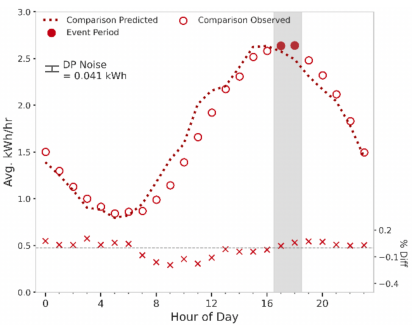
GRIDmeter sampling yields a group of non-participating customers (comparison group) that best represent participants (treatment group). With site matching based on categorical parameters and load shape, the specific comparison pool meters that are the closest representation of an individual treatment meter are selected into the comparison group.

2) FLEXmeter Demand Response Measurement

FLEXmeter calculations utilize a two-stage approach. First, The CalTRACK methods, run via the OpenEEmeter, generate hourly baseline models for all treatment and comparison meters. Next, a comparison group adjustment is generated by the GRIDmeter, and the difference of the differences in the load impact attributable to the program.



3) Energy Differential Privacy



Recurve has employed modern methods to protect customer data from re-identification. Incorporating Energy Differential Privacy into the analysis, a calibrated degree of noise is introduced into aggregate statistics to mask the presence of individuals in underlying datasets. This approach provides privacy protection beyond commonly required aggregation approaches.

B. Study Limitations

This study was enabled by the voluntary participation of DRPs and LSEs and Recurve focused analysis in regions where data for participating and non-participating customers were available. Therefore, results should not be considered representative of a DRP's entire program or the entire state's experience of the events.

Recurve focused sampling where limited comparison pool (non-participant) data were available. These data were not always a fully random representation of an LSE's service territory, and therefore sampling was not always proportional to program enrollment. Recurve did not always have access to participation data for other demand response programs. In a few cases comparison groups exhibit some load reductions during event periods that may be explained by participation in other programs.

At least one program called events that began or ended partway through an hour. In these cases, such hours were treated as event hours without further adjustment. When these hours are labeled as part of the event this approach will tend to account for all savings but would be expected to lower average and percent savings values.

Finally, Recurve notes the difficulty of constructing counterfactuals (predicted usage) for extreme heatwave days when baseline data contain only moderate temperatures, especially with competing Flex Alerts and rotating power outages. However, comparison group adjustments appear to largely capture and account for such exogenous factors, non-linear consumption patterns, and other effects such as customer fatigue and thermal inertia that can be difficult to model.

Despite these caveats, the main objective of this study has been achieved - to test, refine, and publish detailed comparison group methods that are compliant with the CAISO tariff for performance settlement and immediately available for use. With complete data these methods can serve to provide more accurate, consistent, and reliable measurements and lend critical confidence in demand response resources.

C. Summary of Results

A wide variety of programs are investigated here from residential behavioral interventions to direct scheduling and load control of large commercial customer usage. Summary results for event period savings, aggregated by anonymized DRP and sector, are given in Table E1.

Table E1: 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%

Of the 38 distinct DRP/LSE/event date combinations studied, all had positive savings, including 35, 25, and 15 that had event period load reductions of greater than 5%, 10%, and 20%, respectively.

The FLEXmeter methods enable load impact measurements across all hours of the day. Figure E1 summarizes savings by hour type for each combination of DRP and sector. Fractional savings (y-axes) are plotted against hour type (x-axes) for residential programs in the left-hand panels and non-residential programs in the right-hand panels. Event hours are shown in blue and are numbered 1 - 7. (Some events lasted up to 7 hours.) Non-event hours are shown in red and are grouped from left to right into three categories:

- More than 2 hours before or after an event (+/- > 2 Hours)
- 2 hours before or after an event (+/- 2 Hours)
- The hours immediately preceding or following an event (+/- 1 Hour)

Each DRP/sector combination shows a distinct pattern. Some programs have high average event savings but may have fewer customers and/or shorter event periods. Some programs show strong savings persistence over long events while others experience significant degradation. Some programs exhibit high degrees of increased usage in the hours immediately preceding and following events ("the rebound effect" or "take back"), while others show savings in non-event hours.

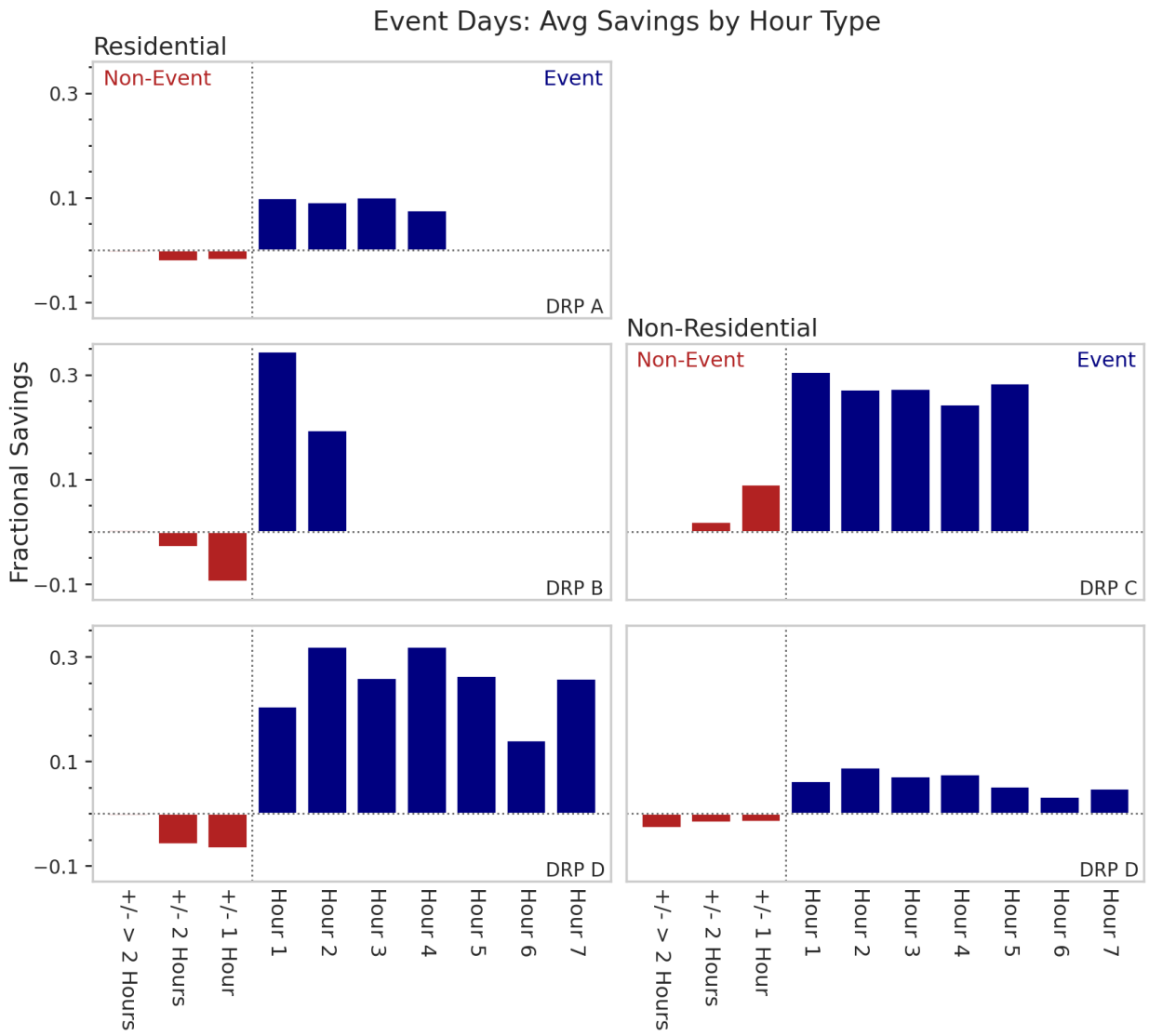


Figure E1: Average adjusted fractional savings by hour type for each combination of DRP and sector. Residential programs are shown in the left-hand panels and non-residential programs in the right-hand panels. Non-event hours are shown in red and event hours are shown in blue.

Because FLEXmeter provides a load impact measurement across all hours, total electricity savings on account of demand response events can be assessed. Table E2 summarizes the total savings observed for the DRP/sector combinations.

Table E2: 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. Three of the five programs delivered more than 2.5% reduction in total electricity consumption on the event days.

D. Comments on Existing Demand Response Measurement Methods

While extensive side by side comparison of FLEXmeter results against those of current common demand response measurement methods was beyond the scope of this work, Recurve provides some comments on "X of Y" baseline methods.[4]

These methods, including "5 of 10" and "10 of 10," average observed hourly consumption during pre-event days, which, themselves, are not event 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 often capped. It is well documented that the caps were too restrictive in the California August 2020 heatwave to capture the magnitude of increased consumption. Regulators have temporarily lifted the adjustment caps, providing relief in cases of extreme weather. However, we perceive several fundamental issues with "X of Y" baseline methods, whether capped or uncapped, which we draw attention to here:

1. **Events impact usage during non-event hours as well as event hours.** While "X of Y" methods may not utilize the single hours immediately preceding and following an event, Figure E1 shows that demand response events can substantially impact usage during non-event hours throughout the day. This "rebound effect" differs widely by program and event and many examples can be seen in the Extended Results Appendix to this report. 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.
2. **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. Consider a typical residential customer; usage in midday and early evening hours will

[4] Additional context for "X of Y" baseline methods are available in: *Measurement and Verification for Demand Response*, M. L. Goldberg and G. K. Agnew, DNV KEMA Energy and Sustainability, 2013.

consist of a high proportion of air conditioning, along with refrigeration and other baseload devices. 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.

3. **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, Recurve observed that 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).
4. **"X of Y" methods disincentivize efficiency and load ongoing load management**. 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.

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 will benefit from continued research and implementation. But by design they are not prone to these significant sources of error and limitation. 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.

E. Advantages of Common and Consistent Measurement

This study demonstrates that the FLEXmeter methods can be implemented across a wide variety of demand response programs. In practice, a consistent application of FLEXmeter could lend confidence and consistency to measurement, putting demand response resources on a more solid footing in energy markets. Consistency in measurement would also yield important advantages, including comparability within and between programs. Comparability in turn enables program optimization as differences in outcomes can be readily attributed to the program itself and not methodological choices. If measurement is viewed as biased, unreliable, or unpredictable, the entire demand response resource cannot be scaled with the confidence of all stakeholders.

While Recurve believes this study takes a large step toward a more robust set of open source methods that can be applied consistently across programs, the methods themselves require data for implementation. These data include hourly non-participant usage across service territories for comparison group sampling, categorical data to ensure comparison groups are representative of treated populations, and event participation data, preferably not only of the program in question but of other programs that may be operating in the same territory. While data can be anonymized, it does need to be at an individual-meter level. In Recurve's experience pre-aggregated samples fail to capture the variety of usage profiles present in the market, can lead to measurement delays, and are difficult to verify.

Below Recurve describes recommendations for data sharing and privacy protection and in the appendices we detail data specifications that demand response providers and load-serving entities can use to enable measurement via FLEXmeter.

F. CAISO Study Objectives and Recommendations

The California Independent System Operator commissioned this study to address four key objectives.

Understand and operationalize the baseline and comparison group methods in relation to existing guidance and practice;

Recurve assessed the FLEXmeter comparison group methods in relation to the FERC-approved control group methodology outlined in the CAISO tariff. The FLEXmeter methods are specified to a greater degree and are compliant with the tariff. The full summary of the methods comparison is provided in Appendix A.

Understand barriers to data access and identify a viable path to overcome them

Recurve demonstrated that with the proper data organization, frequency, and security protections in place, a centralized settlement model can both handle data securely and present results while mitigating the risk of re-identification. Standardized data specifications can streamline the execution of FLEXmeter methods going forward. Recommended specifications for both consumption and categorical data are detailed in Appendix B.

Relying on data donors for non-participant data to enable this study meant that the full cross-section of the state was not available. Recurve recommends that state agencies coordinate to authorize the development of a centralized non-participant data pool to enable comparison group settlement methods. In addition, Energy Differential Privacy enhances customer data privacy while still allowing the extraction of the necessary information to support demand flexibility as a reliable CAISO market resource.

Understand the 2020 heat storm events based on the baseline and comparison group methods implemented by Recurve to inform and support decision making;

The 2020 extreme heat events are the subject of the measurements detailed in this study. Since the FLEXmeter methods comply with the existing tariff, distributed energy resource

providers can utilize them immediately for settlement with access to the necessary data. The methods are observed to work well across geographies and for both residential and commercial demand response, illustrating their viability for wider adoption.

This report can inform and support contemporary and future decision-making across state agencies. Since FLEXmeter methods are available as open-source frameworks, they could serve as a transparent foundation for assessing performance across market actors and agencies. Some near term considerations for each agency include:

- The California Independent System Operator can recognize the CAISO tariff for performance settlement using comparison groups as a tariff compliant methodology for calculating energy measurement of supply-side demand response and refine its business practices in its application. The CAISO may additionally utilize the analysis in further exploration of forecasting or settlement of distributed energy resource demand reduction impacts.

- The California Energy Commission is currently hosting a staff workshop on resource adequacy qualifying capacity of supply-side demand response[5] that will inform the next resource adequacy proceeding. This baseline methodology can support the use of bids (and dispatches and performance) as inputs to any QC methodology, particularly for weather-sensitive DR resources..

- The California Public Utilities Commission is in the midst of considering how to address Summer Reliability for 2022 and 2023. This analysis may support a review of solutions adopted by the Commission to provide visibility to all agencies in understanding the performance capabilities of supply side demand response when used in managing grid operations or in considering alternatives to the Load Impact Protocol process.

Understand impacts of demand response events in 2021 and operationalize methods at scale.

While an analysis of 2021 has not been implemented at this point, Recurve's 2020 analysis shows that the comparison group methods provide a reliable means for the assessment of demand response impacts. It is functional at scale and can be implemented across climate regions and sectors.

CAISO can set the expectation that consistent, transparent, revenue-grade measurement for demand response and other flexible resources is a requirement for understanding grid impacts in the state.

Additional Recommendations

As a result of the outcomes of this report, we offer the following recommendations for the CAISO, CEC, and CPUC to operationalize comparison groups for demand response:

[5] CEC Docket 21-DR-01.

- CAISO should update their existing control group method tariff to add important detail and address minor recommendations in Appendix A.
- Cooperating state agencies should adopt the comparison group methods, approved for settlement in CAISO and operationalized with the full-stack open-source codebase, as the preferred performance measurement approach across demand response resources. This would ensure consistent, robust, and transparent measurement to underpin this important and growing resource statewide.
- The California Public Utilities Commission, in collaboration with the California Energy Commission and CAISO, should authorize secure data access to a non-participant pool for qualified vendors to allow this method to be used.

I. Introduction

As the bridge between utilities, customers, and system operators, demand response providers are a critical hub in transitioning to a modern grid that remains reliable and cost-efficient with high levels of renewable penetration. However, properly valuing demand response assets has proven difficult, and measurement approaches have been controversial.

California's emergency grid events the week of August 14, 2020, exposed the need for greater standardization, transparency, and measurement accuracy to engender shared confidence in the impact and value of demand response resources. During this week, an extreme heatwave led to supply constraints, severe price spikes, and rolling blackouts. These blackouts were the subject of national headlines[6] and a multiagency analysis, requested by the Governor, which identified underperformance of demand response as a root cause, among many others.[7]

However, that same report cataloged no fewer than four distinct frameworks, along with "several subcategories" of demand response measurement methods. It is unclear which methods were utilized to arrive at the load impact values cited or what issues may have arisen in their application. While there may be some value in the availability of multiple methods, it leads to significant challenges in comparing different resources if outcomes reflect methodological choices rather than actual differences in performance.

Some commonly employed demand response methods, including the "10 of 10" baseline,[8] are known to yield inaccurate results when extreme weather creates non-linear customer usage patterns. Adjustment factors have been permitted within these types of "X of Y" baselines but have traditionally been capped.[9] In response to the problems these methods created in 2020, the cap on the same day adjustment has been temporarily lifted. However, this change introduces other risks, including savings inflation or deflation depending on event day non-event hour usage patterns, which are observed in this study to be altered on account of the event itself.

Transparent and standardized comparison group[10] methods offer a path to improved accuracy, consistency, and reliability of the demand response measurement. In this study, we are applying open-source GRIDmeter methods, developed with funding from the US

[6] See for example, *California Expresses Frustration as Blackouts Enter 4th Day*, New York Times, Aug. 17, 2020; Rolling Blackouts in California Have Power Experts Stumped, Aug. 16, 2020

[7] Root Cause Analysis: Mid-August 2020 Extreme Heat Wave, January, 2021

[8] The "10 of 10" baseline method is commonly used when comparison groups are not possible.

[9] It is likely that these adjustment caps were too low for the types of extreme heat experienced in August, 2020 and that is becoming more common due to climate change.

[10] In this report the term "comparison group" is used in the same context as "control group" in some literature. We prefer to reserve the term "control group" specifically for randomized control trials. The adopted tariff uses the terminology "control group" but this term more closely resembles a comparison group per our definition.

Department of Energy to automate the process of implementing standardized comparison groups in the calculation of demand response event impacts.

In its recent report to the California Independent System Operator (CAISO),[11] the Baseline Accuracy Working Group (BAWG) found that "control groups consistently outperformed day and weather matching baselines." Several of the BAWG's recommendations for control group methodologies were subsequently adopted in the CAISO tariff governing demand response.[12] However, the adopted control group methods remain largely untested and, to date have not been used for settlement. Without the refinement that comes from implementation, the ISO tariff's approved control group framework still leaves a good deal of room for interpretation — and, therefore, uncertainty. Adding to these risks, data access constraints, particularly for third parties, have limited the opportunity to leverage comparison groups, which require non-participant data.

This report undertakes three important steps to support and take full advantage of the adopted control group method in the tariff with minor adjustments represented by the comparison group methods outlined in this report for demand response in California.

First, detailed comparison group methods are specified that are compliant with the ISO tariff. To enable transparency and detailed inspectability of application, the code to implement the demand response methods is largely open-sourced.

Second, the methods and code need to be thoroughly tested and refined based on application to demand response events across various cases (different demand response providers, program types, LSE territories, climate zones, sectors, etc.).

Third, modern, mathematically rigorous privacy methods are incorporated to protect the non-participant data needed for utilization of comparison groups.

This report takes on these steps to test, refine, and expand the comparison group methodologies put forward in the CAISO tariff.

II. Methodological Foundations

At the heart of the FLEXmeter are the CalTRACK 2.0[13] Hourly methods and the GRIDmeter[14] methods. These methodologies, utilized for whole building modeling and hourly comparison group load impact adjustments, were developed in open stakeholder processes and are

[11] *Baseline Accuracy Work Group Proposal*, Nexant Report to CAISO, June, 2017

[12] Section 4.13.4.3 of the CAISO tariff specifies the Control Group Methodology. More detail is provided in Section 5.3 of the Demand Response Business Practice Manual.

[13] https://www.caltrack.org/

[14] https://grid.recurve.com/

publicly available. The codebases that implement these methods are also open-source and publicly available.

To safeguard customer data, the FLEXmeter methods also utilize Energy Differential Privacy approaches. The Energy Differential Privacy[15] applications have been detailed in several publications cited below and are also supported by an open-source codebase. Differential privacy is an emerging best practice in consumer protection and has been utilized recently by the US Census and Google, among others. All of these components of the FLEXmeter are described in greater detail below, along with references to where they are fully specified in public resources.

i. CalTRACK 2.0 Hourly Methods for Baseline and Counterfactual Modeling

The CalTRACK 2.0 Hourly methods implemented through the OpenEEmeter Python codebase serve as the starting point for the generation of hourly baseline models and event-day predictions (counterfactuals) at an individual meter level. The CalTRACK hourly model is a Time-Of-Week and Temperature (TOWT) model that is normalized for weather and occupancy.[16] For each hour in the baseline dataset usage is assigned across up to seven temperature bins.[17] 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 dedicated group of subject matter experts and stakeholders worked in collaboration throughout the development process. The CalTRACK methods[18] process is described in full detail at www.caltrack.org, and a general overview of the CalTRACK hourly methods is also available in a recent article on Recurve's website. The OpenEEmeter Python codebase is publicly available under the stewardship of the Linux Foundation Energy[19] as an open-source project using a permissive Apache 2 license.[20] For optimal performance in the

[15] https://edp.recurve.com/

[16] 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.

[17] 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.

[18] 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.

[19] https://www.lfenergy.org/

[20] https://github.com/openeemeter/eemeter/blob/master/LICENSE

measurement of demand response impacts, certain CalTRACK requirements are modified slightly in ways that are discussed in detail in the following sections of this report.

ii. GRIDmeter Comparison Group Selection and Savings Adjustment Methods

The GRIDmeter methods[21] and codebase[22] are also publicly available and are designed to automate comparison group sampling based on meter data. Recurve developed the GRIDmeter methods in partnership with the United States Department of Energy (DOE) and MCE.[23] The GRIDmeter methods detail several comparison group selection procedures, including advanced stratified sampling and site-based matching, and provide recommendations on which procedure should be applied given the data available and precision needs of a measurement. 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 Beyond. 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 the GRIDmeter and CalTRACK methods and code to measure and adjust meter-based savings for numerous demand-side programs across the United States.

iii. Differential Privacy to Protect Customer Data Privacy

To enable this 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. The general data flow structure to enable this report is shown in Figure 1. Data sharing has taken place under strict third-party non-disclosure agreements. To provide confidentiality to these partners, all results are presented anonymously and aggregated as needed.

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

[22] https://github.com/recurve-methods/comparison_groups

[23] https://www.mcecleanenergy.org

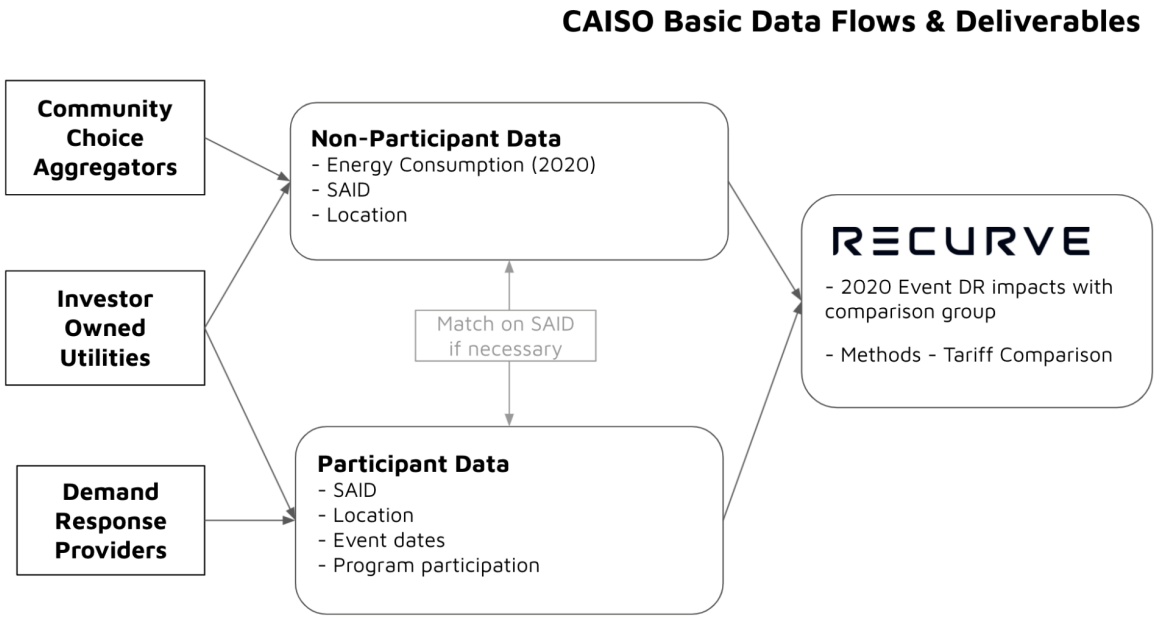


Figure 1: This report's general data flow structure

To minimize the risk of the reidentification of non-participant data, Recurve applies differential privacy[24] algorithms to usage and savings values as well as figures, utilizing methods developed by the Energy Differential Privacy (EDP) project, which is supported by the DOE to enable the secure sharing of energy data.[25] As part of the EDP Recurve has released an open-source library, eeprivacy,[26] and uses the corresponding methods throughout this work.

The EDP techniques are similar to those employed recently by Google to release social mobility data.[27] In much the same way that energy data can reveal important patterns for policymaking and social programs, the Google analysis provided insights into the efficacy of COVID-related stay-at-home orders.[28] Other notable users of differential privacy include the U.S. Census and Facebook.[29]

[24] For a summary of differential privacy for the sharing of energy data, see: *Consultation Paper: Data Access Models for Energy Data Comments prepared for Australian Competition and Consumer Commission,* Open Energy Efficiency (now Recurve) (2019) and *Differential Privacy for Expanding Access to Building Energy Data*, McGee Young (Recurve), Marc Paré (Recurve), and Harry Bergmann (DOE) ACEEE Summer Study 2020

[25] https://www.energy.gov/eere/buildings/energy-data-vault

[26] https://github.com/recurve-inc/eeprivacy

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

[28] Google's Community Mobility Reports are available at: https://www.google.com/covid19/mobility/

[29] See for example, *New privacy-protected Facebook data for independent research on social media's impact on democracy*, Facebook, February, 2020.

In an initial test of the FLEXmeter methods, Recurve applied differential privacy in the measurement of OhmConnect's load impact on Aug. 14 in MCE territory.[3] Recurve also recently applied differential privacy to securely release detailed COVID energy impacts data.[30] Other examples of differential privacy applied to energy data have been recently published.[31]

In addition to differential privacy to protect non-participant data, Recurve has taken certain precautions in this report. First, as mentioned above, all demand response providers and LSEs are anonymized. Second, all results are given in various aggregates. Third, for any event, Recurve only provides results if at least 100 residential treatment and 100 comparison group customers are the subject of a measurement and at least 50 treatment and comparison customers are the subject of a commercial statistic. Additionally, the presence of very high positive or negative consumers in a population can increase the privacy risk of a dataset because of how such customers can impact summary statistics. (Their absence or presence can be more easily detected.) Therefore, in certain cases, Recurve has adopted percentile cutoff values between 0.1% - 0.5% and 99.5 - 99.9% applied to the lowest and highest consumption data points present in an underlying dataset subject to aggregation. These boundaries are described as quantile cutoff lower "qcl" and quantile cutoff upper "qcu", respectively.

III. 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

Appendix C contains a detailed methodological specification. In this section, we walk through these steps for a single event (Aug. 19, 2020, 5 - 7 pm) for a specific participant group (DRP B, LSE 2). Using this example event we describe several important conceptual and technical considerations that will serve as the basis to interpret the results in the remainder of this report.

Baseline Specifications

The concept of a "baseline period" is important to both comparison group selection and the load impacts calculation steps. For our purposes, a "baseline period" can refer to the

[30] *Differential Privacy for Expanding Access to Building Energy Data*, McGee Young, Marc-Antoine Paré, Harry Bergmann

[31] *A Community in Crisis: Unpacking the Impacts of COVID-19 on Building Energy Consumption*, Adam Scheer, Stephen Suffian, Marc Paré, McGee Young Released: July 9, 2020 In partnership with MCE and supported by the U.S. Department of Energy through the "Secure Algorithm Testbed For Energy Data Fusion"

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. Where needed in this report, we qualify the baseline period under reference 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 CAISO tariff[6] provides some definitive baseline guidance, and Recurve has designed the FLEXmeter methods accordingly. In particular, the tariff mandates that a minimum of 20 non-event days is included in the "validation" of comparison groups. The tariff is also clear that comparison group validation must be conducted on the basis of pre-event data only (i.e. no post-event data).

Following this guidance, 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 "blacked 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 per the tariff, no post-event data are used for comparison group selection, though this is a topic that Recurve recommends revisiting since including post-event data can help ensure weather patterns are captured that best represent event day weather. The baseline period for savings calculations utilizes a 45-day pre-event and 15-day post-event window, again with event days blacked out. Additional baseline topics that are specific to the steps above are covered in the sections that follow.

1. 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"). Comparison groups can be sampled from a comparison pool in many ways. Several examples of comparison group selection methods on the basis of meter data are given in reference 21.

In the FLEXmeter methods, comparison groups are selected utilizing the GRIDmeter individual site-level matching approach. With site matching, the specific comparison pool customers that are the closest representation of an individual treatment meter are selected into the comparison group. The site matching is conducted on the basis of modeled average weekly load shape. 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. The FLEXmeter generates a time-of-week and temperature (TOWT) model using data from the 45-day baseline period.[32] In this study, the sampling period is taken as the 45-day window leading to Aug. 14, which is the first event day analyzed here. To avoid the influence of events that occurred during the baseline period, any day in which a treatment customer participated in an event is blacked out, meaning not included in the model. The model outputs an hourly prediction of usage for every meter for every hour in the baseline period, including for blacked-out event days. From these data, the 168-hour average modeled weekly load shape is generated for every meter. This process is applied identically for every treatment meter and every comparison pool meter.[33] 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.

Figure 2 shows an example of the blackout procedure for DRP A LSE 1 for a particular climate zone. The blue dots show the fraction of participating meters retained for a given hour. The red dots represent average hourly temperatures, and the lighter red dots indicate hours that are blacked out. An analogous plot for the comparison group looks very similar.

[32] The CalTRACK 2.0 Hourly methods specify that for every baseline period month an independent model be produced in which data from the month in question is fully weighted while data from the nearest neighboring months are weighted at half. Therefore the 45-day baseline period with no weighting is a modification of this standard.

[33] Despite not participating in the demand response events, the demand response blackout step is applied also to the comparison pool meters such that the days and hours used for modeling are the same between treatment and comparison meters.

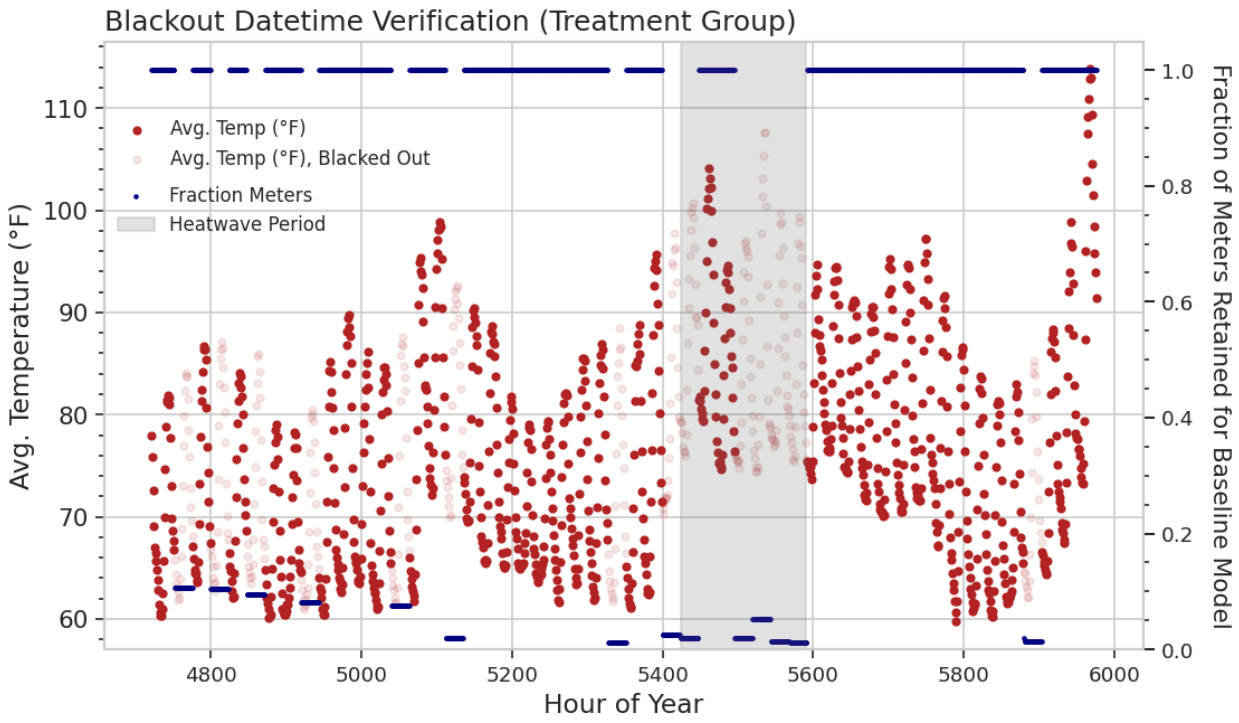


Figure 2: DRP A LSE 1: Example of the blackout procedure. The blue dots show the fraction of participating meters retained for a given hour. The red dots represent average hourly temperatures, and the lighter red dots indicate hours that are blacked out.

Before the assignment of comparison meters to treatment meters, several categorical requirements are established. First, the sector (residential or non-residential) must be shared between treatment and comparison meters. Similarly, solar PV status and climate zone must all also be shared. These sampling requirements are designed to ensure that results can be clearly assigned to particular DRP/LSE combinations, 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.

In effect, each categorical grouping is treated as a distinct treatment group in the difference of differences calculation detailed below. For example, a residential program with solar and non-solar customers spread across two climate zones would be treated as four treatment groups that are matched independently with comparison customers that share these categorical filters. (For example, non-solar, climate zone 1 would be distinct from non-solar, climate zone 2.) After comparison group adjustments are made at this categorical level, they can be aggregated to any degree desired.

With these categorical requirements in place, the sum of squares of the differences in the modeled weekly load shape for every possible combination of treatment and comparison pool meters is calculated. Each treatment meter is simultaneously assigned the comparison group meter with the lowest sum of squares value. After this step, the code removes duplicate meters sampled into the comparison group, and the comparison group meters

assigned are removed from the remaining comparison pool. In this approach, a comparison meter can only be assigned to a single treatment meter.

This process is repeated until the desired comparison group size is reached. At this point, Recurve has not been able to conduct a rigorous statistical power analysis related to sample sizing but can provide recommendations based on the stability of samples produced throughout this work.

- For participant groups between 15 to 1,000, Recurve recommends utilizing a comparison pool with at least a factor of 20 more meters than the participant group and sampling at least four comparison meters per treatment meter.
- For participant groups of 1,000 to 4,000 meters, Recurve recommends sampling from a comparison pool of at least 20,000 meters such that at least 4,000 comparison group meters are selected.
- For participant groups larger than 4,000 meters, if a comparison pool of at least 8 times the size of the participant group is available, Recurve recommends sampling at a 2:1 ratio. If a comparison group of between 4 to 8 times the size of the participant group is available, Recurve recommends sampling at a 1:1 ratio.
- For any comparison pools smaller than 4 times the size of a participant group, Recurve recommends random sampling from the treatment group such that a 4:1 ratio is obtained and then sampling at a 1:1 ratio. In these cases, savings should be scaled to the full population of participants.

When using modeling approaches, it is important to establish data sufficiency and model fit eligibility criteria as well as how to assign savings for ineligible meters. These considerations are detailed in footnote [34].

Figure 3 shows the results of the site-based meter matching for the example event. This treatment group consisted of about 1,500 customers (Table 3 below) and was matched to about 5,300 non-participants. Though the matching is conducted based on average weekly load shape, in this figure Recurve has aggregated to average daily load shape in order to

[34] To qualify for the savings calculation, both treatment and comparison group meters are required to have 85% of all possible meter readings populated in the savings calculation baseline period. Additionally, at least one meter reading must be present in all 168 hours of a week. Finally, the baseline period CalTRACK hourly Coefficient of Variation of the Root-Mean-Squared Error (CVRMSE) should be between -2 and 2. Outside these bounds customer usage can be quite erratic. For residential meters eliminated on account of these eligibility criteria, Recurve recommends assigning the average hourly savings for eligible meters taking into account which event hours the eliminated meters participated in. For non-residential meters such an approach is not recommended because of the much greater variability in total consumption and smaller participant groups. In cases where sufficient baseline period data exists, savings for ineligible meters can be assigned based on the percentage savings achieved for each hour within the eligible participant group applied to the average baseline period hourly consumption of the ineligible meter for the duration of the event called for that meter.

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.

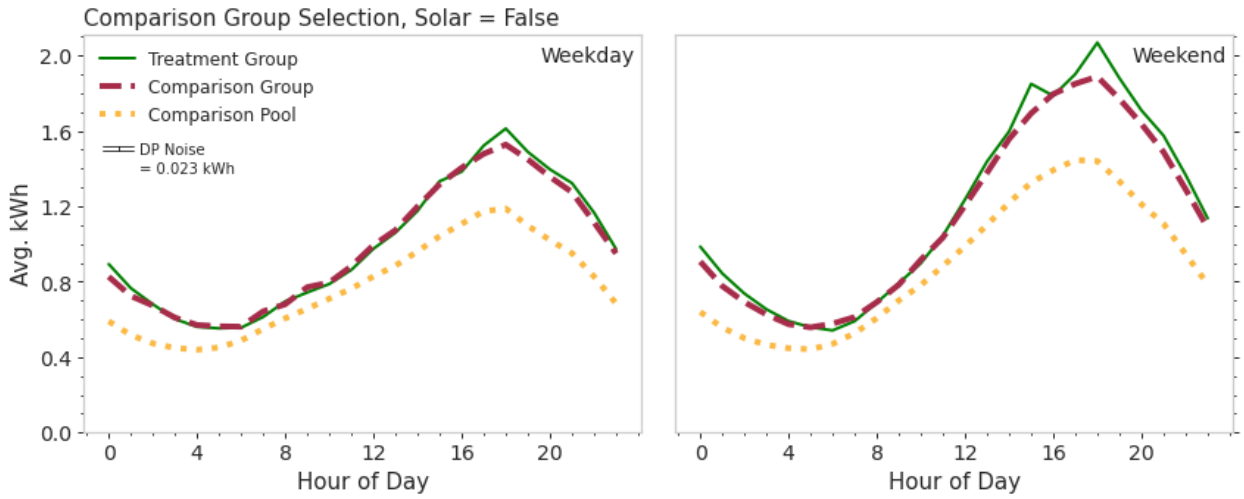


Figure 3: The average weekday (left) and weekend (right) load shape of a meter in the treatment group (solid green), comparison pool (dotted orange), and comparison group (dashed red).

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.

Importantly, this match is not the result of a fortuitous cancellation of errors but rather reflects that quality matches have been achieved across the entire spectrum of treatment customers. Figure 4 shows the average modeled daily load shape for treatment (solid green) and comparison pool (dotted orange) customers grouped into quartiles of total baseline consumption. 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.

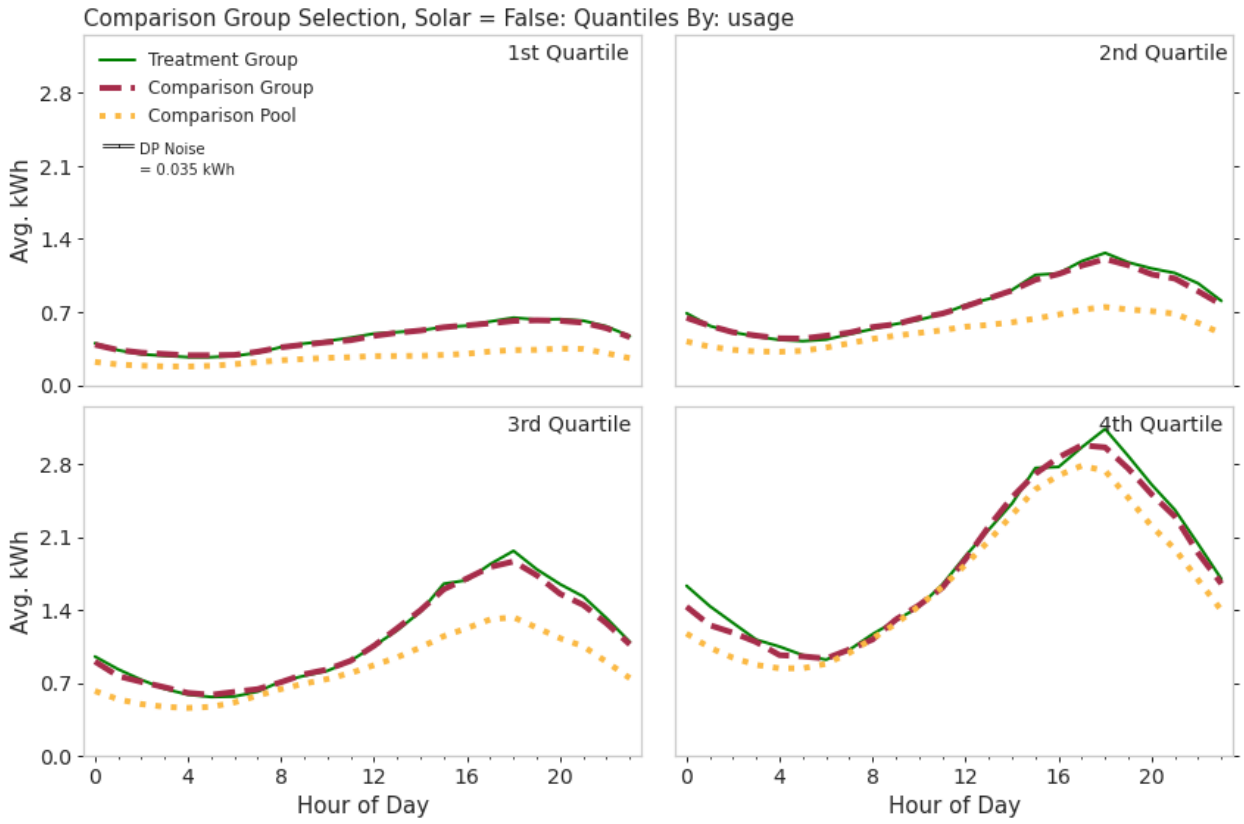


Figure 4: The average load shape of a meter in the treatment group (solid green), and comparison pool (dotted orange) broken out by baseline usage quartiles. Matched comparison group average load shapes are also shown (dashed red). The lowest quartile is in the top left panel and the highest quartile is shown in the bottom right panel.

Despite the difference in total usage showcased in Figure 4, these customer groups all have relatively homogeneous average load shapes. Of course, individual customers exhibit a wide range of different usage patterns and we can check that the comparison selection has also succeeded in matching across different relative load shapes. Figure 5 is an analogous quartile plot where the grouping has been conducted on the ratio of a customer's daily average minimum hourly consumption to average maximum hourly consumption. Each of these groups now show a different load profile (for example customers in the lowest quartile have higher minimum consumption compared to customers in the highest quartile despite the much greater peak and total usage of the latter), yet the matching has performed well for each group.

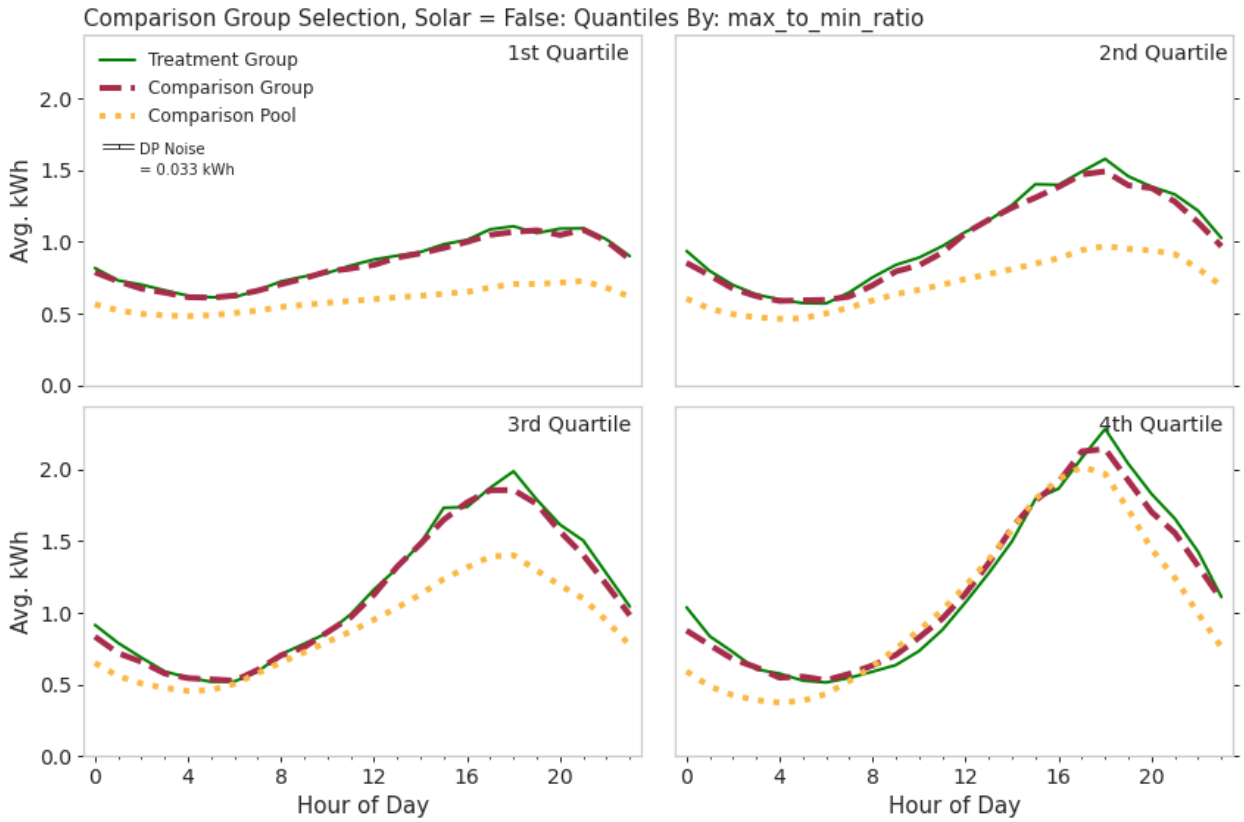


Figure 5: The average load shape of a meter in the treatment group (solid green), and comparison pool (dotted orange) broken out by average daily maximum to minimum quartiles. Matched comparison group average load shapes are also shown (dashed red). The lowest quartile is in the top left panel and the highest quartile is shown in the bottom right panel.

With this comparison group in place, we now turn attention to the components of the demand response savings calculation.

2. 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 6 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.

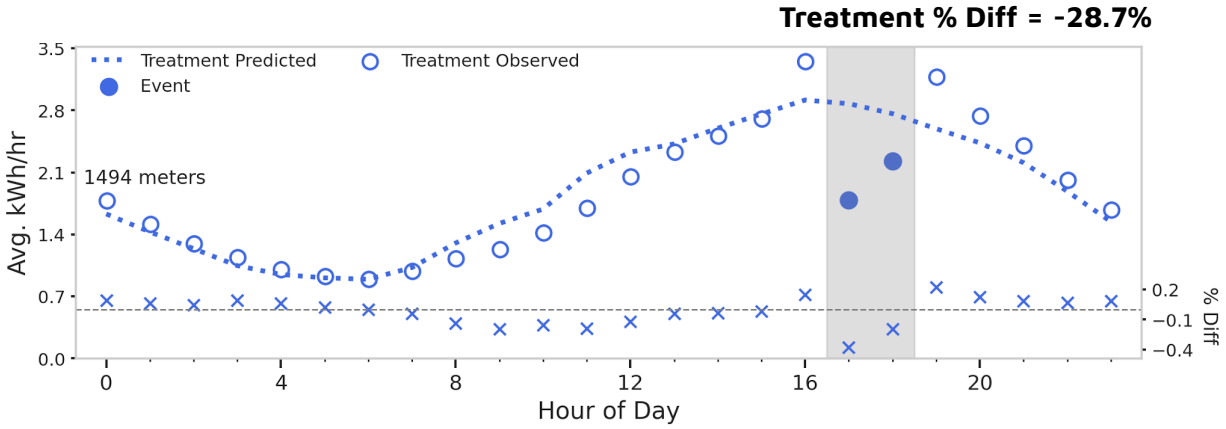


Figure 6: The dotted trace gives the average counterfactual (model-predicted) usage of a participating customer. The circles are the average observed meter readings (open = non-event hour, filled = event hour). The fractional difference between the two traces is shown as ✕'s and refer to the right-hand axis.

Finally, the ✕'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 7.

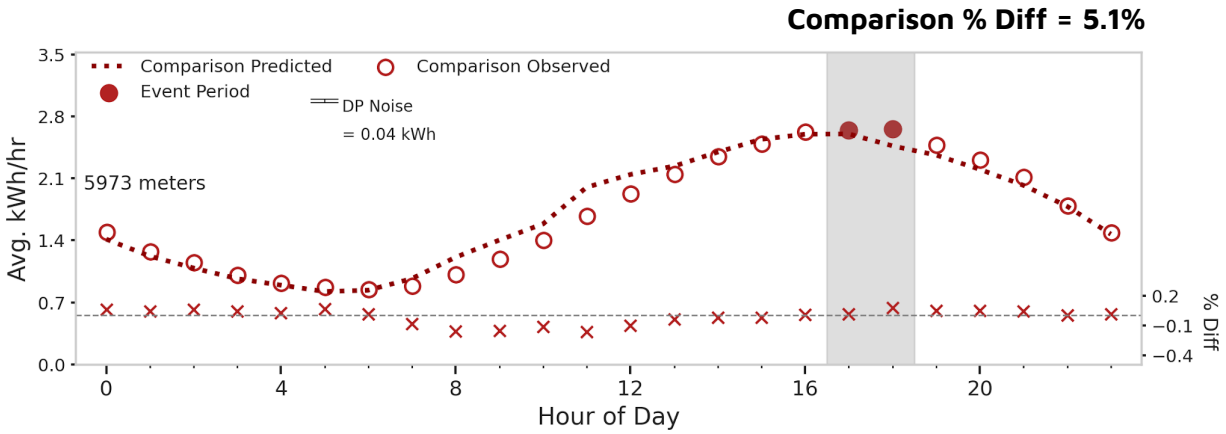


Figure 7: The dotted trace gives the average counterfactual (model-predicted) usage of a comparison group customer. The circles are the average observed meter readings (open = non-event hour, filled = event hour). The fractional difference between the two traces are given as X's and refer to the right-hand 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. Recurve 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.

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 7 red ×'s) is calculated by hour as follows:

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

Where again the subscript i refers to the hour of day.

3. Adjusted Load Impact Calculations Via % Difference of Differences

Comparison group adjustments are made to the treatment group savings via equations 3 and 4.

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

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

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. In so doing, 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 8 and 9.

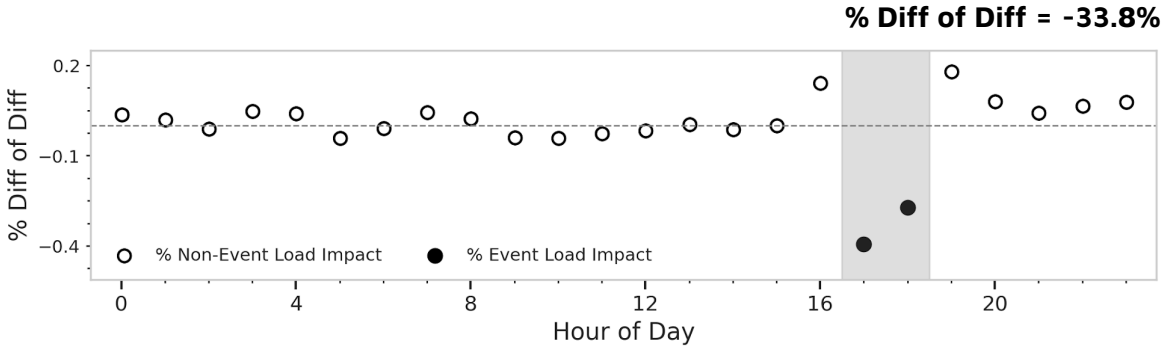


Figure 7: The fractional difference of differences (Equation 3) between treatment and comparison group. Open and filled circles represent non-event and event hours, respectively. This plot represents the comparison group adjusted fractional load impact of the participants.

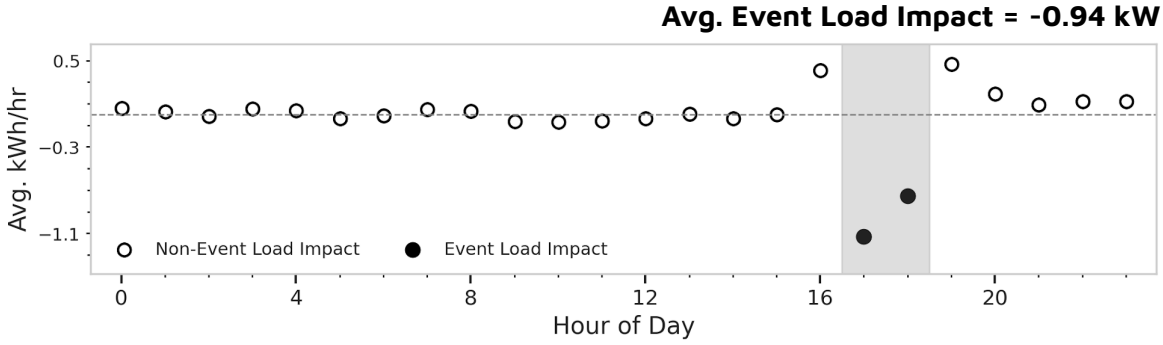


Figure 9: The final average participant comparison group adjusted hourly load impacts (Equation 4). Open and filled circles represent non-event and event hours, respectively.

The components of Figures 6 - 9 can be effectively condensed into a summary plot (Fig. 9) 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):

$$Adjusted\ Counterfactual = Counterfactual_{Treatment,i} \times Observed_{Comparison,i} / Counterfactual_{Comparison,i}$$

(5)

This adjusted counterfactual should be considered the FLEXmeter prediction of hourly consumption absent the intervention.

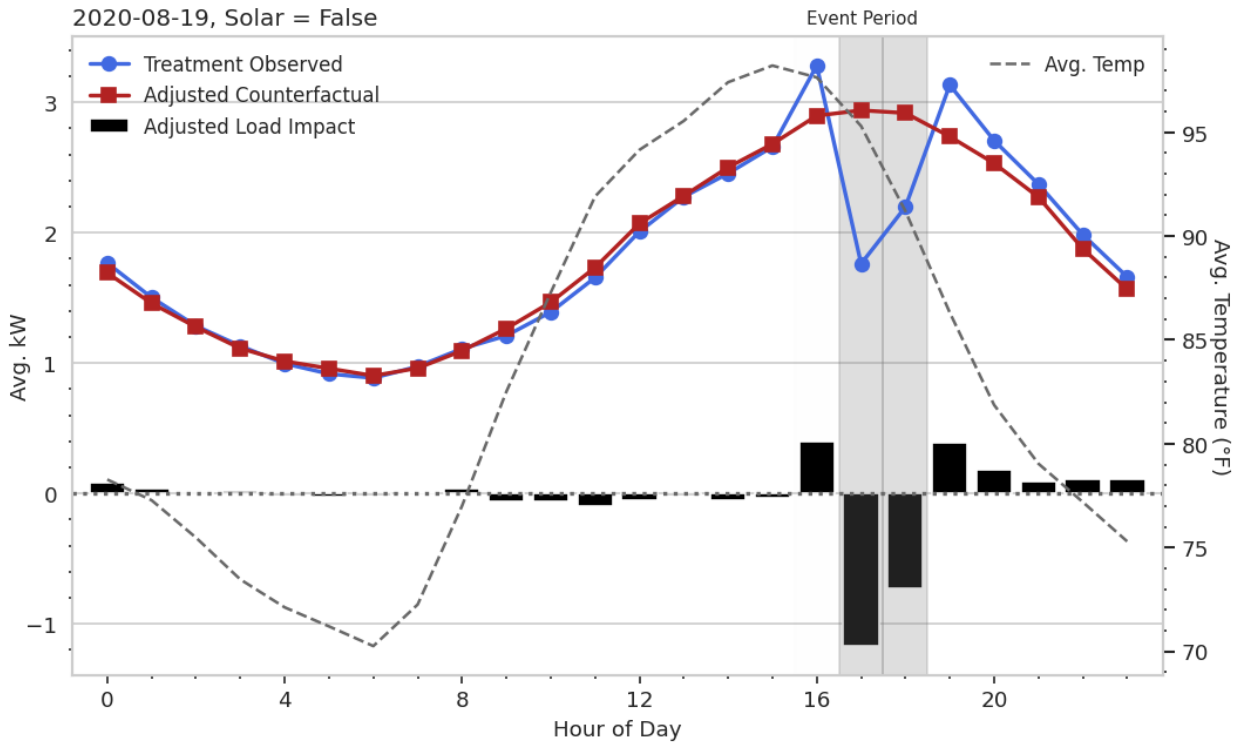


Figure 10: Summary figure comparing average participant event day usage (blue circles) and adjusted counterfactual (Eq. 5; red squares) along with hourly load impacts (black bars). The adjusted counterfactual in this figure has been privatized with ε = 10 as in Figure 7.

Also included in Figure 10 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.

It is reasonable to ask why generating models and producing counterfactuals are necessary for reliable comparative analysis. Consider that 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 we were making an assessment only 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 program. In this case, we see from the treatment and comparison counterfactuals that we *expected* a difference of about 0.17 kWh (12%). Therefore only a small portion of the difference (about 0.09 kWh) should be associated with the program. 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.

Similarly, one may ask why the computation of load impacts should be done first on a percentage (normalized) basis and then applied to counterfactual usage to produce a final adjusted load impact assignment. The answer here also relates to the fact that in not every case can a nearly perfect match between treatment and comparison groups be attained. Especially when treatment customers are very 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, and we will see some examples of this below. 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*.[21] With the normalization step, differences in the relative sizes of treatment and comparison customers do not manifest directly as under- or over-counted savings.

The entire load impact calculation can be summarized succinctly:

The event day usage of participating demand response customers is compared to a modeled prediction of these customers' usage in the absence of the event. An identical calculation is conducted for a group of non-participating customers of similar types and usage patterns. The hourly load impacts attributed to the event are taken as the load impacts calculated for the participating customers adjusted for any impacts seen in the non-participant sample.

This calculation is done on a relative basis in order to minimize error on account of imperfect matching between participant and non-participant groups. So if the treatment group achieves an average event period load reduction of 20% and the comparison group experiences an event period load increase of 5%, the total percentage load impact calculated for the event will be -20% - 5% = -25%. If the treatment customers were predicted to have used 1 MWh during the event, the total load impact is calculated to be -0.25 x 1 MWh = -250 kWh.

4. Application of Differential Privacy

With a privatized dataset, the maximum degree to which an attacker can gain knowledge about whether an individual's data was even included is determined by a parameter, ε (epsilon). Epsilon effectively sets a privacy budget. At the point the budget is stretched too far, an attacker could theoretically determine whether or not an individual building's data

was included. By reducing the threshold ε, the data is made safer, at the expense of introducing greater noise into reported statistics.[35] No such privacy guarantees are offered by traditional aggregation methods.[36]

With differential privacy, data are often represented along with an indication of the maximum noise added to a value (at the 95% confidence interval). Datasets with a very large set of fairly homogeneous individuals can be properly protected with relatively little noise added to aggregated statistics, while smaller datasets with high outliers need significantly greater noise added to attain the same level of privacy protection. Additionally, statistics with multiple data points require the injection of more noise at any given point than does the reporting of a single value. For instance, to release an average 8,760 hourly annual load shape for a population will require adding more noise at each point than would a simple single average annual total consumption value at a given level of protection.

The application of differential privacy is evident in figures where bars indicate the maximum degree of possible noise added to a value and in summary statistics, which are also listed with the maximum possible noise at the 95% confidence level.[37] It should be noted that at this time, specific best practices for the application of differential privacy to energy data are not fully established. In this report, we seek to balance strong privacy protection while not losing the utility of the data in an industry where increased uncertainty can equate to decreased value. Finally, since results are provided in a static report and not a queriable dataset of individuals' consumption, we have full confidence in the protections against any risk of re-identification. Therefore, we generally apply differential privacy in such a way that noise introduces 5% or less uncertainty into the final statistics.

In Figure 7 Recurve used an ε value of 10. As a result, noise was added such that each observed and counterfactual data point was modified within a range of +/- 0.04 kWh away from its true mean. This range is indicated by the "DP Noise" annotation in the figure. The load impact results of Figures 8 and 9 are also affected accordingly. Figure 11 shows the impact of lowering ε by an order of magnitude to 1.0. The corresponding noise increases by a factor of 10, and this effect is immediately clear in the more "choppy" traces. For a settlement-quality computation, this degree of noise would likely not be acceptable. Even

[35] When publishing multiple pieces of information figured from an underlying dataset, the total privacy impact can be determined with Basic Composition Theorem, adding up the ε for each output statistic. In comparison group sampling amplification of the privacy budget may be achieved by the "Secrecy of the Sample." In short, if the population under study is a sample of a larger population, and it can be expected that the sample used for analysis will remain secret, then the privacy impact can be reduced by the fraction of the total population that this sample represents. While this sampling scenario is often the case in this work, we do not adopt this practice here.

[36] In a 15/100 aggregation approach, an aggregation of buildings must be composed of at least 100 buildings. No building can contribute to more than 15% of the total consumption. With these conditions met the aggregation is considered publishable and the data safe. This is just one example of an aggregation threshold. Jurisdictions vary widely on the applicable aggregation requirements.

[37] This practice is known as the global model of differential privacy, in which noise to the final statistic value before output.

when optimizing the privacy factor based on a minimum usability, adding some noise still provides significant privacy enhancement beyond simple aggregation.

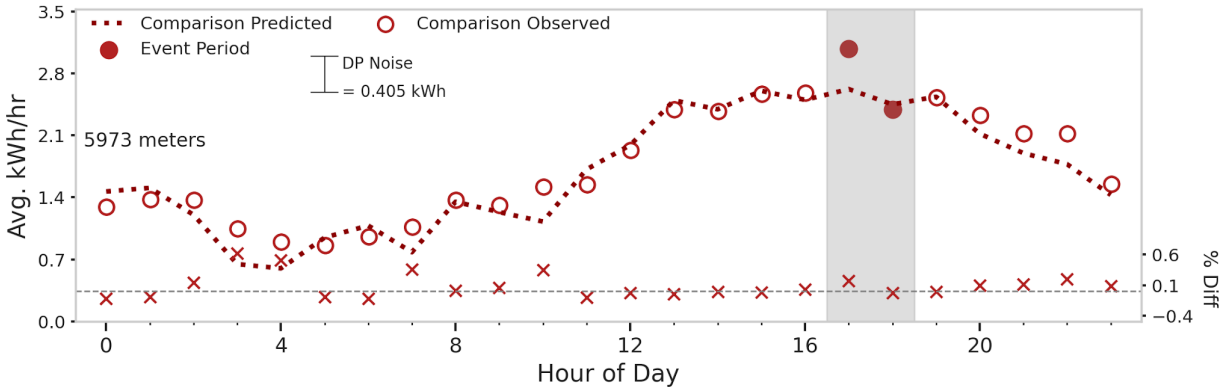


Figure 11: Regeneration of Figure 7 with ε = 1. The reduction of ε by a factor of 10 leads to greater noise in the corresponding traces.

IV. Summary Results

The figures and tables that follow summarize the FLEXmeter measurements across the August 13 - 20, 2020 events. The program design differs across these demand response providers, but the consistently applied methods offer a universal measure of impact no matter the program design.

It is important to reiterate that this analysis represents a comparison group performance assessment based on the CAISO Tariff. It is not structured as a qualifying capacity methodology that complies with the CPUC Load Impact Protocols.[38] While the methods fully comport with requirements for "ex-post analysis" in the Load Impact Protocols, the **results** presented in this report have a few important differences:

1. An open-source codebase is available to replicate the methods (and reproduce the results if one has access to the same data).

2. The breadth of this analysis for any given DRP in this study is limited to the geographic locations where non-participant pool data were also available. We did not have matches for every participant for every DRP, and the participant pool was primarily available in cooler coastal climate zones.

3. The analysis was limited to the heatwave events in 2020; it did not capture every event in the year.

[38] The CEC (in Docket 21-DR-01) is currently considering updates to the Resource Adequacy Qualifying Capacity of Supply-Side Demand Response in an industry working group forum. The methods presented in this report may also be considered as part of that process.

4. This analysis does not make an ex-ante forecast of performance and does not account for changes in participant enrollment and composition and future weather (extreme or otherwise).

Tables are broken out by different demand response providers and different combinations of sector and solar status among the customer bases. Rows displaying event results are shaded in blue and tagged as "TRUE." Average non-event results are also given in the non-shaded rows and tagged as "FALSE." All savings values have noise added as per the differential privacy protocols described above with an ε of 2.0. The maximum degree of noise is indicated though in practice, a greater degree of uncertainty is introduced on account of limited significant figures. The percent savings columns and comparison group savings adjustment columns also reflect differential privacy protection. The associated maximum noise values have been omitted to keep the tables readable. These summary tables designate the demand response providers with capital letters and the LSEs with numbers. So A3 would represent demand response provider 'A' in load-serving entity '3' service territory.

For each combination of demand response provider, LSE, sector, and solar status, the figures show average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. The fractional savings (savings/counterfactual usage) are shown in orange dots and refer to the right-hand axis.

Additional tables are available that provide a breakdown of demand response savings for each event day by specific temporal categories:

- Event +/- > 2 Hours: All hours of the day more than 2 hours away from an event
- Event +/- 2 Hours: Hours 2 hours away from the event
- Event +/- 1 Hour: Hours 1 hour away from the event
- Event Hour 1
- Event Hour 2
- Event Hour 3
- ...

These tables are too large to reasonably include here but can be made available online.

Table 1: FLEXmeter Results for Demand Response Provider A, Residential, Non-Solar

DRP	LSE	Date	Event	Avg. Temp (F)	Participants	Avg. Hourly Savings (kW)	Max. DP Noise (kW)	% Savings	Comp Group Savings Adjustment (kW)
A	1	2020-08-14	FALSE	85.6	4,934	0.02	0.0003	1.4%	-0.032
A	1	2020-08-14	TRUE	94.7	4,933	0.29	0.0017	12.1%	-0.005
A	1	2020-08-17	FALSE	84.9	4,940	0.03	0.0003	2.1%	0.064
A	1	2020-08-17	TRUE	92.6	4,938	0.31	0.0020	12.7%	0.063
A	1	2020-08-18	FALSE	86.6	4,803	0.01	0.0004	0.9%	0.030
A	1	2020-08-18	TRUE	91.1	4,797	0.28	0.0015	11.9%	0.045
A	1	2020-08-19	FALSE	86.9	4,971	0.00	0.0003	-0.3%	0.072
A	1	2020-08-19	TRUE	89.8	4,969	0.20	0.0026	8.5%	0.102
A	1	2020-08-20	FALSE	85.1	4,967	-0.06	0.0003	-4.2%	0.082
A	1	2020-08-20	TRUE	87.5	4,960	0.13	0.0028	6.1%	0.192
A	2	2020-08-14	FALSE	84.4	3,358	0.00	0.0005	0.2%	-0.084
A	2	2020-08-14	TRUE	95.7	3,355	0.24	0.0024	12.3%	-0.063
A	2	2020-08-19	FALSE	75.8	112	0.04	0.0089	2.9%	0.028
A	2	2020-08-19	TRUE	85.5	112	0.17	0.0709	8.4%	0.043
A	3	2020-08-14	FALSE	80.1	4,444	-0.02	0.0003	-1.9%	-0.081
A	3	2020-08-14	TRUE	86.0	4,443	0.14	0.0019	7.9%	-0.056
A	3	2020-08-17	FALSE	78.3	4,501	-0.01	0.0003	-1.2%	0.018
A	3	2020-08-17	TRUE	83.0	4,501	0.16	0.0016	9.1%	0.037
A	3	2020-08-18	FALSE	80.5	4,384	-0.02	0.0004	-1.3%	0.015
A	3	2020-08-18	TRUE	84.2	4,379	0.14	0.0016	7.6%	0.204
A	3	2020-08-19	FALSE	81.7	4,544	-0.01	0.0004	-0.6%	0.081
A	3	2020-08-19	TRUE	82.3	4,541	0.16	0.0025	9.0%	0.199
A	3	2020-08-20	FALSE	80.2	4,536	-0.02	0.0002	-1.8%	0.105
A	3	2020-08-20	TRUE	79.9	4,525	0.05	0.0023	3.1%	0.217

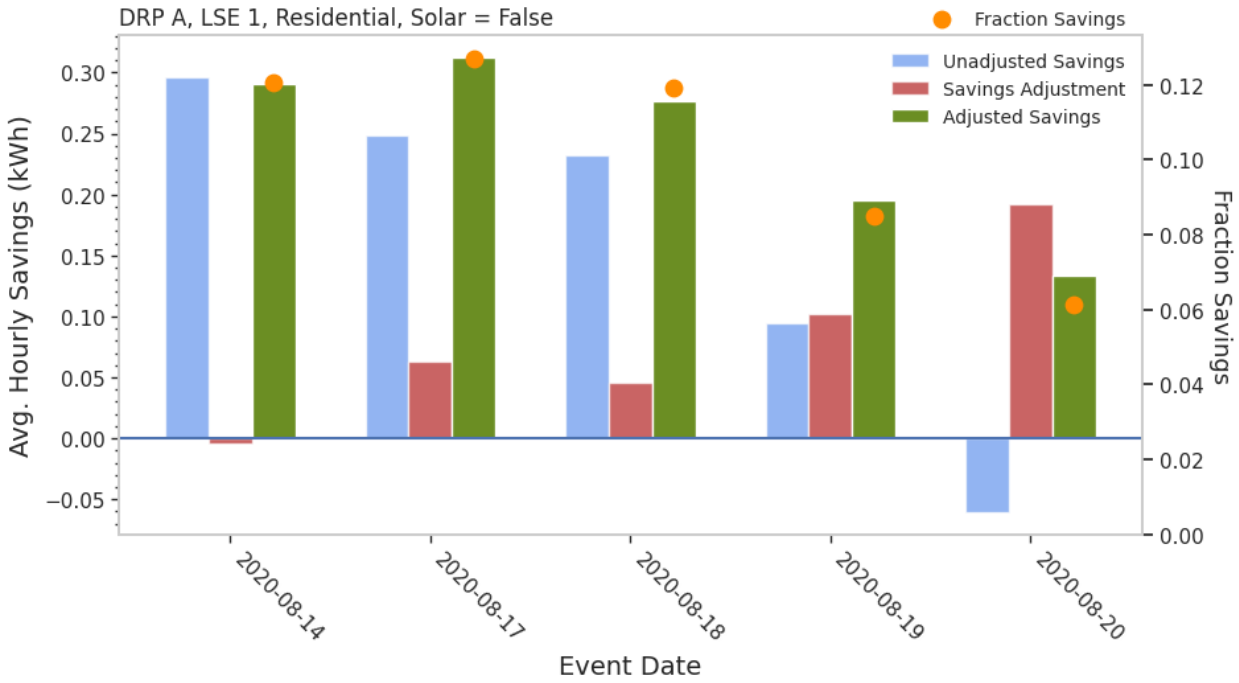


Figure 12: DRP A, LSE 1, Non-Solar. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown as orange dots and refer to the right-hand axis.

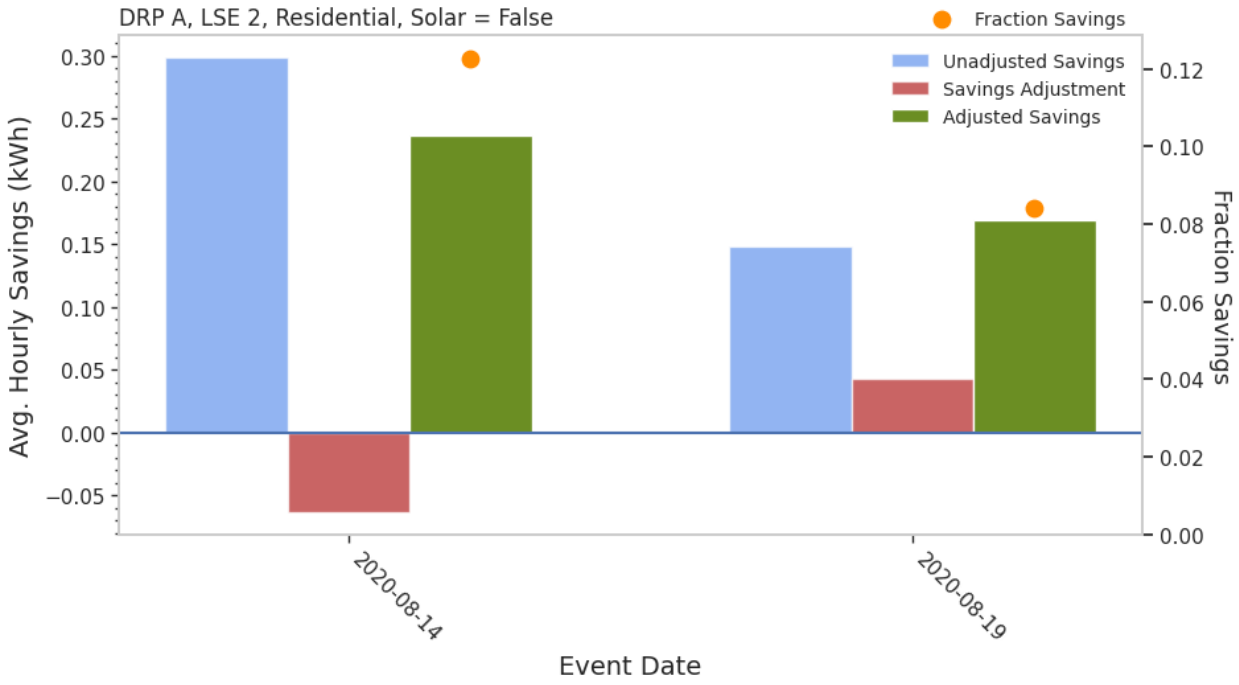


Figure 13: DRP A, LSE 2, Non-Solar. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown in orange dots and refer to the right-hand axis.

Figure 14: DRP A, LSE 3 Non-Solar. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown in orange dots and refer to the right hand axis.

Table 2: FLEXmeter Results for Demand Response Provider A, Residential, Solar

DRP	LSE	Date	Event	Avg. Temp (F)	Participants	Avg. Hourly Savings (kW)	Max. DP Noise (kW)	% Savings	Comp Group Savings Adjustment (kW)
A	2	2020-08-14	FALSE	84.9	106	0.02	0.0132	1.9%	-0.129
A	2	2020-08-14	TRUE	96.8	106	0.60	0.0731	17.1%	-0.166
A	3	2020-08-14	FALSE	82.1	223	0.02	0.0060	1.6%	-0.145
A	3	2020-08-14	TRUE	89.7	223	0.30	0.0334	9.8%	-0.203
A	3	2020-08-17	FALSE	80.6	231	-0.01	0.0048	-0.8%	0.088
A	3	2020-08-17	TRUE	86.2	231	0.58	0.0341	18.7%	0.109
A	3	2020-08-18	FALSE	82.2	221	0.11	0.0072	6.6%	0.093
A	3	2020-08-18	TRUE	88.9	221	0.51	0.0308	15.6%	0.256
A	3	2020-08-19	FALSE	83.3	234	-0.02	0.0075	-1.1%	0.104
A	3	2020-08-19	TRUE	84.8	234	0.46	0.0536	13.4%	0.322
A	3	2020-08-20	FALSE	81.7	236	-0.07	0.0044	-4.7%	0.127
A	3	2020-08-20	TRUE	82.4	236	0.21	0.0401	6.7%	0.377

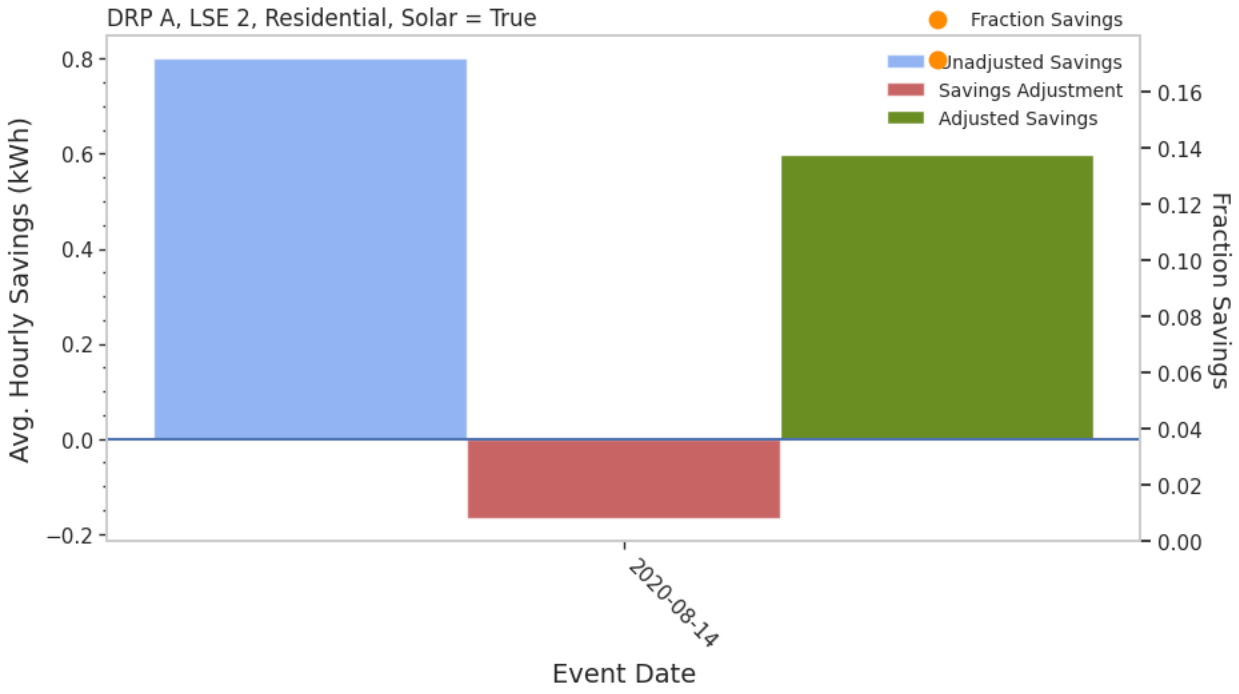


Figure 15: DRP A, LSE 2 Solar. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown in orange dots and refer to the right hand axis.

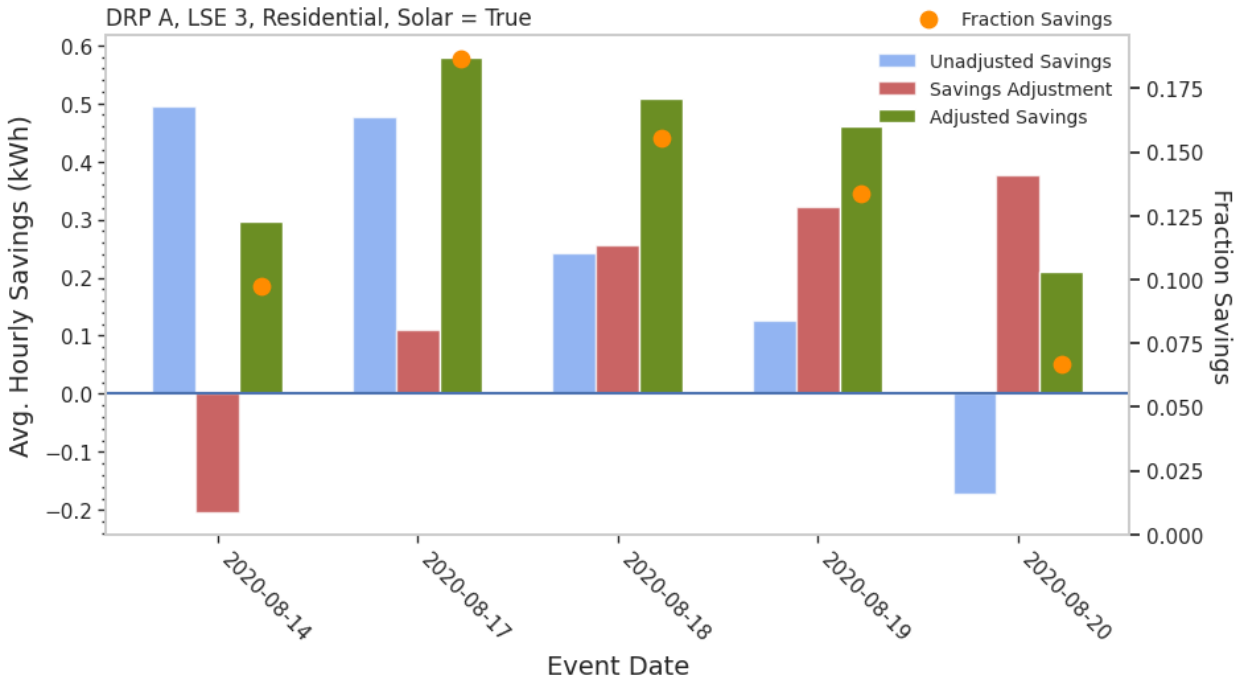


Figure 16: DRP A, LSE 3 Solar. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown in orange dots and refer to the right hand axis.

Table 3: FLEXmeter Results for Demand Response Provider B, Residential, Non-Solar

DRP	LSE	Date	Event	Avg. Temp (F)	Participants	Avg. Hourly Savings (kW)	Max. DP Noise (kW)	% Savings	Comp Group Savings Adjustment (kW)
B	2	2020-08-14	FALSE	84.1	1,507	-0.04	0.0010	-2.3%	-0.087
B	2	2020-08-14	TRUE	98.9	1,507	0.74	0.0044	23.0%	-0.136
B	2	2020-08-16	FALSE	84.7	1,506	-0.04	0.0012	-1.7%	0.081
B	2	2020-08-16	TRUE	98.7	1,506	1.07	0.0122	31.2%	0.337
B	2	2020-08-17	FALSE	72.6	29	-0.03	0.0178	-5.1%	-0.009
B	2	2020-08-17	TRUE	72.6	29	0.15	0.0842	17.3%	0.003
B	2	2020-08-18	FALSE	83.4	1,530	-0.04	0.0008	-2.3%	0.032
B	2	2020-08-18	TRUE	96.7	1,529	0.91	0.0082	28.9%	-0.040
B	2	2020-08-19	FALSE	82.3	1,529	-0.05	0.0008	-2.7%	-0.023
B	2	2020-08-19	TRUE	93.2	1,529	0.94	0.0093	32.3%	0.146
B	4	2020-08-14	FALSE	83.5	1,235	-0.01	0.0011	-0.4%	-0.096
B	4	2020-08-14	TRUE	95.3	1,235	0.72	0.0066	24.3%	-0.170
B	4	2020-08-16	FALSE	83.6	1,231	-0.04	0.0013	-1.8%	0.056
B	4	2020-08-16	TRUE	91.3	1,231	0.55	0.0132	18.8%	0.271
B	4	2020-08-18	FALSE	80.0	1,223	-0.02	0.0010	-1.1%	0.034
B	4	2020-08-18	TRUE	91.5	1,222	0.90	0.0089	31.9%	0.036
B	4	2020-08-19	FALSE	77.4	1,222	-0.02	0.0012	-1.3%	0.079
B	4	2020-08-19	TRUE	84.6	1,222	0.66	0.0082	30.3%	-0.004

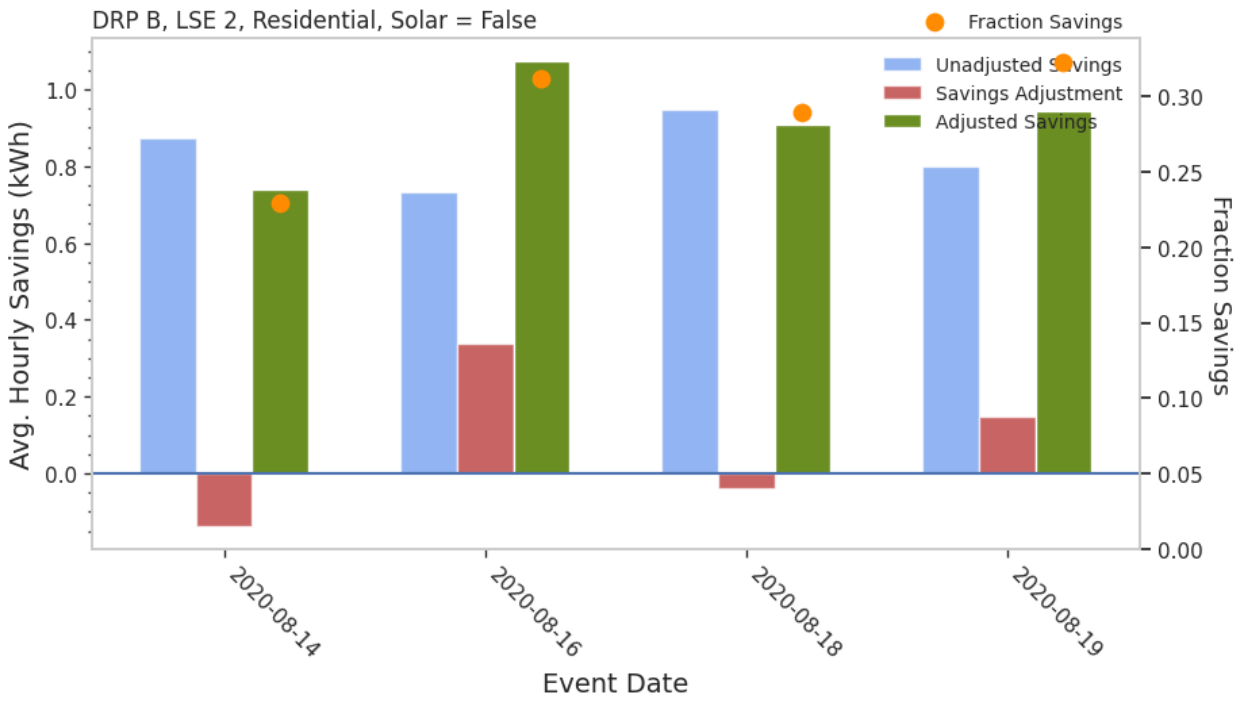


Figure 17: DRP B, LSE 2 Non-Solar. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown in orange dots and refer to the right hand axis.

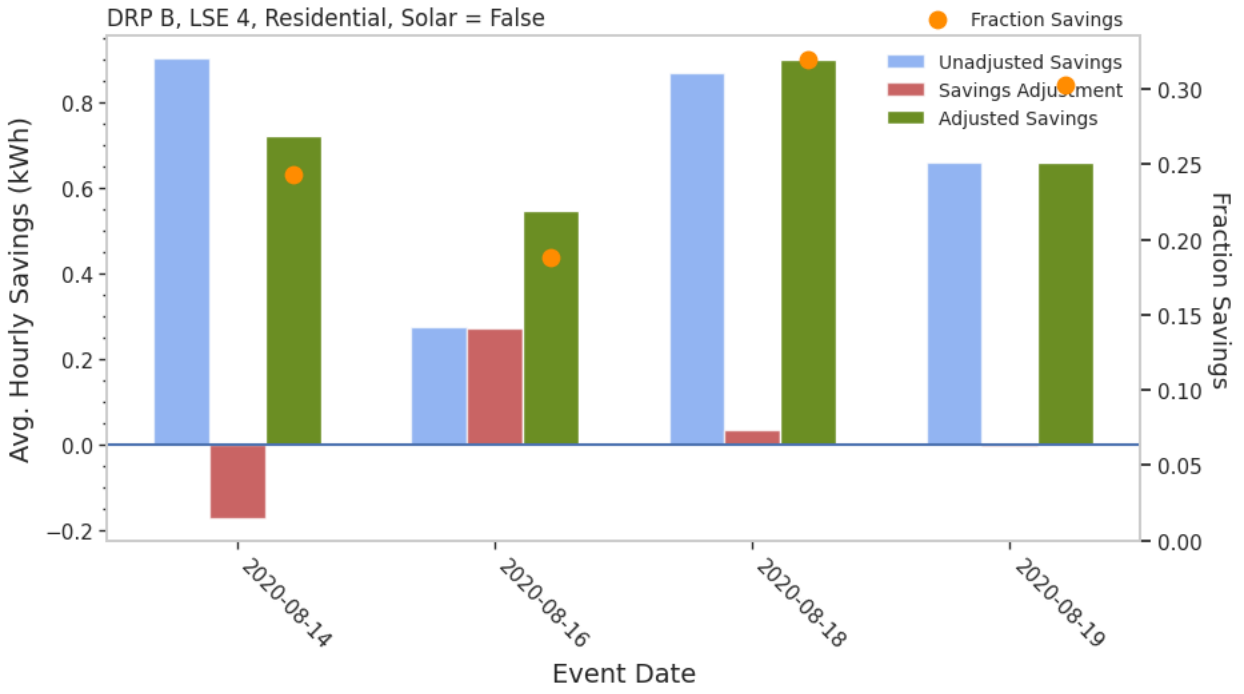


Figure 18: DRP B, LSE 4 Non-Solar. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown in orange dots and refer to the right hand axis.

Table 4: FLEXmeter Results for Demand Response Provider C, Non-Residential, Non-Solar

DRP	LSE	Date	Event	Avg. Temp (F)	Participants	Avg. Hourly Savings (kW)	Max. DP Noise (kW)	% Savings	Comp Group Savings Adjustment (kW)
C	1	2020-08-13	FALSE	84.6	86	0.65	0.7487	0.5%	-5.513
C	1	2020-08-13	TRUE	93.0	86	50.62	9.6646	41.8%	-10.022
C	1	2020-08-14	FALSE	84.1	137	0.93	0.4390	0.7%	-5.037
C	1	2020-08-14	TRUE	92.2	137	24.54	1.7418	18.6%	-10.822
C	1	2020-08-17	FALSE	82.3	137	2.77	1.0038	2.1%	-0.380
C	1	2020-08-17	TRUE	89.7	137	42.32	2.1759	31.1%	-9.063
C	1	2020-08-18	FALSE	84.8	137	-0.45	0.8690	-0.3%	-0.242
C	1	2020-08-18	TRUE	90.4	137	35.43	1.9359	26.3%	-5.774
C	1	2020-08-19	FALSE	85.5	137	2.78	0.6192	2.1%	-0.132
C	1	2020-08-19	TRUE	90.3	137	43.57	3.5925	33.1%	-3.167

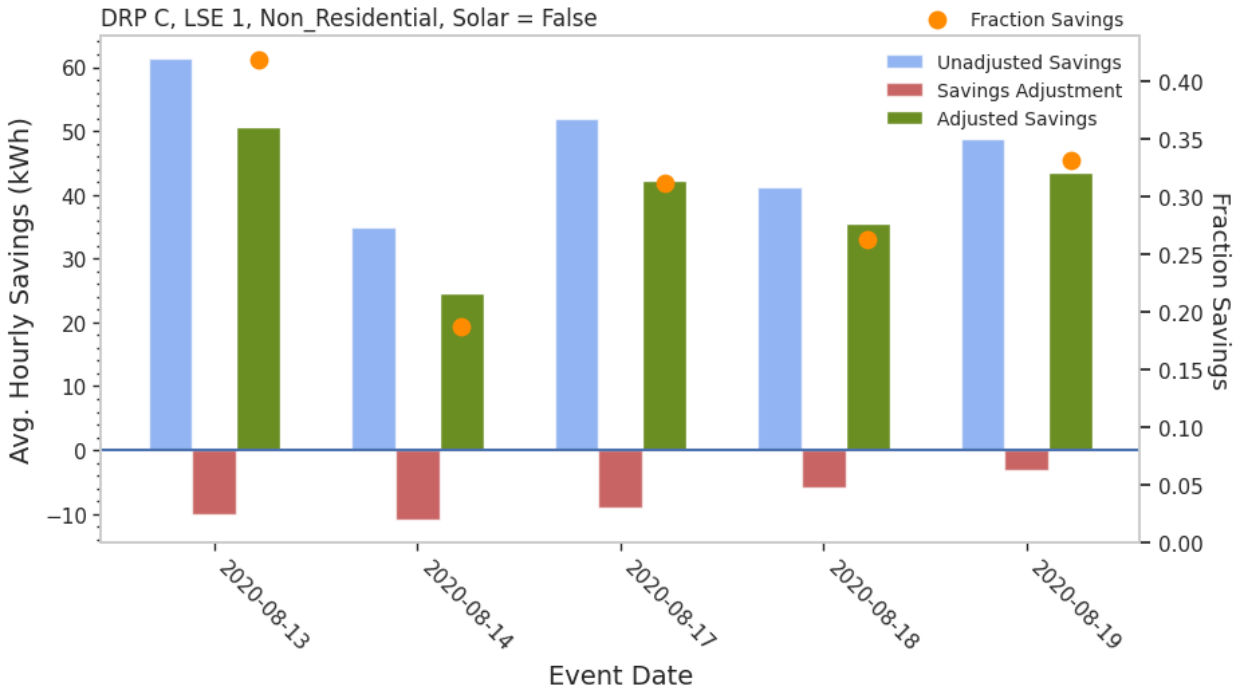


Figure 19: DRP C, LSE 1 Non-Solar. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown in orange dots and refer to the right hand axis.

Table 5: FLEXmeter Results for Demand Response Provider D, Residential, Non-Solar

DRP	LSE	Date	Event	Avg. Temp (F)	Participants	Avg. Hourly Savings (kW)	Max. DP Noise (kW)	% Savings	Comp Group Savings Adjustment (kW)
D	1	2020-08-13	FALSE	83.6	5,077	0.06	0.0002	4.1%	-0.118
D	1	2020-08-13	TRUE	88.0	5,076	0.28	0.0015	11.5%	-0.133
D	1	2020-08-14	FALSE	84.9	5,051	-0.01	0.0003	-0.5%	-0.037
D	1	2020-08-14	TRUE	93.4	5,050	0.81	0.0011	26.9%	0.038
D	1	2020-08-15	FALSE	85.5	5,070	-0.09	0.0003	-5.0%	0.051
D	1	2020-08-15	TRUE	96.0	5,070	1.06	0.0015	32.8%	0.083
D	1	2020-08-16	FALSE	85.1	5,068	-0.08	0.0002	-4.1%	0.120
D	1	2020-08-16	TRUE	90.0	5,067	0.44	0.0016	15.4%	0.192
D	1	2020-08-17	FALSE	83.8	5,072	-0.03	0.0002	-2.0%	0.062
D	1	2020-08-17	TRUE	94.7	5,071	0.91	0.0011	29.5%	0.101
D	1	2020-08-18	FALSE	83.7	5,069	-0.09	0.0003	-5.2%	0.027
D	1	2020-08-18	TRUE	97.5	5,069	0.85	0.0010	27.1%	0.017

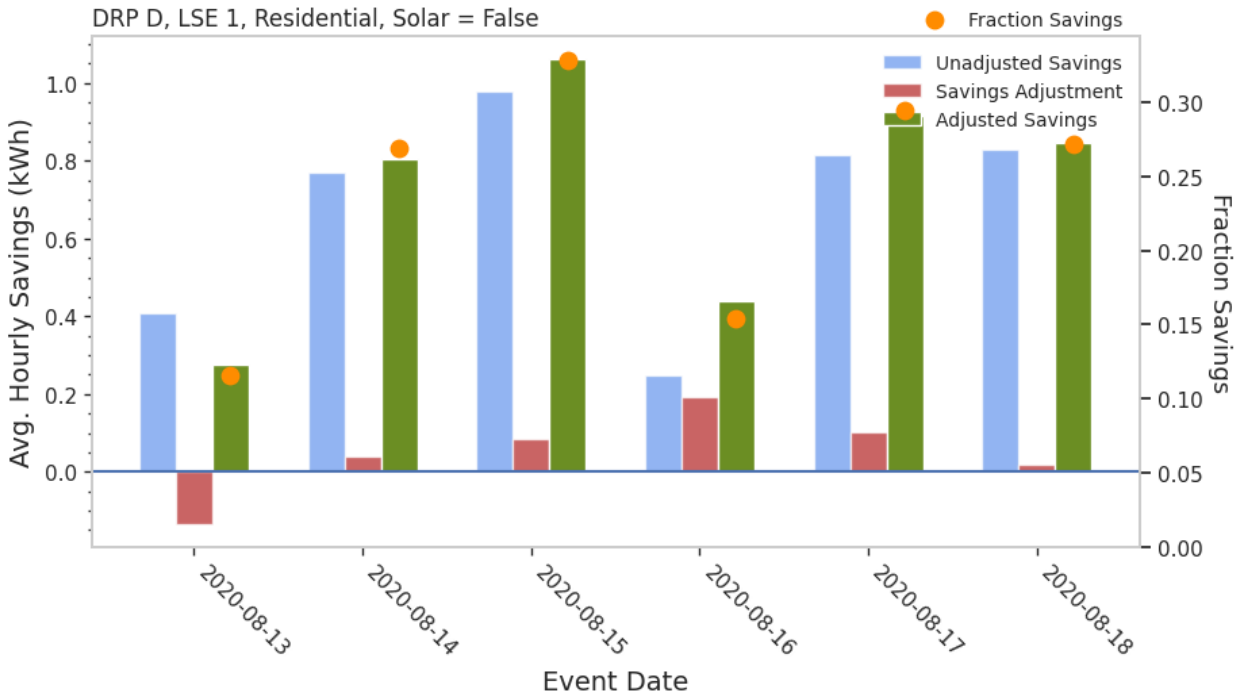


Figure 20: DRP D, LSE 1, Non-Solar, Residential. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown in orange dots and refer to the right hand axis.

Table 6: FLEXmeter Results for Demand Response Provider D, Non-Residential, Non-Solar

DRP	LSE	Date	Event	Avg. Temp (F)	Participants	Avg. Hourly Savings (kW)	Max. DP Noise (kW)	% Savings	Comp Group Savings Adjustment (kW)
D	1	2020-08-13	FALSE	81.7	2,748	-0.66	0.0147	-3.8%	-0.967
D	1	2020-08-13	TRUE	84.3	2,746	0.12	0.0433	0.8%	-1.291
D	1	2020-08-14	FALSE	83.3	2,747	-0.97	0.0187	-5.4%	-0.573
D	1	2020-08-14	TRUE	90.3	2,747	1.17	0.0435	6.7%	-1.043
D	1	2020-08-15	FALSE	83.9	2,747	0.10	0.0103	0.7%	-0.365
D	1	2020-08-15	TRUE	93.6	2,747	1.68	0.0356	10.1%	-0.516
D	1	2020-08-16	FALSE	83.5	2,748	0.01	0.0082	0.1%	-0.105
D	1	2020-08-16	TRUE	87.6	2,746	0.42	0.0308	3.0%	0.108
D	1	2020-08-17	FALSE	82.1	2,755	-0.70	0.0240	-3.7%	0.161
D	1	2020-08-17	TRUE	92.5	2,755	1.28	0.0673	6.3%	-0.383
D	1	2020-08-18	FALSE	81.9	2,754	-0.20	0.0153	-1.1%	0.052
D	1	2020-08-18	TRUE	95.4	2,754	1.97	0.0358	8.1%	-0.497

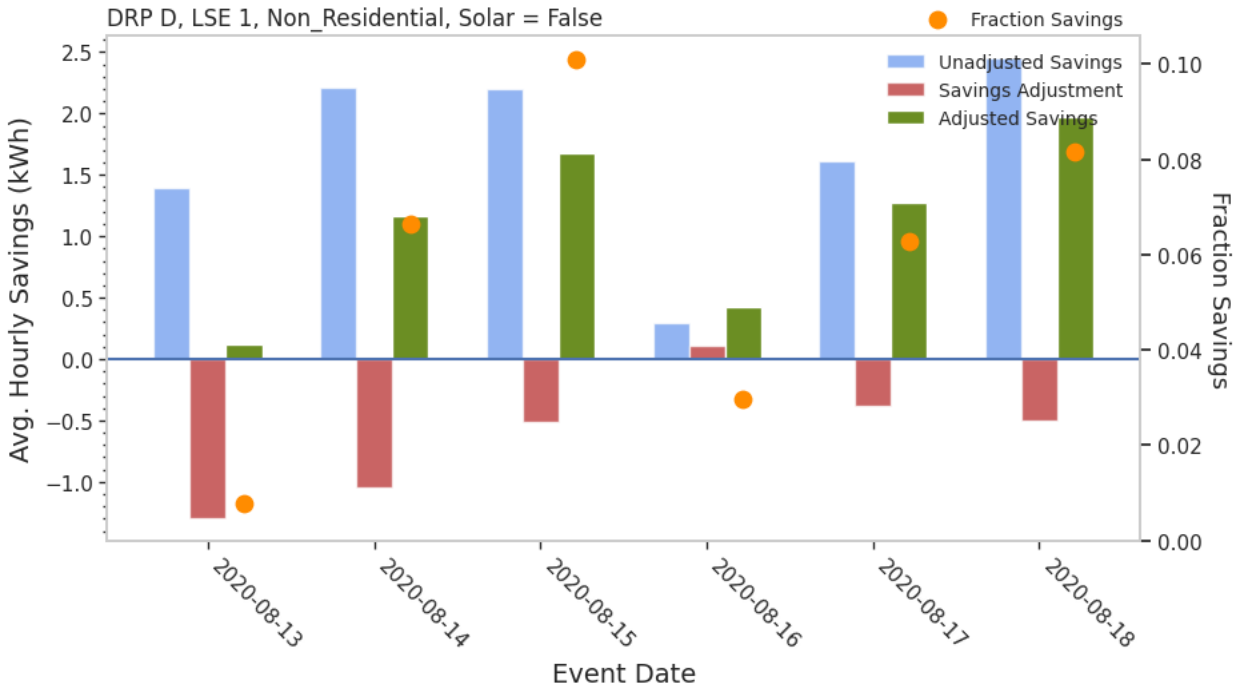


Figure 21: DRP D, LSE 1, Non-Solar, Non-Residential. Average participant hourly unadjusted savings (blue bars), the comparison group adjustment (red bars), and the adjusted savings (green bars) for each event. Fractional savings (savings/counterfactual) are shown in orange dots and refer to the right hand axis.

V. Case Studies

A more comprehensive set of results and figures are provided separately in the Extended Results Appendix. In this section, we walk through key results for a couple important cases. First, we assess a residential program with a large number of solar and non-solar participants (A3). We then cover a commercial program with a smaller number of high-consumption participants.

i. Residential: Group A3 Solar and Non-Solar

For this case study, we focus on A3 (DRP A in LSE territory 3 - see Tables 1 and 2). Group A3 consisted of both solar and non-solar customers. Figures 22 and 23 show the results of comparison group matching. As described earlier, solar PV status is one of the categorical factors that must match between a treatment customer and the assigned comparison customers in the FLEXmeter methods. For both non-solar and solar groups, the comparison group matching step produces a clear improvement in the load shape similarity between treatment and comparison samples. The resulting average comparison group load shapes are a close match to the respective treatment group load shapes for both weekday (left) and weekend (right) periods.

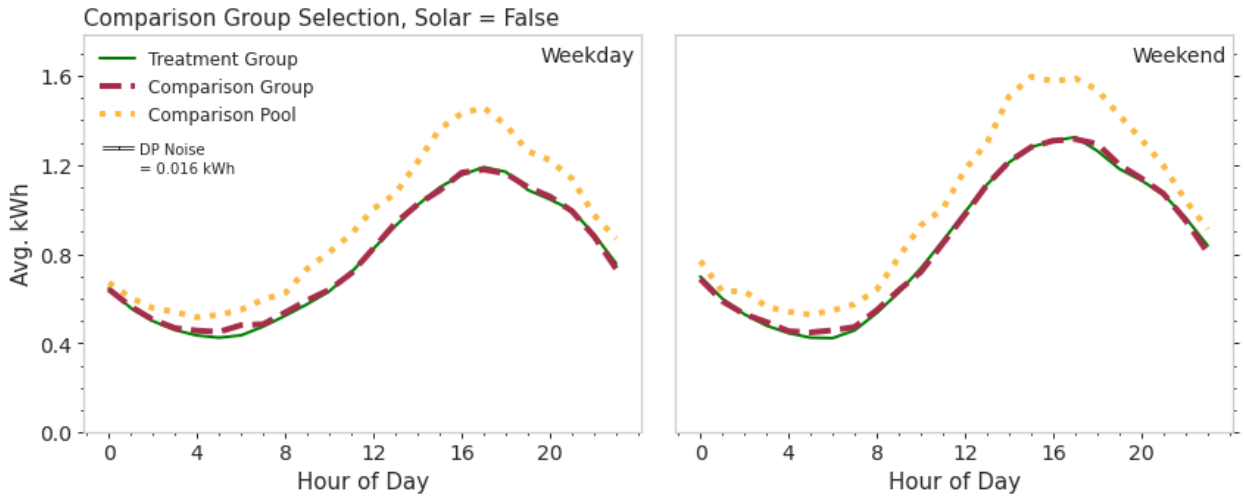


Figure 22: Group A3, Non-Solar - The average weekday (left) and weekend (right) load shape of a meter in the treatment group (solid green), comparison pool (dotted orange), and comparison group (dashed red).

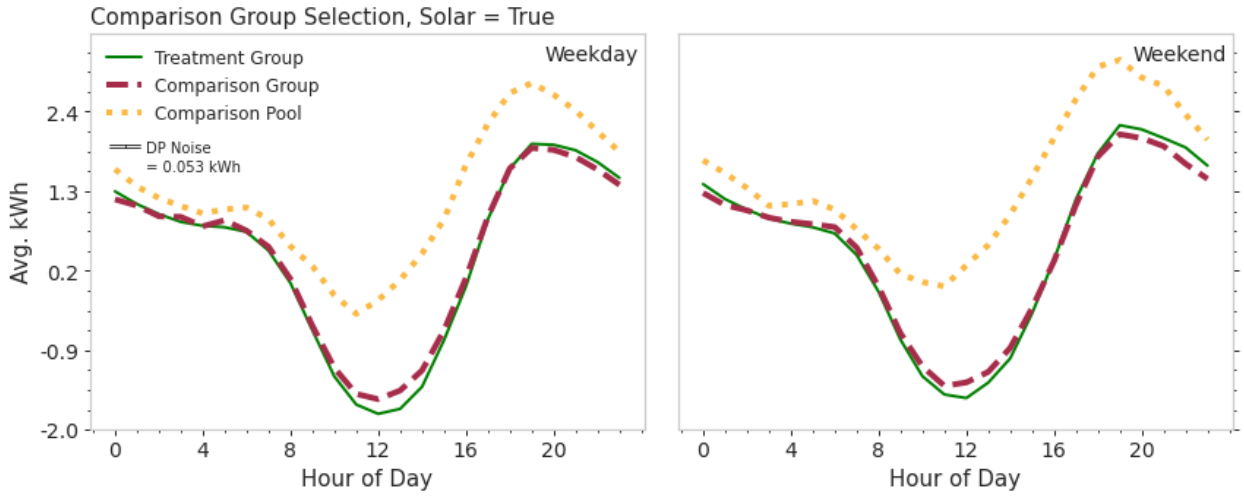


Figure 23: Group A3, Solar - The average weekday (left) and weekend (right) load shape of a meter in the treatment group (solid green), comparison pool (dotted orange), and comparison group (dashed red).

Figures 24 and 25 break out the non-solar treatment customers by usage quartiles and daily average hourly min-to-max ratio quartiles, respectively. These figures demonstrate that the average load shape matches of Figure 20 results from close matching throughout the range of participants.

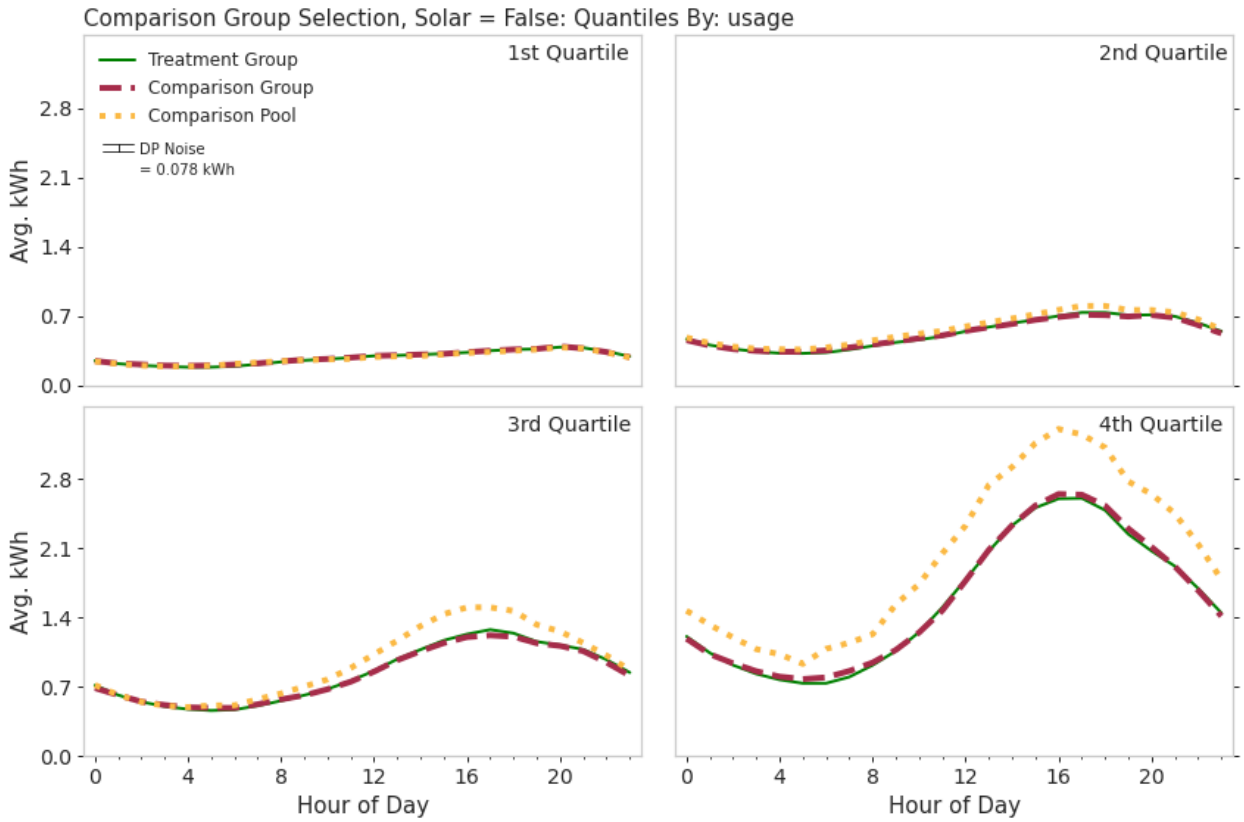


Figure 24: Group A3, Non-Solar - The average load shape of a meter in the treatment group (solid green), and comparison pool (dotted orange) broken out by baseline usage quartiles. Matched comparison group average load shapes are also shown (dashed red). The lowest quartile is in the top left panel and the highest quartile is shown in the bottom right panel.

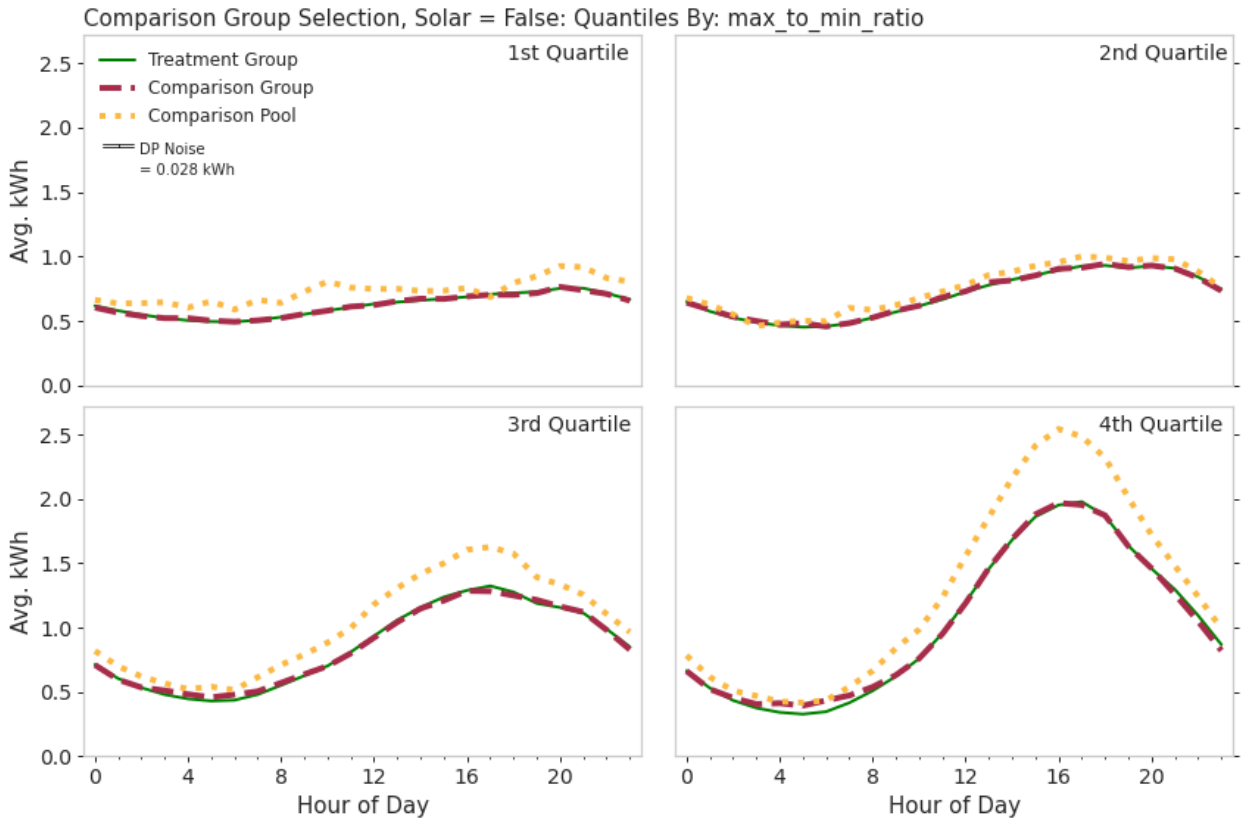


Figure 25: Group A3, Non-Solar - The average load shape of a meter in the treatment group (solid green), and comparison pool (dotted orange) broken out by average daily maximum to minimum quartiles. Matched comparison group average load shapes are also shown (dashed red). The lowest quartile is in the top left panel and the highest quartile is shown in the bottom right panel.

Similarly, the group of solar customers shows a good match throughout the distribution of participants. Figure 26 shows how the matching performs across different quartiles of evening ramp steepness.

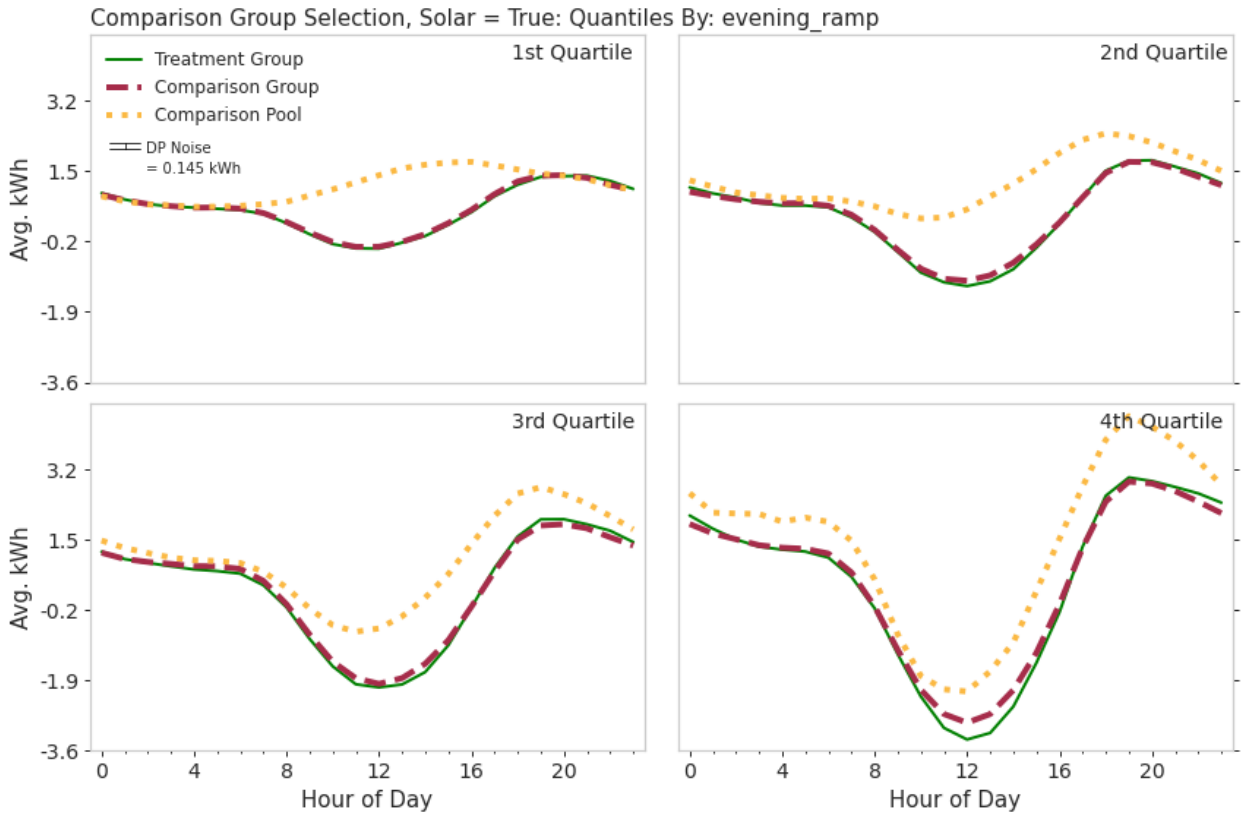


Figure 26: Group A3, Solar - The average load shape of a meter in the treatment group (solid green), and comparison pool (dotted orange) broken out by evening ramp steepness. Matched comparison group average load shapes are also shown (dashed red). The lowest quartile is in the top left panel and the highest quartile is shown in the bottom right panel.

Tables 1 and 2 show that Group A3 had 5 events during the week of Aug 14, 2020. Figure 27 shows the average temperature and hourly savings by participant for event and non-event periods for these five events. Average participant hourly savings hovered between 0.08 kW and 0.14 kW for all events with the exception of the hottest event day in which savings dipped. As would be expected, average temperatures during non-event hours were lower. Figure 27 shows a small but consistent degree of negative savings during these times, a topic discussed further below.

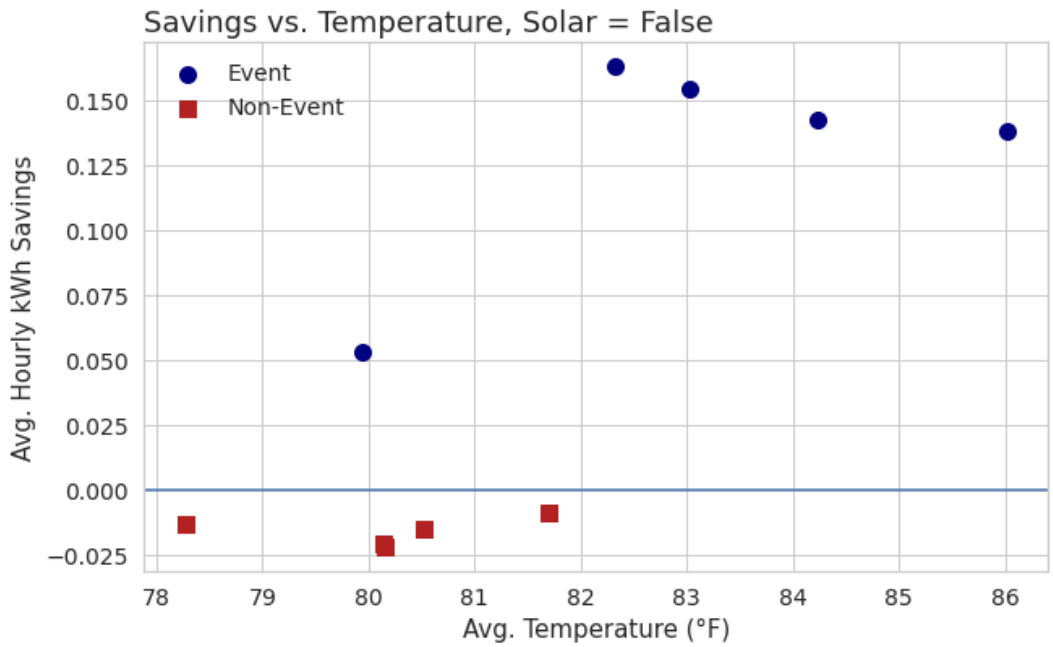


Figure 27: Group A3, Non-Solar - Average participant hourly savings during event (blue circles) and non-event (red squares) periods as a function of Temperature. Maximum differential privacy noise fraction = 0.0013

Solar customers (shown in the Extended Results Appendix) also produced event period savings and exhibited lesser impacts during non-event periods, though with a far smaller group size much greater variation was observed.

Figure 28 breaks down load impacts by hour type. Slight negative savings (< 1%) are observed in hours that are beyond 2 hours preceding or following the event. However, in the 2 hours bracketing the event on both sides, a higher degree of negative savings are observed (2%). This type of "rebound" effect may be expected but it is important to understand and quantify since, depending on the event hours, the added load may occur when the grid is still under stress.

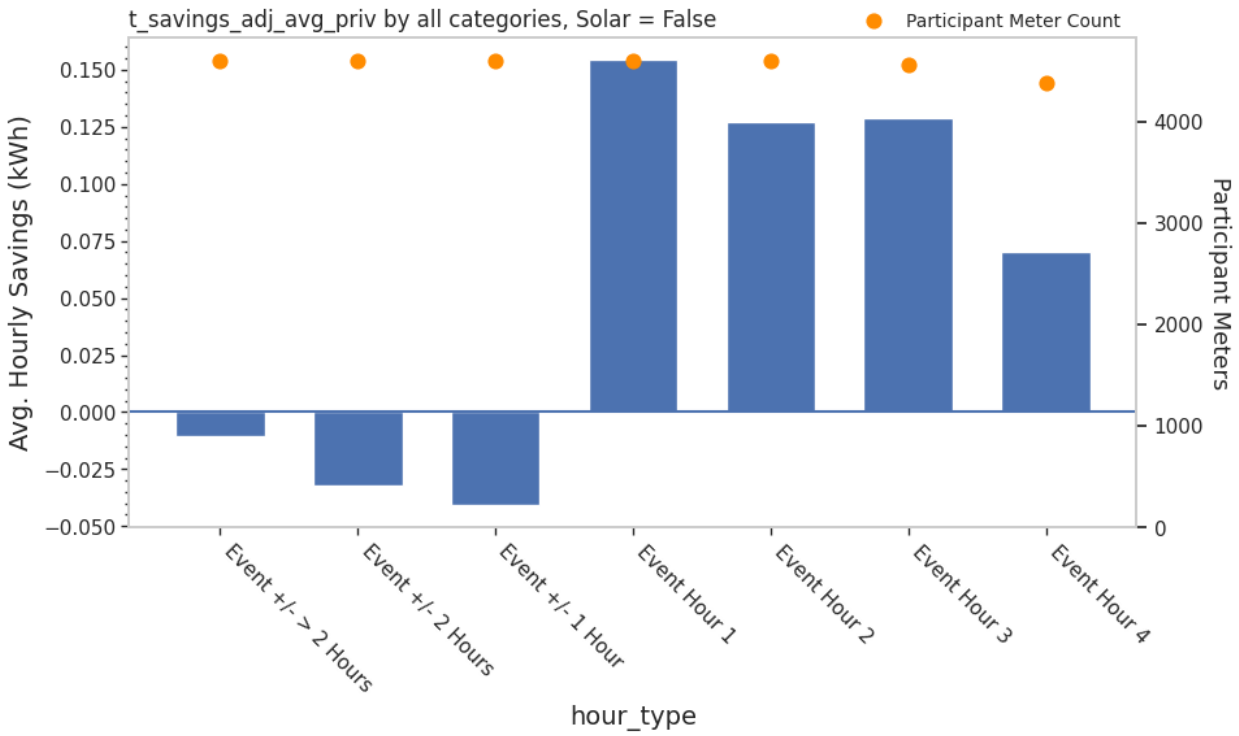


Figure 28: Group A3, Non-Solar - Average participant hourly savings during the indicated hour types (x-axis). The participant counts for each category are shown as orange circles and refer to the right-hand axis.

Event hour 1 shows the strongest response with nearly 8% savings. (For an analogous graph that shows fractional savings see the Extended Results Appendix). However, diminishing returns are evident as the event persists. Upon reaching the fourth event hour, savings are roughly half of the first hour.

While it is useful to see averaged or summed savings values, these statistics do not tell the whole story. With large numbers of participants, some customers will respond aggressively to an event while others will not alter their behavior or may even override integrated systems. Additionally, events do not take place in a bubble; individual customers will use more or less during any given hour due to a host of non-program factors related to both baseline and event period timeframes. Therefore, to understand program performance it is essential to gauge the shift in the *distribution* of savings across the population of participants.

With comparison group adjustments being applied at an aggregated population level, it is difficult to see directly the adjusted savings that serve as a final barometer of program impacts in a histogram. Nevertheless, it is instructive to assess the distribution of unadjusted savings and to compare event periods to non-event periods. To that end, Figure 29 shows the distribution of unadjusted savings from the population of program participants during all hours of the events (blue) as well as the non-event hours (red) that were assessed during the

week of August 14. This plot is normalized such that the heights of all bars in a given distribution add to 1. In this way, the different shapes reveal the comparisons of interest.

Somewhat surprisingly, the average unadjusted savings is very nearly zero for both periods, each displaying highly symmetric distributions centered around zero. The tails of the event period distribution are wider than the non-event profile, but this can be explained by higher and likely more varied consumption during evening hours in general.

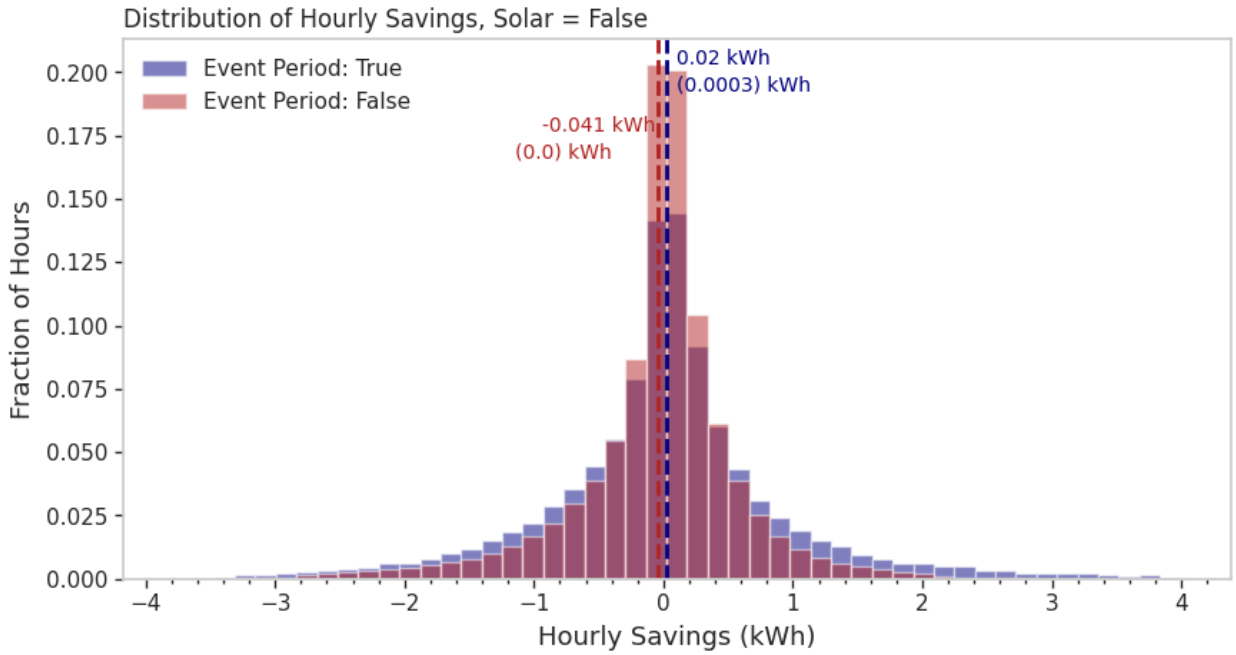


Figure 29: Group A3, Non-Solar - Distribution of participant-level hourly *unadjusted* savings during the event period (blue) and non-event period (red).

However, when focusing the analysis on the first event hour and the hours immediately preceding and following the event, a clear trend emerges. Figure 30 shows how the "rebound" hours are shifted to negative savings.

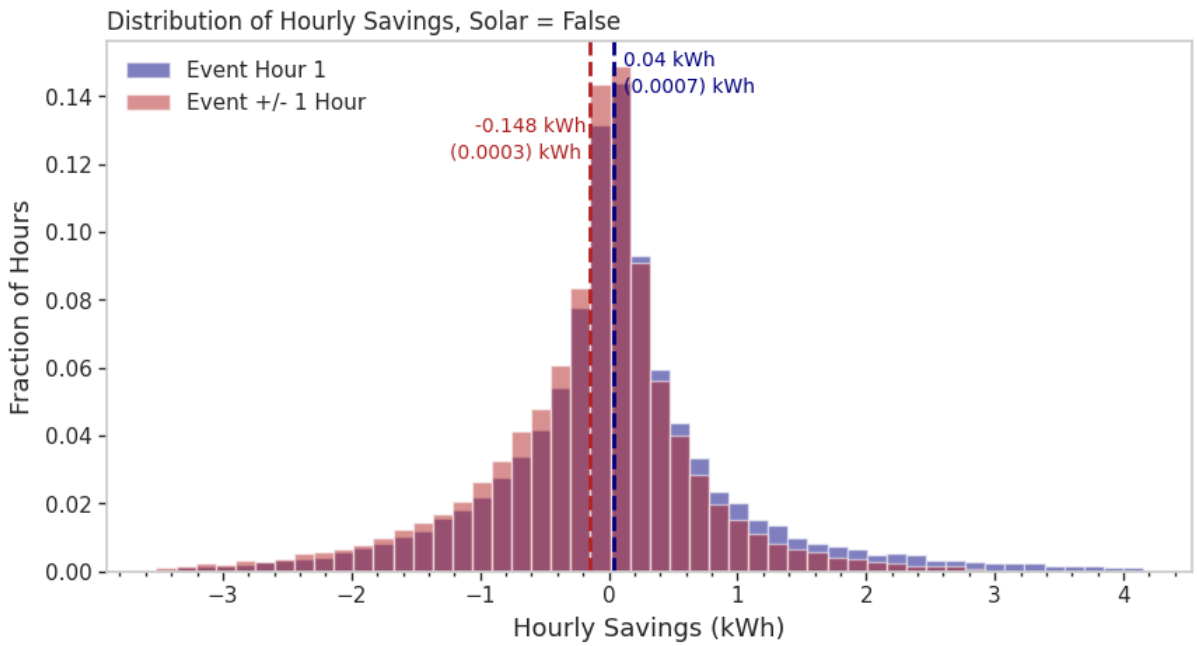


Figure 30: Group A3, Non-Solar - Distribution of participant-level hourly *unadjusted* savings during the first event hour (blue) and hours immediately preceding or following the event (red). Vertical lines indicate respective means.

Similar results are observed among the solar customers, though the limited number of these participants causes increased variability in the plot (Figure 31).

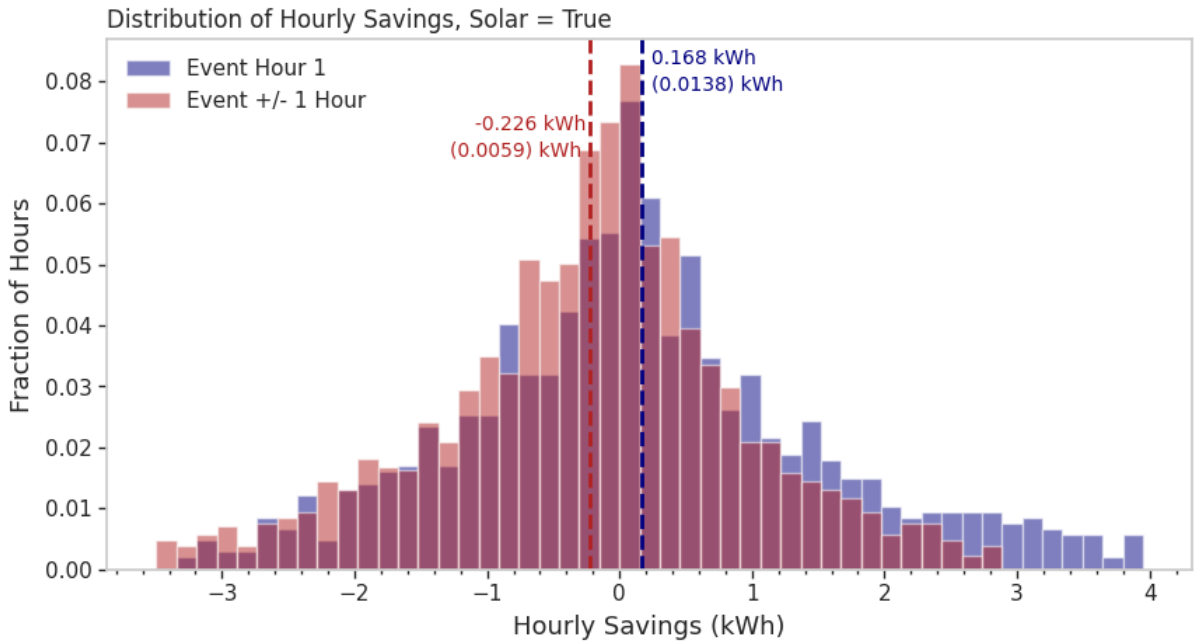


Figure 31: Group A3, Non-Solar - Distribution of participant-level hourly *unadjusted* savings during the first event hour (blue) and hours immediately preceding or following the event (red). Maximum histogram noise = 0.0015.

How can it be that customers with apparently no collective event response would exhibit a rebound effect? To answer this question we can turn to the comparison group analysis.

Figure 32 summarizes the *% Difference of Differences* load impact measurement, also across all event days. This figure combines several elements of the stepwise FLEXmeter methods walkthrough covered in Section III. Because Figure 32 summarizes event data across multiple days during which events were called over different hours with potentially different enrollment, it is helpful to indicate the degree to which a particular hour was subject to an event. This is done with darker shading for hours with higher event overlap and lighter shading for hours with little (but not zero) event participation.[39]

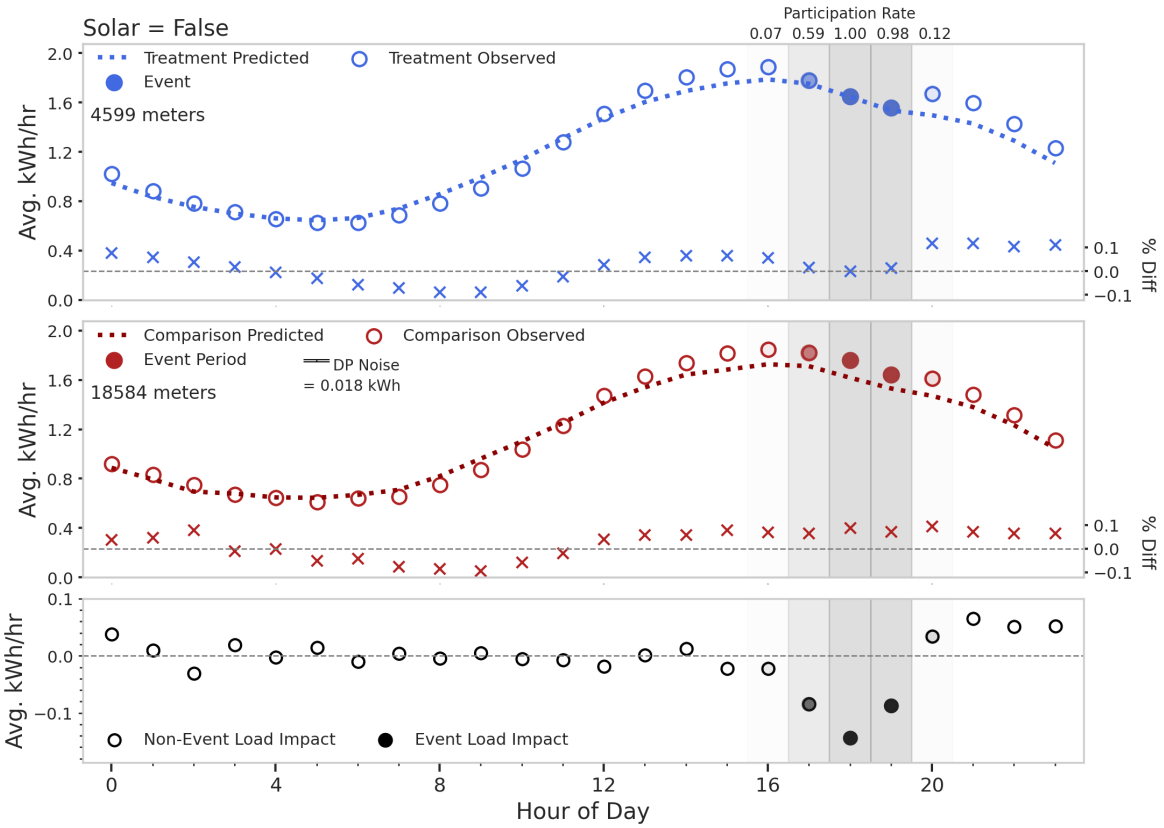


Figure 32: Group A3, Non-Solar - Top: The dotted trace gives the average counterfactual (model-predicted) usage of a participating customer. The circles are the average observed meter readings (open = non-event hour, filled = event hour). The fractional difference between the two traces are shown as ✕'s and refer to the right-hand axis. Middle: Analogous results for the comparison group. Bottom: Average hourly load impacts across all events. The shading and the "participation rate" values at the top of the plot indicate the fraction of hours associated with an event.

[39] The participation rate is calculated as the fraction of hours the participating customer base was subject to an event relative to all hours. For instance, if a treatment group consists of 1,000 customers, 500 may be called for an event on a particular day and of these maybe only 250 are called during hour 18. In this case the participation rate for hour 18 would be 25%. A participation rate of 1.0 indicates that the full treatment group was called during that particular hour every day in which an event was called during the study window.

In the middle panel of Figure 29, we see that the comparison group actual consumption during the event periods is higher than the model prediction (counterfactual). The fact that the participants do not exhibit this trend but rather fall close to the counterfactual indicates the events indeed had an impact. In fact, during the hour of maximum participation (18), savings of greater than 10% were achieved!

The result of this case study illustrates the importance of using a comparison group to properly attribute savings. A similar effect is seen amongst the solar group, shown in Figure 33.

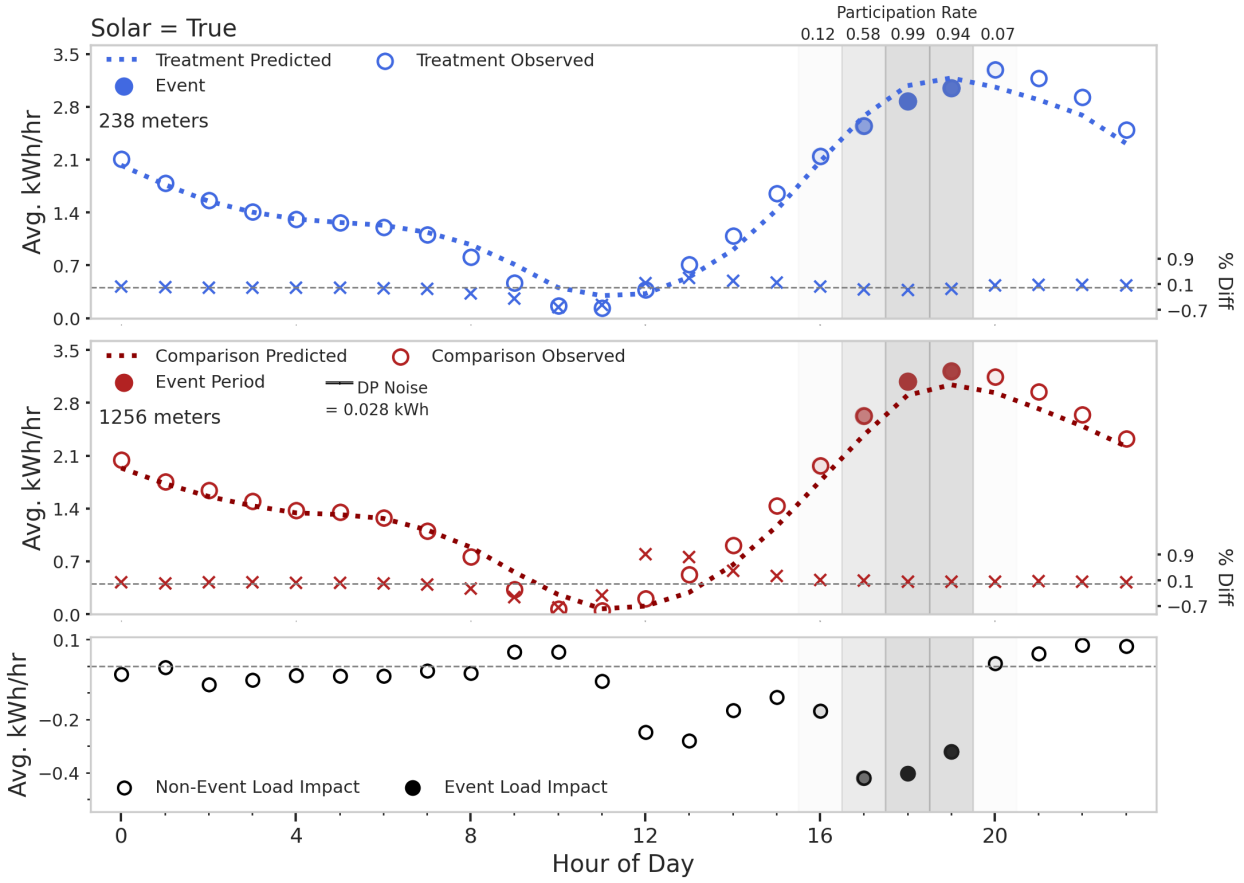


Figure 33: Group A3, Solar - Top: The dotted trace gives the average counterfactual (model-predicted) usage of a participating customer. The circles are the average observed meter readings (open = non-event hour, filled = event hour). The fractional difference between the two traces are shown as ✕'s and refer to the right-hand axis. Middle: Analogous results for the comparison group. Bottom: Average hourly load impacts across all events. The shading and the "participation rate" values at the top of the plot indicate the fraction of hours associated with an event.

Figure 33 shows some instability in the midday hours of the savings calculation for solar customers. This arises because counterfactual usage can be very near zero for daytime hours where solar production nearly balances building consumption. This can cause the denominator of the % Diff calculation to be very large, which is then reflected in the final savings values. For the vast majority of customers, this issue does not pose a threat to event

period savings, as solar PV production is usually much lower or zero during these hours. The remedy for this issue is continued methodological development such that the TOWT model can also handle variables predictive of solar PV output (such as cloud cover and solar irradiance). In the interim, we must be mindful of this effect. For event-specific figures see the Extended Results Appendix.

ii. Non-Residential

Group C1 provides a good case study of a non-residential program that serves large customers. In C1 there were both non-solar and solar participants but the latter group was small and is omitted here. We begin again by inspecting the comparison group matching step (Figures 34 - 36), applied across 138 participating customers.

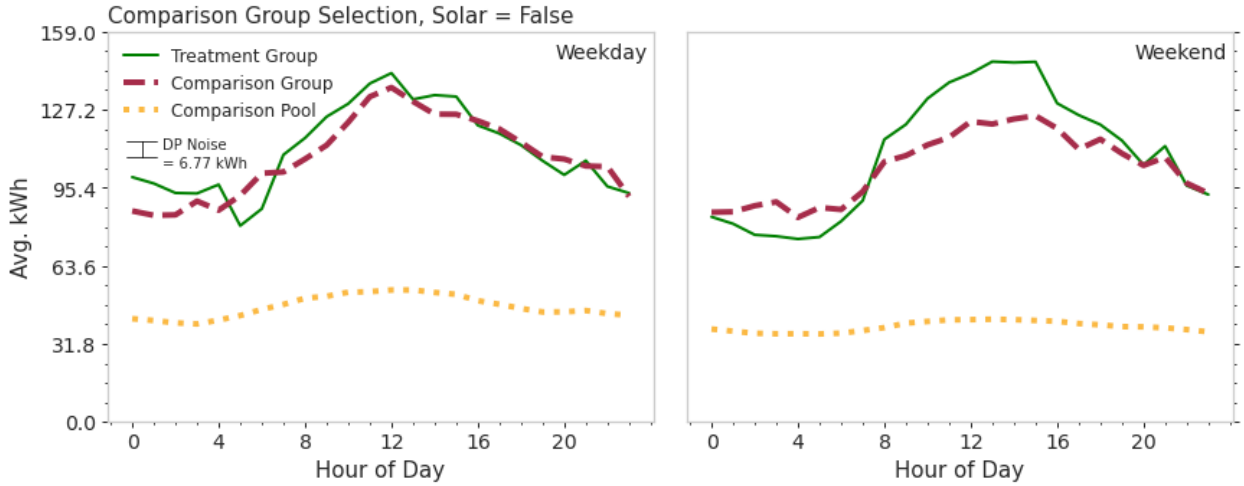


Figure 34: Group C1, Non-Solar - The average weekday (left) and weekend (right) load shape of a meter in the treatment group (solid green), comparison pool (dotted orange), and comparison group (dashed red).

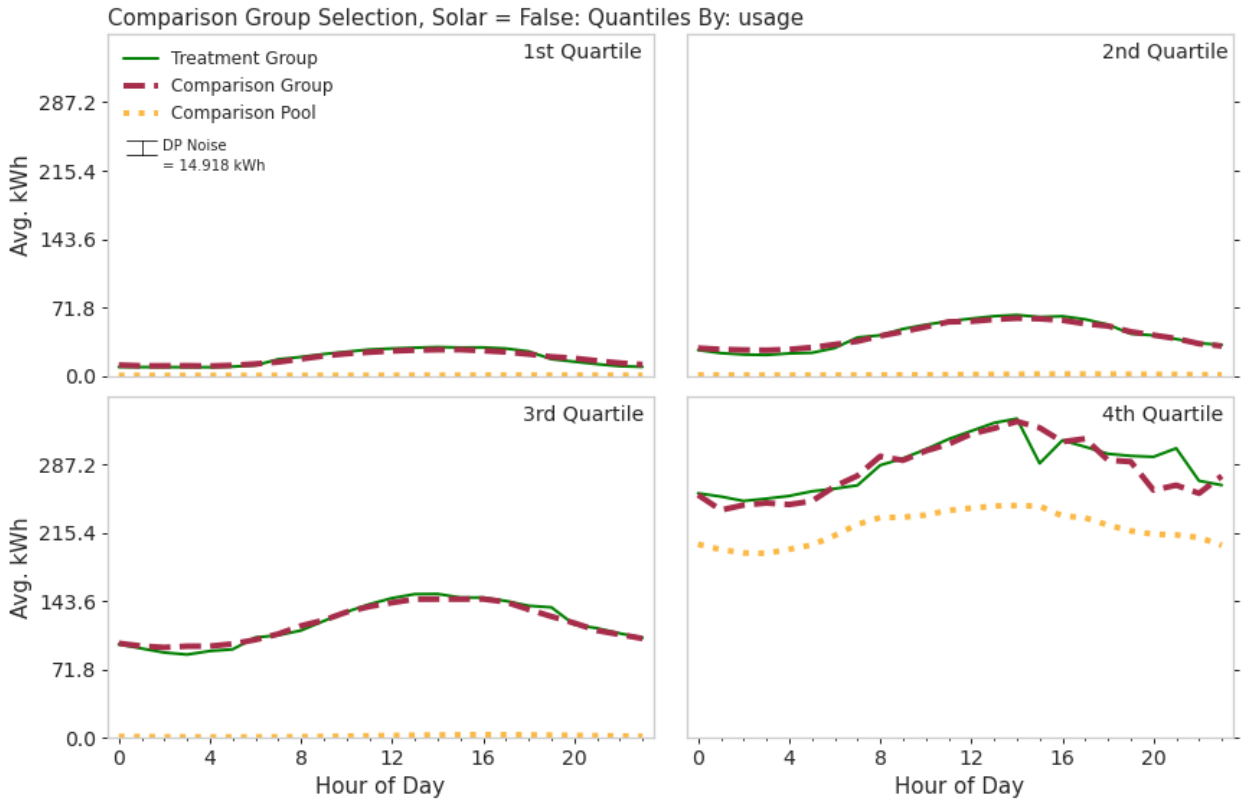


Figure 35: Group C1, Non-Solar - The average load shape of a meter in the treatment group (solid green), and comparison pool (dotted orange) broken out by baseline usage quartiles. Matched comparison group average load shapes are also shown (dashed red). The lowest quartile is in the top left panel and the highest quartile is shown in the bottom right panel.

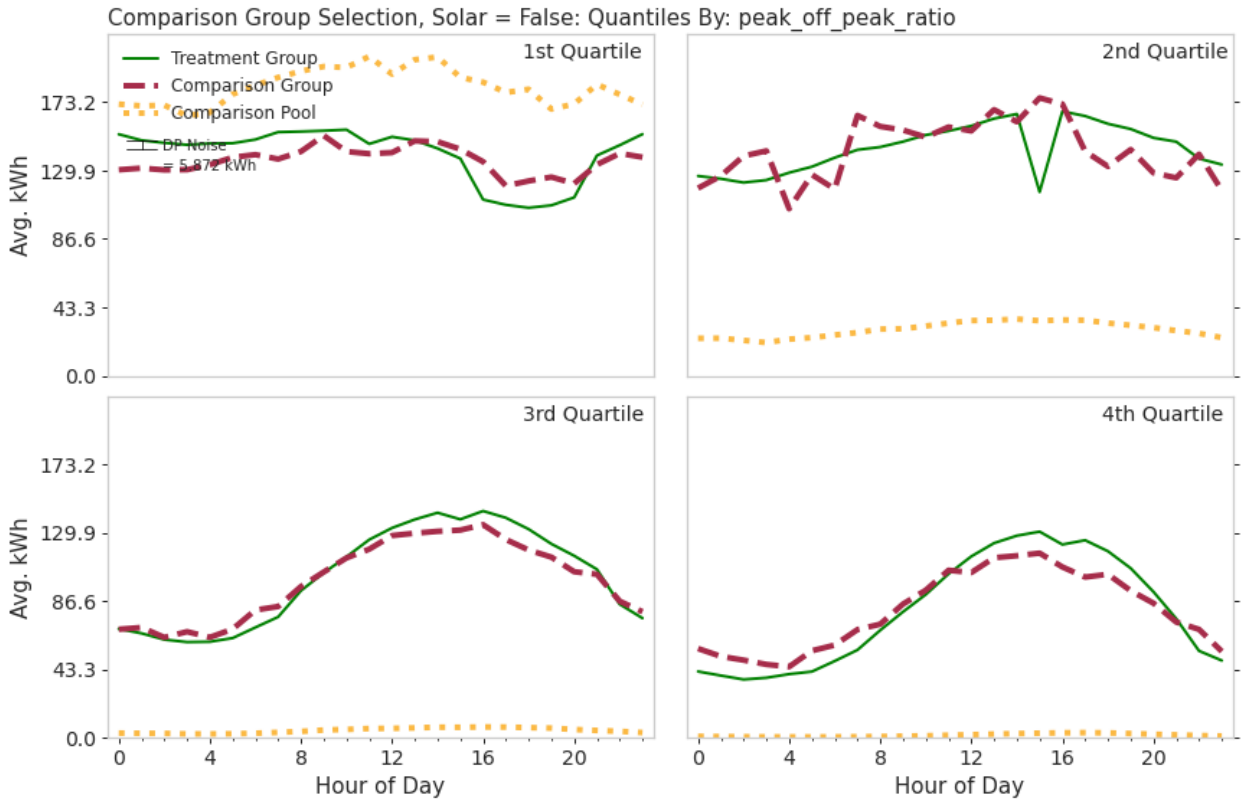


Figure 36: Group C1, Non-Solar - The average load shape of a meter in the treatment group (solid green), and comparison pool (dotted orange) broken out by average evening ramp quartiles. Matched comparison group average load shapes are also shown (dashed red). The lowest quartile is in the top left panel and the highest quartile is shown in the bottom right panel.

These figures show that a reasonable match between treatment and comparison groups has been achieved. It is clear that the participants are substantially larger on average than a typical non-residential customer. However, the site-based matching appears to have succeeded in pulling larger customers with similar load profiles into the comparison group. With only 138 treatment meters we observe a greater degree of hour-over-hour variation in average consumption and with a smaller pool of customers (large commercial) to sample from it is not surprising that some daylight persists between treatment and comparison groups. In cases such as this, the normalized basis upon which the comparison group savings adjustment calculations are conducted is particularly important.

Table 4 details results for the 5 events studied for group C1. Figure 37 shows the average temperature and hourly savings by participants for event and non-event periods. Again, both average temperatures and savings are consistently higher for event periods. Non-event periods showed near zero program impacts.

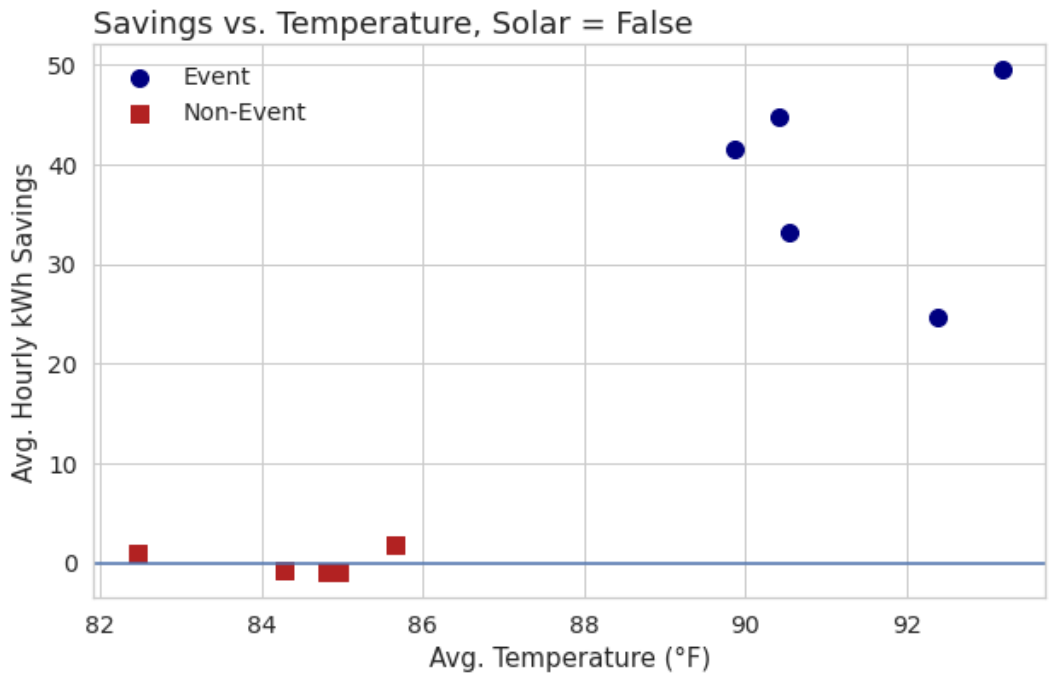


Figure 37: Group C1, Non-Solar - Average participant hourly savings during event (blue circles) and non-event (red squares) periods as a function of Temperature.

Figure 38 shows the distribution of unadjusted hourly savings for all event days broken out but event and non-event periods. Unlike the first case study where little difference was observed in the analogous distributions, in this case, we observe a significant shift during the event period.

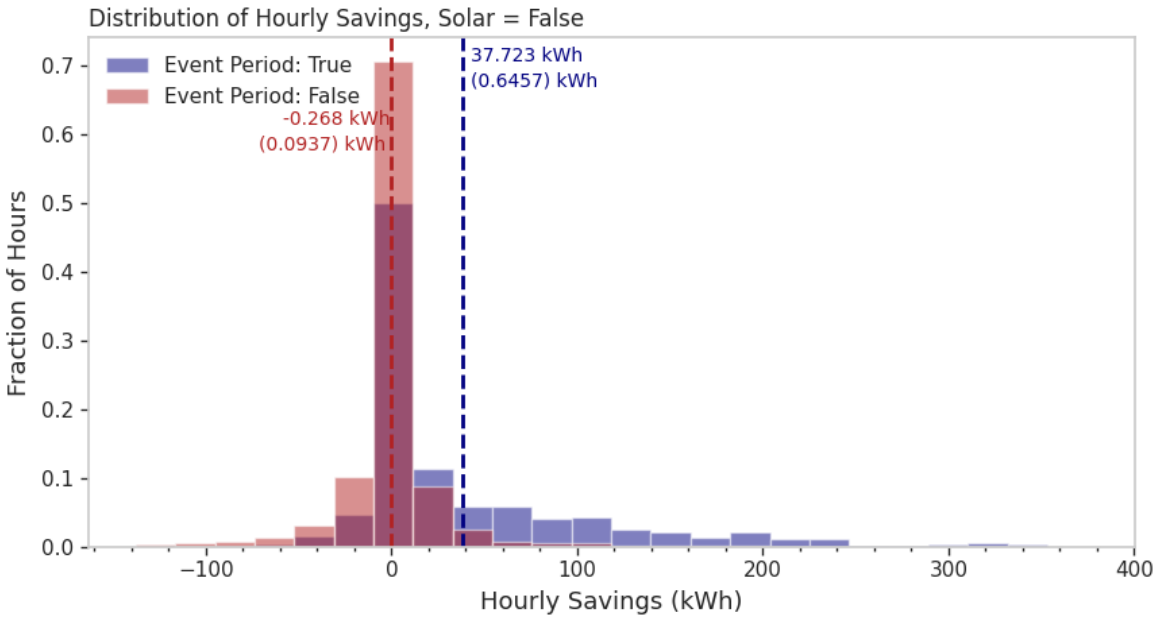


Figure 38: Group C1, Non-Solar - Distribution of participant-level hourly *unadjusted* savings during the event period (blue) and non-event period (red).

Figure 39 catalogs savings by hour type. While savings dip slightly after the first event hour they remain strong for the duration, even into events that persisted for five hours. Interestingly, no evidence of takeback is seen in the hours immediately surrounding the event. In fact, positive savings are observed during these times. With large commercial customers this raises an important question: Are customers simply shedding load or is some of this load shifted to other days? The latter case could present baseline issues and further study may be warranted to understand if such a shift is taking place.

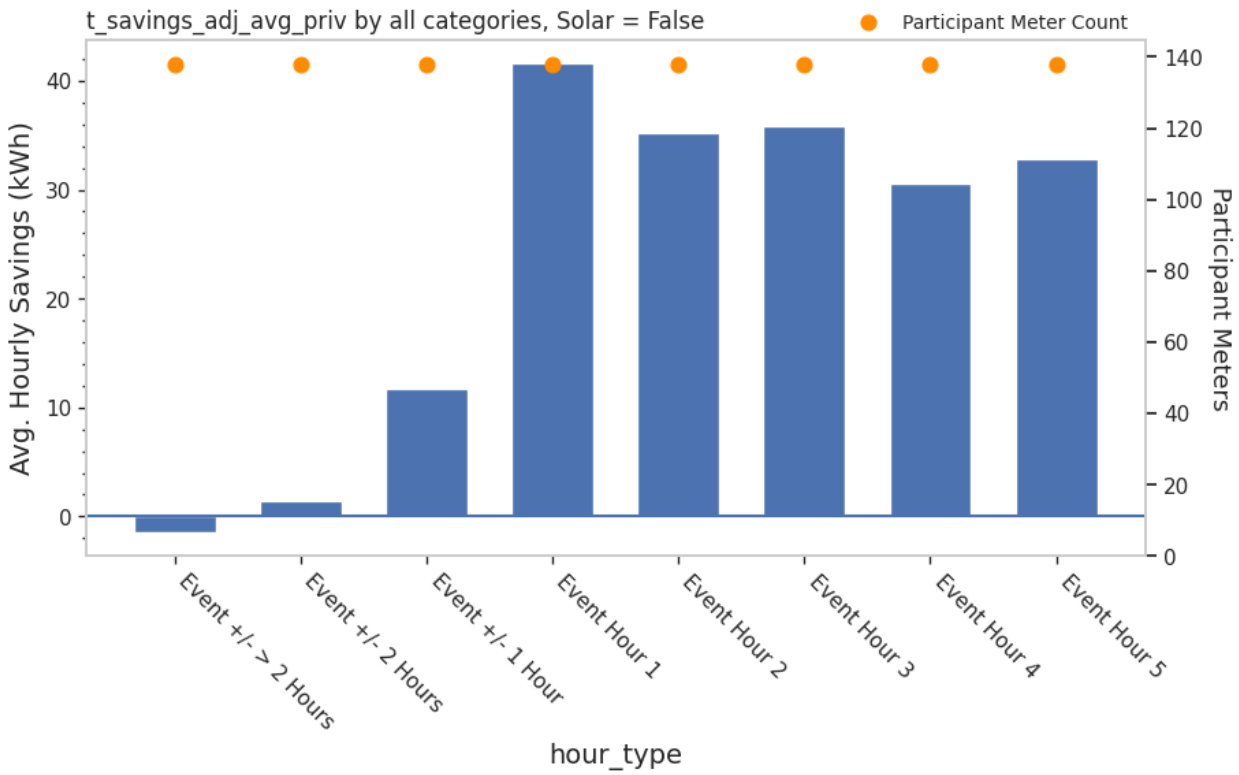


Figure 39: Group C1, Non-Solar - Average participant hourly savings during the indicated hour types (x-axis). The participant counts for each category are shown as orange circles and refer to the right-hand axis.

Figure 40 provides an analogous breakout but plots fractional instead of absolute savings.

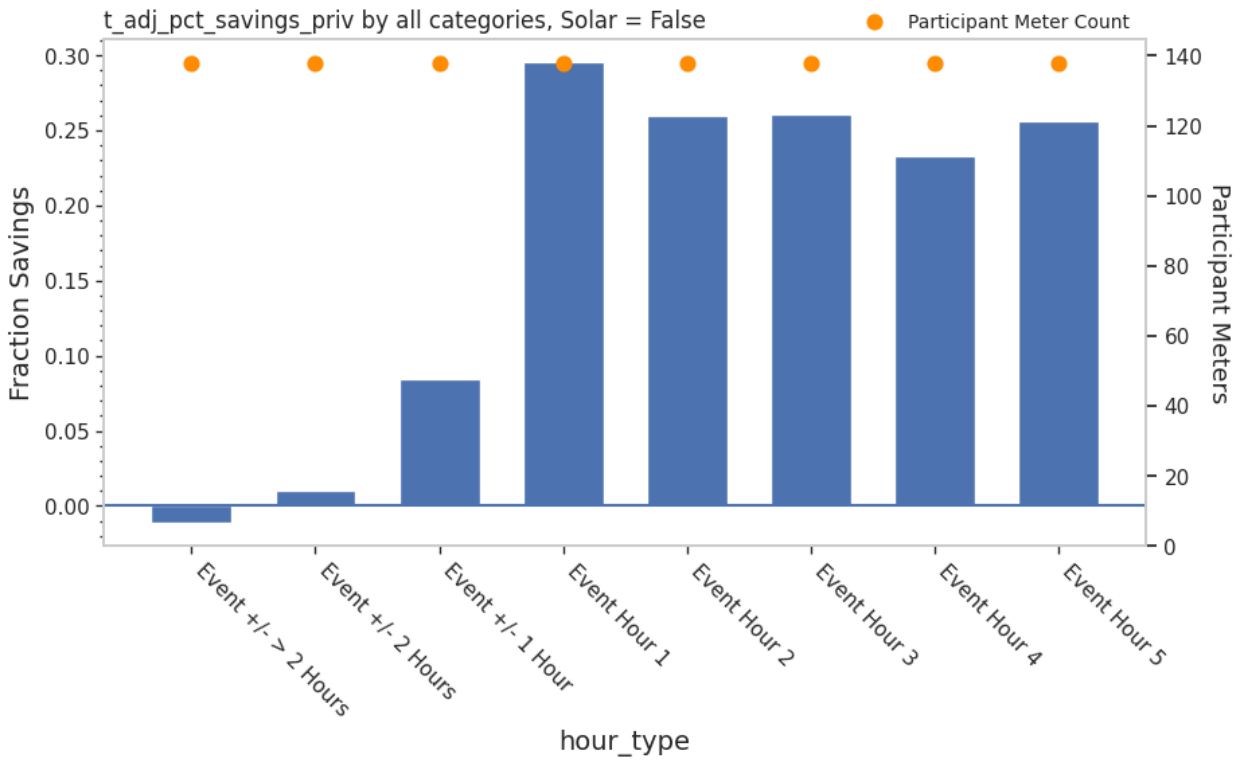


Figure 40: Group C1, Non-Solar - Average participant hourly savings during the indicated hour types (x-axis). The participant counts for each category are shown as orange circles and refer to the right-hand axis.

The events for group C1 are summarized in Figure 41. The fractional participation by event hour shows a clear pattern; savings were strongly correlated with the degree of event participation.

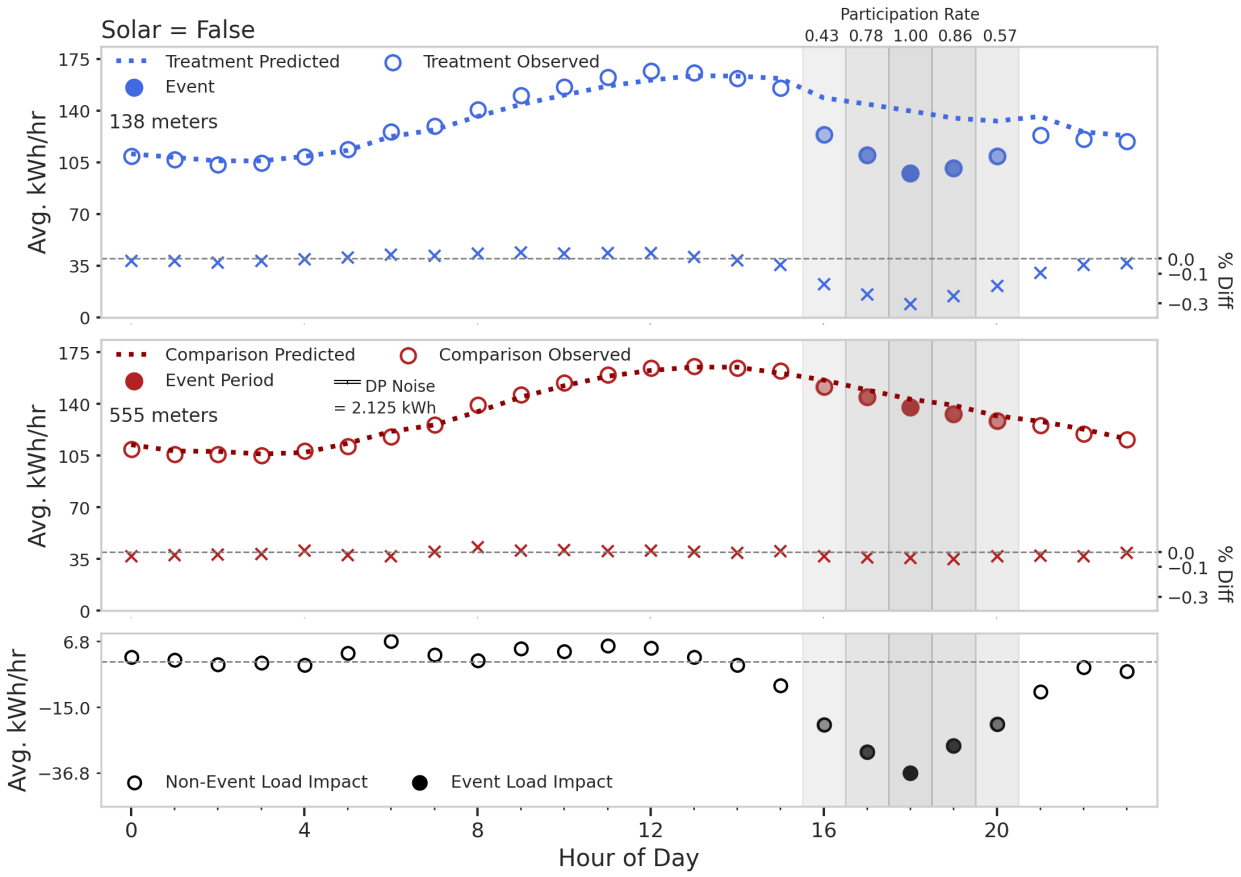


Figure 41: Group C1, Non-Solar - Top: The dotted trace gives the average counterfactual (model-predicted) usage of a participating customer. The circles are the average observed meter readings (open = non-event hour, filled = event hour). The fractional differences between the two traces are shown as X's and refer to the right hand axis. Middle: Analogous results for the comparison group. Bottom: Average hourly load impacts across all events. The shading and the "participation rate" values at the top of the plot indicate the fraction of hours associated with an event.

VI. Error, Outliers and Uncertainty

While rigorous error analysis is not included for every statistic reported here, assessments have been undertaken to check for bias and to understand and quantify model error. Models were evaluated using in-sample and out-of-sample tests to determine the accuracy of fit and savings estimation. Here we detail the process of error measurement and provide an example of its application.

Dummy events were created to provide a testbed for analysis. A dummy event is defined as a timeframe of similar hours as actual events but on days where no demand response event was called. Dummy events were used to measure out-of-sample error by taking the difference between the model fit and the hourly observed consumption values. Without the influence of demand response on the dummy event days, the "load impact" value can be interpreted as model residual. The dummy event information is blacked out for each dummy

event model fit, meaning information about the period does not inform the model. The dummy residuals thus represent the out-of-sample error for each participant.

Figure 42 is a residual plot for all dummy events for DRP B LSE 2 and Figure 43 summarizes these data in a histogram. The residuals are heteroskedastic and appear random, which suggests that the linear regression model is being used appropriately. Furthermore, the residuals are normally distributed and centered around zero, which suggests that conducting statistical inference on these samples can produce reliable results. Residual analysis did reveal outliers, which upon further investigation likely stemmed from measurement error. More details on outliers are available below, but for the purposes of illustration residuals were plotted with an outlier cutoff.

The error analysis was limited to DRP B LSE 2 due to time constraints, but was done to ensure that a standard evaluation of model fit is available for all future methods application. This analysis can also be applied to outlier detection to identify and isolate irregularities. The standard methodology for error analysis will be added to the open source package. Importantly, error analysis allows for the uncertainty in the model to be measured. In this particular case the model error has a mean of .008 kWh/hr and falls within 0.008 ± 0.006 kWh/hr at the 95% confidence level.

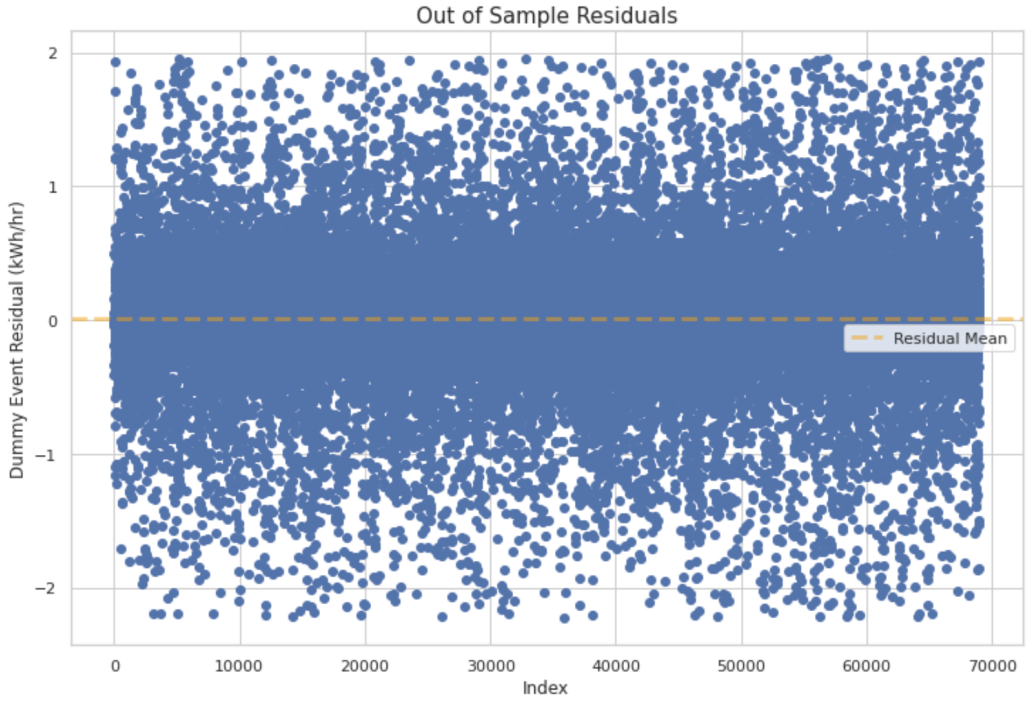


Figure 42: Out of sample residuals for DRP B LSE 2. The residuals are randomly distributed and centered around 0. There does not appear to be any trends in the residuals and they appear heteroskedastic.

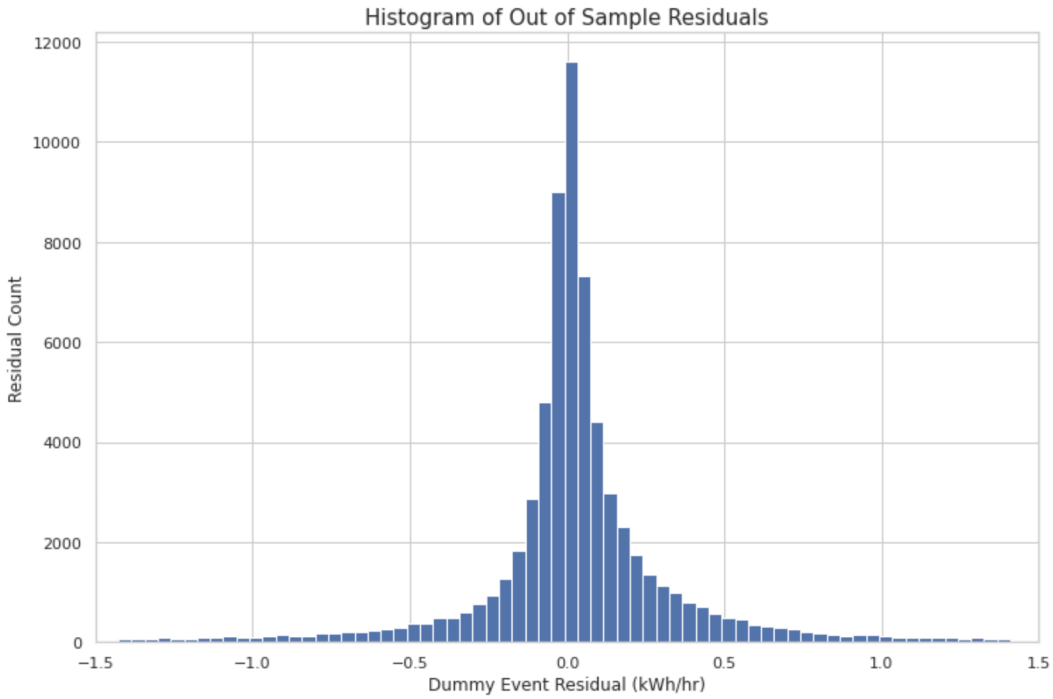


Figure 43: Histogram of residuals for DRP B LSE 2. The out of sample error is centered around 0 and normally distributed.

Outlier detection and removal is an important detail of any methodology. Recurve found that significant issues can arise from the presence of outliers. In order to ensure replicable results, Recurve recommends further research into the standardization of outlier removal. An open-source methodology for outlier removal will be extremely valuable in ensuring stable and reasonable results are produced from these methods.

Three types of outliers were generally encountered in this study. The first was a magnitude outlier: customers who use much more energy than their counterparts. These outliers are not necessarily erroneous. However, magnitude outliers can be difficult to match with an appropriate comparison group. It is recommended that such outliers are examined on a case-by-case basis and if appropriate analyzed separately.

The second outlier type commonly encountered was erratic usage. These outliers often arise due to issues in raw consumption data. They are more challenging to detect and often present themselves as unrealistic models with high error. Recurve implemented model fit checks as a way of finding erratic usage outliers and eliminated meters with CalTRACK Hourly CVRMSE outside the bounds of -2 to 3. Further research can help refine these limits or provide alternative outlier metrics associated with model fit. Finally Recurve implemented an outlier filter to eliminate hours in which Equation 1 (fractional between observed and counterfactual hourly usage) was outside the bounds of -10 to 10.

A common question in population-level meter based programs (including energy efficiency and demand response) is With both treatment groups and comparison groups as well as hourly modeling, comprehensive uncertainty analysis is more challenging. As a proxy for

uncertainty Recurve also conducted Monte Carlo analyses on residential and non-residential results for DRP D using the following approach:

1. Randomly split participants into two equal sized samples
2. Calculate sample 1 percent event savings
3. Calculate sample 2 percent event savings based on growing random samples
4. Calculate difference between sample 1 and sample 2 percent event savings at every step
5. Repeat analysis 100 times

Each scan is unique. As the sample 2 size grows statistical fluctuations in the difference between sample 1 and sample 2 will tend to decrease. However, some samples will tend to align even with small sample sizes while others will be misaligned.

Figure 44 shows results of all scans for the residential sector. The August 15 events are the basis for this analysis. The x-axis (log) plots the number of sample 2 meters on which sample 2 percent savings and the difference with sample 1 is being calculated. The y-axis gives the absolute value of the difference between sample 1 and 2. This difference (labeled "Fraction Residual") is on the basis of total consumption. So if sample 1 value is 25% savings and sample 2 is 22% savings then the difference will be 3% or 0.03 on the y-axis.

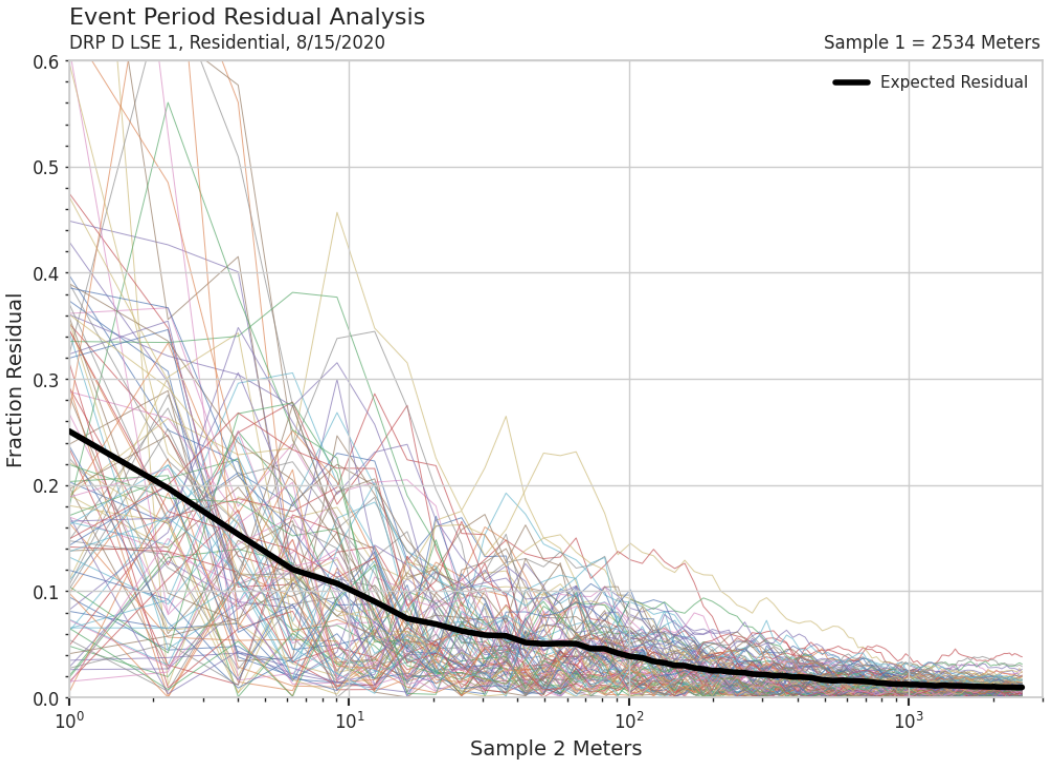


Figure 44: Monte Carlo analysis results of the difference between percent savings of random samples for DRP D LSE 1, residential. See text for additional details.

Each of the 100 traces is shown in Figure 44 with the average value plotted as the thick black line. With these data, Recurve then rank ordered the residuals at each sample 2 step. Using this ranking we then plotted the expected residuals at the 5th, 50th and 95th percentiles (Figure 45).

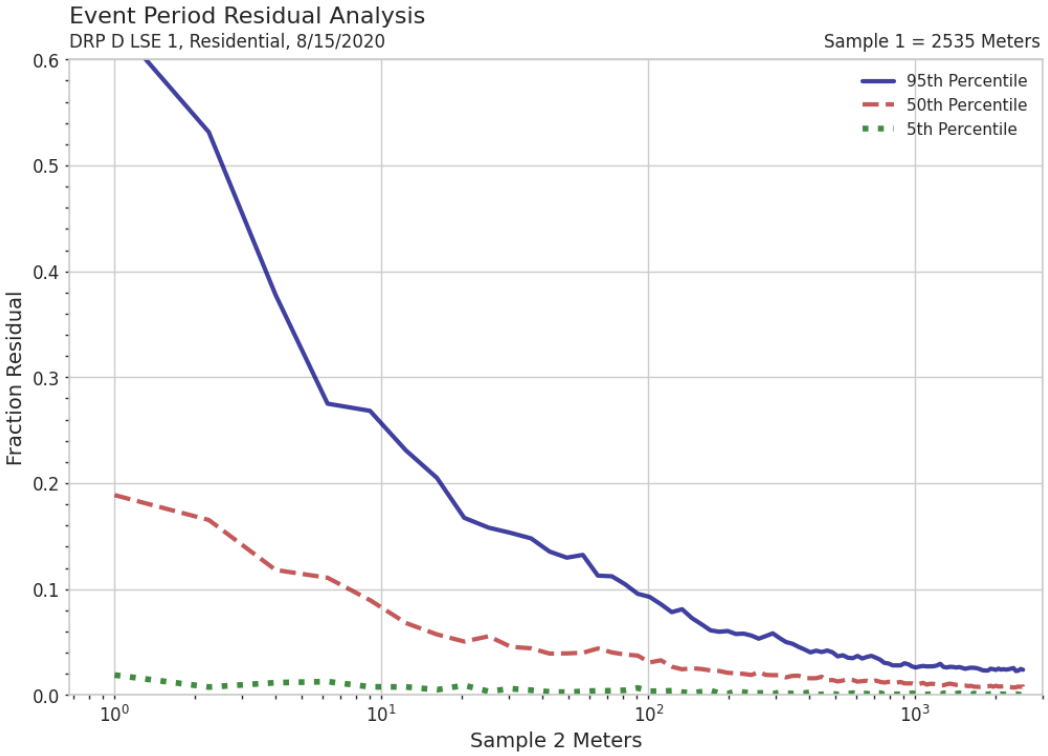


Figure 45: Monte Carlo analysis results of the 5th, 50th and 95th percentile differences between percent savings of random samples for DRP D LSE 1, residential. See text for additional details.

While different programs and different territories will show different results, Figure 45 provides an important check and reveals the following ballpark figures for residential demand response programs:

- In most cases about 100 meters are needed to measure event savings within an uncertainty bound of approximately 3% of predicted event period usage.
- However, to be 95% confident that a sample can achieve this level of certainty, more than 1,000 participants may be required.
- Stated a different way, for a program achieving 20% event period savings, for an event savings measurement to achieve +/- 15% uncertainty at a 95% confidence level on the order of 1,000 meters should be targeted.

Figure 46 shows analogous results for the small/medium business sector. Statistical variability is slightly higher than in the residential sector.

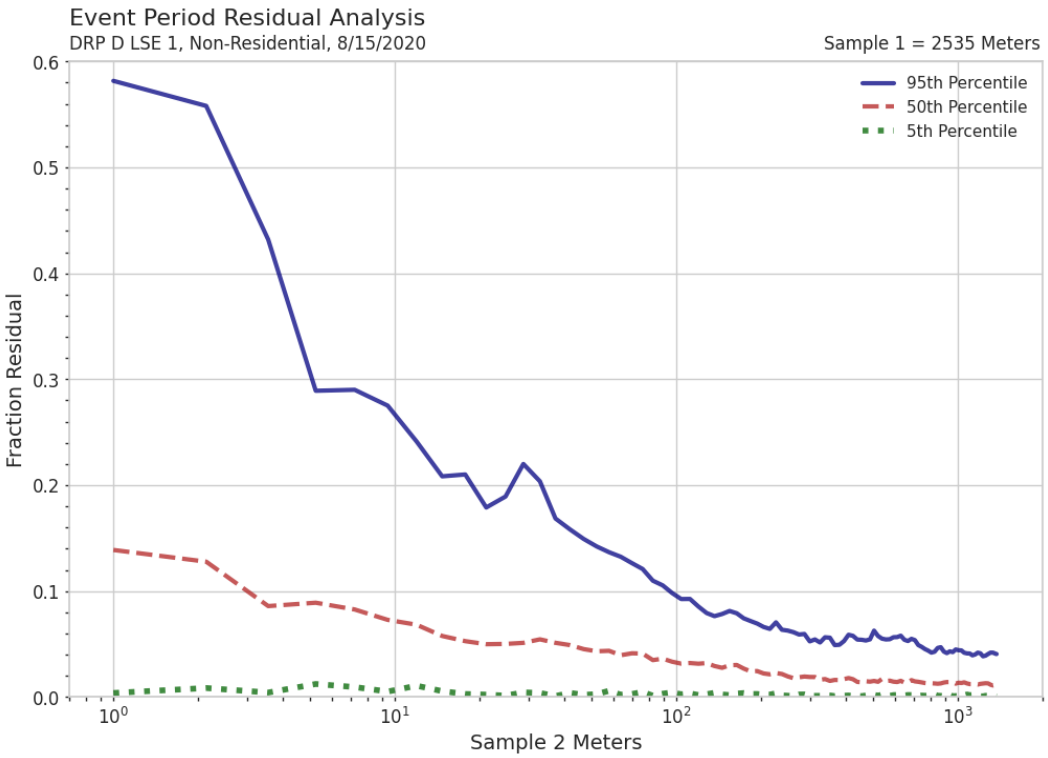


Figure 46: Monte Carlo analysis results of the 5th, 50th and 95th percentile differences between percent savings of random samples for DRP D LSE 1, residential. See text for additional details.

VII. Data Access and Recommended Future Pathways

Comparison group methods outlined in this paper are dependent on access to non-participant data. While energy consumption data is a primary component of the required data, other information regarding non-participant customers, including their sector and location, enables a proper comparison group to be formed. 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.[40] Data protections are

[40] 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.

provided through non-disclosure agreements with third parties[41] to ensure security provisions are in place and can be contractually enforced.

These provisions apply 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).

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. Given the importance of demand response to the California grid and as a means to enable compensation to support utility customers' contributions, the California Public Utilities Commission has the authority to classify demand response performance analysis as a "primary purpose," thereby clarifying that load-serving entities can provide non-participant data as it directly supports grid operations.

Recurve supports whatever process is most efficient to enable the flow of data for comparison group analysis. This may be a centralized data repository where the necessary non-participant data can be stored to minimize transaction costs of building data pipelines from multiple entities. The CEC's centralized data repository may be a resource for the future. Key considerations are the frequency of the data transfers from the utilities and the specification of the data held by the CEC. It may also be a decentralized model with secure and robust standardized data pipelines with the largest utilities to enable the analysis.

It is essential that entities entrusted with handling non-participant data have proper security credentials and can share data securely without undue exposure to non-participant data. Most non-utility demand response providers are not interested in securing the necessary infrastructure to handle non-participant data. Third parties doing the analysis are the more likely candidates to handle these data. Using Recurve as an example, we hold a SOC2 security certification and have embedded differential privacy algorithms in the outputs to protect non-participant privacy, as demonstrated in this report.

VIII. Conclusions

The ISO Tariff method and the Matched Control Group and Weather Matching Baseline methods are all designed and intended to 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.

By incorporating both a comparison group and weather-normalized hourly modeling applied equivalently to treatment and comparison customers, the FLEXmeter approach blends the

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

strengths of both approved tariff methods. The model is sensitive to weather and the comparison group provides adjustments for exogenous effects and directional model error (for example nonlinearity in energy response to temperature at very high temperatures). After careful review, we believe the FLEXmeter approach to be aligned with the content and intent of the ISO Tariff.

As mentioned, non-participant data availability and data privacy concerns have proven barriers to the implementation of comparison groups in demand response measurement. However, data privacy methods and security practices have advanced in recent years. This has been demonstrated here with differential privacy protections adopted to protect non-participant data used for the comparison group. Similarly, privacy protections and rigorous data security practices can be put in place going forward to minimize the risk of customer re-identification.

Using site-level matching to produce comparison groups, incorporating meter-level hourly models, specifying a normalized difference of differences adjustment, enabling hourly savings assessment for both event and non-event periods, and fully specifying methods rooted in open source code are all important enhancements to the tariff's current description of control group methods. These steps and the demonstration of application in this study are hoped to provide parties the confidence to move forward with a standardized approach.

Appendix A: CAISO Tariff and FLEXmeter Methods

Introduction

This appendix assesses the compliance of theFLEXmeter methods with the existing ISO Tariff governing demand response.[42] If the methods proposed here are compatible and deemed compliant with the ISO Tariff, they can be implemented to measure the impacts of the demand response resources called upon to meet California's grid reliability needs.

This document will cover the following key points to demonstrate how comparison group methods using standardized CalTRACK and GRIDmeter methods and code comport with the control group requirements described in the ISO Tariff. Topics covered include:

- Origin of Existing ISO Tariff regarding comparison groups
- Origin of proposed standardized comparison group method
- Detailed methodological comparison, including a point by point accounting of tariff requirements and methods:
    - Comparison groups and statistical equivalence
    - Geographic and weather similarity
    - Comparison group sample size
    - Baselining
    - Savings calculation
    - Treatment and comparison group evolution
    - Verification, auditability, and reproducibility
- Findings & Conclusion

Existing Comparison Group Requirements in the ISO Tariff

In 2017 CAISO adopted a control group methodology for calculating savings for Proxy Demand Resources or Reliability Demand Response Resources. The method is outlined in the ISO Tariff and is further articulated in the Demand Response Business Practice Manual Section 5.3. A baseline accuracy working group (BAWG) was foundational to developing the control group method in the ISO Tariff and the final positions of that working group are included in the California ISO Baseline Accuracy Work Group Proposal completed in June of 2017 by Nexant.

[42] Section 4.13.4.3 of the CAISO tariff specifies the Control Group Methodology. More detail is provided in Section 5.3 of the Demand Response Business Practice Manual.

Methods Comparison

In the discussion that follows, the ISO Tariff is referred to as "tariff", the Business Practice Manual for Demand Response is abbreviated to "BPM." The terms "treatment group" and "participating customers" are used interchangeably and align with the concept of a demand response "resource." The DOE report "Comparison Groups for the COVID Era and Beyond", referenced above, is referred to as the "DOE Report." Finally, the term "comparison pool" refers to the broader population data from which a "comparison group" is selected. The "comparison group" is used synonymously with the term "control group" in the tariff.

Recurve has divided the tariff and BPM language into distinct sections in order to provide clear comparisons with the proposed FLEXmeter methods.

1. **Comparison groups and statistical equivalence**

    a. **Tariff language and interpretation**

    The tariff calls for a "randomized control group of End Users." The tariff goes on to state that "the control group must have nearly identical Demand patterns in aggregate as the Proxy Demand Resources or Reliability Demand Response Resources…[T]he control group must statistically demonstrate (i) lack of bias and (ii) sufficient statistical precision with (iii) sufficient confidence."

    Given this clear direction for statistical equivalence, the term "randomized" does not imply a random selection of non-participating customers. Instead, this term can be interpreted to refer to one of two things:

    1. A portion of a demand response provider's resources that are randomly withheld to serve as the control group for an event (i.e. a randomized control trial or RCT),

    or

    2. A group of non-participants was selected on the basis of statistical equivalence to the treatment group.

    b. **FLEXmeter alignment**

    The proposed FLEXmeter methods are fully aligned with these requirements.

    c. **Discussion**

    Though performing savings calculations and comparison group adjustments within an RCT is straightforward, Recurve anticipates that most demand response providers will not wish to take this step since it introduces program complexity and diminishes the available resource for any given event. The proposed FLEXmeter methods, therefore, focus on sampling from a pool of non-participant meters. The GRIDmeter code

enables sampling based on individual meter matching algorithms or advanced stratification techniques, which are detailed in the DOE report. Both of these methods minimize the load shape discrepancy between treatment and comparison groups across the entire distribution of treatment customers. This approach ensures an optimal comparison group representation and, given a sufficient comparison pool, statistical equivalence for not just the average treatment meter load profile, but across the entire range of treatment group meters.

Recurve did not find definitions or thresholds for statistical equivalence testing in the tariff or the BPM. The BAWG report does provide information and examples on how to test for statistical equivalence on the basis of aggregated treatment and comparison group load profiles and focuses on the hours of 12 pm - 9 pm for this testing.

Recurve recommends statistical equivalence metrics that focus on the distributions of treatment and comparison groups instead of the aggregate. Doing so will help ensure that average load profiles do not mask large discrepancies on a more granular level. GRIDmeter methods produce optimal samples gauging the full range of participating customers and explicitly test for distribution equivalence. Recurve uses the Kolmogorov-Smirnov test[43] (distribution) as well as the T-test[44] (average) to verify the sampled distribution is equivalent to the treatment group.

Recurve proposes applying the T-test and KS-test to the distribution of each hour of an average weekday load shape in the baseline period. If needed this process can be completed for the average baseline period weekend daily load shape.

2. Geographic, weather, and sector similarity

a. Tariff language and interpretation

The tariff states, "The control group must be geographically similar to the Proxy Demand Resources or Reliability Demand Response Resources such that they experience the same weather patterns and grid conditions." Further, the BPM states, "The Control group cannot combine/co-mingle Residential and Non-Residential."

b. FLEXmeter alignment

The proposed FLEXmeter methods are fully aligned with these requirements.

c. Discussion

As the tariff implies, it is critical that comparison groups are generated from customers that experience similar weather. Drawing from similar geographic locations can also help ensure that treatment and comparison customers are experiencing

[43] https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

[44] https://www.britannica.com/science/Students-t-test

similar exogenous factors such as public safety power shutoffs and public messaging encouraging conservation during extreme weather and grid events. The CAISO and demand response providers can help inform where treatment populations should be measured independently. For each of the geographically distinct treatment groups, an independent comparison group assignment should be drawn and savings calculations carried out. Recurve also notes that comparison group customers should always be drawn from the same sector (residential, commercial, industrial, etc.) as the treatment group, in alignment with the BPM..

3. **Comparison group sample size**

    a. **Tariff language and interpretation**

    The tariff states, "The control group must consist of 150 distinct End Users or more."

    b. **FLEXmeter alignment**

    The proposed FLEXmeter methods are fully aligned with these requirements.

    c. **Discussion**

    Based on experience testing comparison groups to account for the hourly energy impacts of COVID, Recurve would recommend a minimum sample size of significantly more than 150 meters wherever possible. In cases in which fewer than 500 comparison group meters are available that meet the requirements described above, Recurve would recommend the M&V provider be able to issue a rationale and recommendation to CAISO to relax the statistical equivalence thresholds in order to expand the sample size and limit the risk associated with a small comparison group.

4. **Baselining**

    a. **Tariff language and interpretation**

    The tariff contains several elements relevant to constructing a baseline on which comparison groups are to be selected and equivalence assessed. The tariff states "[T]o validate the control group, Meter Data of the control group and the Proxy Demand Resources or Reliability Demand Response Resources from the previous seventy-five (75) days must be evaluated, excluding days where the Proxy Demand Resources or Reliability Demand Response Resources provided Demand Response Services or participated in a utility demand response program. Using the most recent days, at least twenty (20) eligible days of Meter Data must be used for validation. From these days, an average of the hourly load profile from 12 p.m. to 9 p.m. must be developed for the Proxy Demand Resources or Reliability Demand Response Resources and the control group by day and by hour."

The tariff also states, "The control group's aggregate Demand during the same Trade Date and Trading Hour(s) as the Demand Response Event, divided by the relevant number of End Users, will constitute the Customer Load Baseline."

Regarding the timing of comparison group development and validation, the tariff states that the procedure should be conducted "Prior to any Demand Response Event."

b. FLEXmeter alignment

The proposed FLEXmeter methods largely align with these requirements. Below, Recurve notes where slight improvements would be recommended.

c. Discussion

If necessary, the FLEXmeter baselining and equivalence analysis can be limited to 12 - 9 pm on the most recent 20 non-event days within a 75-day pre-event window. However, Recurve notes that utilizing non-event days that both predate and postdate an event can help ensure that treatment and comparison group meters experience and react similarly to weather patterns that are most likely to resemble the event day weather. For example, an event may occur toward the beginning of a heat wave. In this case, incorporating post-event days into the baseline and equivalence analysis ensures that weather conditions most representative of the event day are included in the comparison group selection process. Therefore, Recurve recommends relaxing the requirement that comparison groups be selected prior to the event date.

In addition, while the hours of 12 - 9 pm are likely to encompass any demand response events, Recurve believes it is also important to both differentiate between weekdays and weekends and assess all hours of a day in comparison group selection. Many customers have similar weekday vs. weekend usage patterns and many have very different weekday vs. weekend patterns. Therefore averaging across different day types is not advisable. Regarding the hourly window, taking into account all hours of the day can help ensure treatment and comparison of customer similarity including the proportion of load that can be shifted in an event. Recurve's suggested approach to utilizing the full 24-hour period is discussed in point 1 above.

5. Savings Calculation

a. Tariff language and interpretation

When utilizing a comparison group, the BPM defines the Demand Response Energy Measurement (demand responseEM) as:

> *demand responseEM = {(hourly avg of control group load data) – (hourly avg of treatment group load data)} x (#locations in treatment group) = {(total load*

of control group/# locations in control group) – (total load of treatment group/#locations in treatment group)} x #locations in treatment group

In other words, the demand response event savings are determined by subtracting the average treatment customer usage from the average comparison group usage (presumably for the event hours only) and then multiplying by the number of treatment group customers.

b. FLEXmeter alignment

The FLEXmeter approach combines a weather-based modeling approach and comparison group adjustment step to minimize both error and bias in savings estimates. At its core, the BPM savings calculation is similar to these proposed methods. However, Recurve finds that the BPM definition would benefit from greater detail and that certain elements of the savings calculation can be made more stable via the FLEXmeter approach.

c. Discussion

The BPM savings calculation takes the direct difference of averages between the treatment and comparison group. This puts enormous pressure on the comparison group to be a near-perfect match in both magnitude and shape to the treatment group. While the FLEXmeter methods produce optimal comparison groups for both these parameters, there is always some residual bias that exists between treatment and comparison groups, no matter the sampling process. By taking a direct difference between treatment and comparison, any discrepancy between these groups will manifest directly as bias in the savings calculation.

Recurve instead recommends conducting savings calculations via a percent difference of differences calculation in which the CalTRACK model[45] is used to produce a prediction (counterfactual) for the hourly energy consumption of both treatment and comparison customers during event hours.

In this approach, the CalTRACK model generates an hourly counterfactual for both treatment and comparison groups. Subtracting the observed meter readings from the counterfactual for each group and then dividing by the counterfactual produces the treatment and comparison percentage differences. The comparison group % Diff is then subtracted from the treatment group % Diff and the resulting % Difference of Differences is then multiplied by the treatment group counterfactual to arrive at hourly savings. This methodology is described in detail in Chapter 4 of the DOE report and the relevant equations are reproduced below. The subscript i indicates a summation over individual meters in the respective groups:

[45] http://docs.caltrack.org/en/latest/methods.html

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

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

$$\%\,Diff\ of\ Diff_{i} = \%\,Diff_{Treatment,i} - \%\,Diff_{Comparison,i}$$

$$Savings = \%\,Diff\ of\ Diff_{i} \times Counterfactual_{Treatment,i}$$

The BPM methods do not specify whether the savings calculations should be conducted on an hourly basis during the event period, summed across the event period, or over some other period and granularity. For maximum visibility into the resource performance, Recurve recommends conducting the calculation on an hourly basis across all hours of the day. While payments may be made exclusively on the basis of event period performance, determining impacts for all hours provides an important check and can provide insight into important effects such as snapback.

Finally, the BPM definition utilizes the total hourly consumption of the treatment and comparison groups. In Recurve's experience, procedures need to be in place to ensure data are cleaned for spurious meter readings and that a standardized procedure be followed to interpolate between null or zero values. The CalTRACK methods detail (and the OpenEEmeter codebase operationalizes) data cleaning and missing data interpolation.

6. Treatment and comparison group evolution

a. Tariff language and interpretation

The tariff states, "For Proxy Demand Resources or Reliability Demand Response Resources whose number of End Users have not changed by more than ten (10) percent in the prior month, the control group must be re-validated every other month. For Proxy Demand Resources or Reliability Demand Response Resources whose number of End Users have changed by more than ten (10) percent in the prior month, control groups must continue to be revalidated monthly."

b. FLEXmeter alignment

The proposed FLEXmeter methods are fully aligned with these requirements.

c. Discussion

For this step, it will be important for demand response providers to provide timely participation data. Depending on the availability and reliability of data it may be prudent to re-sample and re-validate comparison groups every month.

7. Verification, auditability, and reproducibility

a. Tariff language and interpretation

The tariff states, "[A]ll Demand Response Event calculations are subject to CAISO audit for three (3) years from the date of Demand Response Event. All results must be reproducible, including underlying interval data, randomization, validation, bias, confidence, precision, and analysis."

b. FLEXmeter alignment

The proposed FLEXmeter methods are fully aligned with these requirements.

c. Discussion

N/A

Appendix B: Recommended Standardized Data Specification

Without clean and reliable data, highly specified and sophisticated methods will not yield reliable or timely measurements. In working with demand response providers, CCAs, and IOUs, Recurve ingested data with many formats and at different levels of quality and cleanliness. In some instances, data issues hindered our ability to produce reliable results in time to include in this study, so incomplete or poor-quality datasets were omitted. Having a data formatting and sufficiency standard in place can ensure that all the data needed to perform comparison group matching and savings calculations are available and help prevent errors and limit delays or excess evaluation spending.

Recurve encountered several data issues in the course of this work that may be avoidable in the future. These include:

- Lack of Solar PV status metadata - When a dataset lacks an explicit flag for solar PV customers, Recurve often assigns solar by the presence of negative meter readings. However, not all solar customers feed energy back to the grid and are therefore missed by this approach.
- Lack of sector metadata - In some instances, data was not provided to catalog customers as residential, commercial, industrial, etc. Without this information, Recurve relied on common residential and non-residential rate codes to assign the correct sector. However, rate code data is often not definitive and may not always be up-to-date or complete.
- Lack of LSE assignment - Recurve often found that demand response provider participant data did not contain an identification of LSE (CCA, IOU, etc.), which can lead to a risk that customers are miscategorized and cause errors in assessing load impacts by service territory.
- Lack of sufficient location information. Customers should be linked to the most granular location data available. In some cases, Recurve identified zip codes in customer data that appear to be non-existent. In these cases, Recurve mapped to climate zones and weather stations in territories expected to be nearby.
- In some cases, common customer identifiers (SAID) were not present in demand response provider participant data. Without this information, these customers cannot be reliably removed from comparison pools, and there is a risk that participants are sampled into comparison groups.
- Mapping of customers to appropriate program levels. Recurve received feedback that some programs operate at the sub-lap level. Additionally, DRPs will often deploy resources in groups and for specific programs. Program, sub-lap, and any other information required for customer stratification should be included at the event participation level.

Application of the recommended data specification outlined below may help prevent these issues from arising again.

Data Requirements

- Fifteen (15) days before the first event of the event season under analysis through 15 days after the last event of the analysis.
- Weekday and weekend data must be included.
- Minimum granularity: hourly.

Additional Meter Metadata Fields Required from demand response providers

Field	Category	Reason
Date_of_installed_solar | is_solar_present	Match stratification	Necessary for comparison group matching. Otherwise, solar meters may get matched to non-solar meters due to similar-looking load-shapes, but will not respond similarly to weather or demand response events.
Demand response Lat/Lng | zip-code+ climate zone	Match stratification	Accurate location information is critical for both matching to a specific weather station in order to get temperature data for the model, as well as matching to only comparison groups within the same climate zone.
source_lse	Match assignment	This is important for matching the meter to its appropriate comparison pool
SAID	Match assignment	This is important for making sure that we remove this meter from the comparison pool. Otherwise, there is a possibility that the meter will match to itself.
Sector | Rate code with mapping	Match stratification	This is important to make sure residential meters are only assigned to other residential meters. Sometimes this is done by assigning different rate codes to a sector. However, it is important

		that we are able to get a mapping from the source LSE about which rate codes match with a given sector.
Optional: NAICS Code (if non-residential)	Match stratification	This is important to match similar non-residential meters together. Sometimes we will aggregate to the 2-digit NAICS code or other NAICS groups, depending on group size.

Demand Response Provider (DRP) Data Format

Data Guide

The data transfer from demand response providers should match the below format. A single file containing all meter and customer information and another detailing all events each customer participated in along with event details are minimum requirements.

This data format revolves around 5 core types of data:

1) Sites - Physical locations where a building or meter exists.
2) Events - The datetime and duration of demand response calls.
3) Event Participation - A list of events each site/meter participated in.
4) Meters - Devices measuring and recording usage data. (OPTIONAL)
5) Meter Trace Records - A time series of energy consumption at the meter. (OPTIONAL)

Each of these 5 file types has required columns.

Sites

A site is the location of a building or project, and carries address fields, geocoded latitude/longitude coordinates, and other metadata (such as building age). The site must contain a link to the LSE meters. Below is an example of a table with the minimum viable information.

Sites are used to link projects with meters, to find appropriate sources of weather data, and to tie outcomes to the grid.

Field Name	Required	Type	Unit	Example
SiteID	Required	UTF-8 String		res-188034
Latitude	Required	Numeric		37.894677
Longitude	Required	Numeric		-122.149389
Zipcode	Optional	Numeric		94608
Link to LSE	Required	UTF-8 String		5198724012AB (SAID, SPID, etc)
Name of LSE	Optional	UTF-8 String		SCE

Sector	Optional	UTF-8 String		Residential
NAICS Code	Optional	UTF-8 String		311352

Events

An event denotes a call for demand response as measured by the program. Each event must have an EventID, a datetime (the date and time the event was called), and a duration (the length of time the event lasted). If preferred, an event start time and end time can be provided.

Field Name	**Required**	**Type**	**Unit**	**Example**
EventID	Required	UTF-8 String		Event_001
EventName (type of event)	Optional	UTF-8 String		august_blackouts
EventStart	Required	ISO 8601	datetime	2020-08-14 16:00:00
Duration	Required	ISO 8601	Timedelta (hours)	4:00
EventEnd	Optional	ISO 8601	datetime	2020-08-14 20:00:00

Delivery should be a single csv that lists each event. Ex:

EventID	EventName	EventStart	Duration
Event_001	august_blackouts	2020-08-14 16:00:00	4:00
Event_002	september_sunsets	2020-09-24 17:00:00	2:00
Event_003	october_oopsies	2020-10-05 16:00:00	3:00

Meter Participation

There must be a mapping between every event and each participant. The minimum viable information is a csv for each event with the meterID of every participant. The preferred delivery is a table of every meter as a row and each event as a column. If a meter participated in a given event, the cell value will be true. In the example below meter4321 participated in the first and third event only. Meter 1234 participated in only event_003 while meter5555 participated in every event.

Preferred delivery format:

MeterID	**EventID**
meter4321	Event_001
meter4321	Event_003
meter1234	Event_003
meter5555	Event_001
meter5555	Event_002
meter5555	Event_003

An alternative format is a single csv for each event, three in this example. The CSV for Event_001 would be titled Event_001.csv and should contain the event name in the first cell. It contains meter4321 and meter5555 as these are the two meters that participate in the event.

Event_001
meter4321
meter5555

The minimum format for a single table is two columns, one with meterIDs and the other with EventIDs.

MeterID	Required	UTF-8 String		meter4321
EventID	Required	UTF-8 String		Event_001

Single csv with two columns

Meter ID	**Event_001**	**Event_002**	**Event_003**
meter4321	TRUE	FALSE	TRUE
meter1234	FALSE	FALSE	TRUE
meter5555	TRUE	TRUE	TRUE

Meters

A meter is a representation of a physical device for measuring energy consumption.

Defining a meter is extremely simple and requires only 3 elements (1 optional):

1) Meter ID - A unique identifier for a meter.
2) Site ID - A site representing where the meter is located.
3) Meter Type - Type of usage being measured, either "electric" or "gas".
4) Net Metering Flag - If not provided, assumed to be "FALSE".

Field Name	**Required**	**Type**	**Unit**	**Example**
MeterID	Required	UTF-8 String		meter4321
SiteID	Required	UTF-8 String		PremID
MeterType	Required	UTF-8 String		electric/gas
NetMeteringFlag	Optional	Boolean		True/False

Meter Trace Records

A meter trace is a time series of energy consumption at the meter, typically given in raw form as automated reads with daily, hourly, or sub-hourly frequency. Each record needs a start and end datetime.

Field Name	Required	Type	Unit	Example
MeterID	Required	UTF-8 String		meter4321
Unit	Required	UTF-8 String		electric/gas
Interval	Optional	UTF-8 String		billing-monthly/daily/hourly
Start	Required	ISO 8601	datetime	2020-08-01 00:00:00
End	Required	ISO 8601	datetime	2020-08-01 01:00:00
Value	Required	Numeric	kWh/Therms	0.12

Sites/Events/Meters Mapping

The 3-way combination of a site, event, and meter allows analysis to be performed.

Data Format for LSEs

Data should be provided 15 days before the first event of the event season under analysis through 15 days after the last event of the analysis. In some cases, program administrators may have demand response participant information in addition to AMI data. In this case, both participant and non-participant program and AMI data should be delivered as detailed below. As for all options, the non-participant data will be used to formulate comparison groups.

Minimum data requirements

For each customer, the following should be provided:

- Customer, Premise, Account, and Meter identifiers (SAID, SPID, etc) ○ If using Option 2, identifiers may be replaced with pseudo identifiers for non-participants only

- Sector (residential/commercial)
- NAICS code (if available)
- Program participation, if any
- Location, as any of the following, for mapping to weather station
    - Latitude/Longitude (preferred)
    - ZIP code (preferred)
    - Address

Appendix C: Detailed Methods Specification

I. Data Formatting

Treatment group and comparison pool data should be formatted according to the specification in **Appendix B: Recommended Standardized Data Specification**. The remainder of this methods specification assumes data meets these formatting standards.

II. Apply Initial Eligibility Criteria:

i. Remove treatment and comparison pool meters that do not have at least 85% of all possible meter readings populated in the Sampling Period (defined below).

ii. Remove treatment and comparison pool meters that do not have at least one meter reading present in the sampling period across all 168 hours of a week.

iii. Filter the comparison pool to remove any possible treatment meters or known meters participating in other demand response programs during the baseline or measurement periods.

III. Comparison Group Sampling

A. Establish treatment groups and comparison pools

i. Group by category: A treatment group should be largely homogeneous. It should consist of customers with the same:

1. Sector (Residential or Non-Residential)
2. Climate Zone (or Climate Region if outside CA)
3. Solar PV Status

ii. For each combination of these three variables the associated comparison pools must match all characteristics.

B. Generate sampling period modeled average weekly load shapes

i. Generate a Sampling Period CalTRACK 2.0 Hourly Model for all treatment meters.

1. Sampling Period consists of a 45-day window leading to the first event to be measured.
2. The Sampling Period must end no later than 1 week prior to an event being measured. This means that the same sampling period is valid for a maximum of a 7-day window where events are to be measured.
3. For each treatment group meter:

a. All days in the Sampling Period in which a treatment meter was subject to an event are "blacked out," meaning excluded from the CalTRACK hourly model development.

b. Apply eligibility criteria:

    i. Remove treatment meters with Sampling Period CalTRACK hourly Coefficient of Variation of the Root-Mean-Squared Error (CVRMSE) outside the range of -2 to 3. For more information on CVRMSE see CalTRACK methods 4.3.2.2.

c. Compute the average 168-hour weekly model load shape. This is done by averaging the Sampling Period model for each hour of the week. Despite blacking out event days in the Sampling Period, the CalTRACK 2.0 model will still generate modeled usage during those days. These days are included in the calculation of the average weekly modeled load shape.

4. For each Comparison Pool meter:

    a. All days in the Sampling Period in which the Treatment Group was subject to an event are "blacked out," meaning excluded from the CalTRACK hourly model development.

    b. Apply eligibility criteria:

        i. Remove comparison pool meters with Sampling Period CalTRACK hourly Coefficient of Variation of the Root-Mean-Squared Error (CVRMSE) outside the range of -2 to 3. For more information on CVRMSE see CalTRACK methods 4.3.2.2.

    c. Compute the average 168-hour weekly model load shape. This is done by averaging the Sampling Period model for each hour of the week. Despite blacking out event days in the Sampling Period, the CalTRACK 2.0 model will still generate modeled usage during those days. These days are included in the calculation of the average weekly modeled load shape.

C. Comparison group selection via GRIDmeter site based matching[46]

[46] For more information on Site Based Matching see Chapter 3 of the report *Comparison Groups for the COVID Era and Beyond*

i. For every treatment meter, calculate the sum of squares of the differences in the modeled weekly load shape of every comparison pool meter

ii. For every treatment meter, select the comparison pool meter with the lowest sum of squares computed in step i.

iii. Check for and remove duplicates sampled from the comparison pool in step ii.

iv. Repeat this process until the desired number of comparison meters has been sampled. This sample will serve as the **comparison group** for the calculation of load impacts covered below.

v. The following recommendations should guide comparison group sizing:

1. For participant groups between 15 to 1,000, Recurve recommends utilizing a comparison pool with at least a factor of 20 more meters than the participant group and sampling at least 4 comparison meters per treatment meter.
2. For participant groups of 1,000 to 4,000 meters, Recurve recommends sampling from a comparison pool of at least 20,000 meters such that at least 4,000 comparison group meters are selected.
3. For participant groups larger than 4,000 meters, if a comparison pool of at least 8 times the size of the participant group is available, Recurve recommends sampling at a 2:1 ratio. If a comparison group of between 4 to 8 times the size of the participant group is available, Recurve recommends sampling at a 1:1 ratio.
4. For any comparison pools smaller than 4 times the size of a participant group, Recurve recommends random sampling from the treatment group such that a 4:1 ratio is obtained and then sampling at a 1:1 ratio. In these cases, savings should be scaled to the full population of participants.

IV. Event Day Treatment Group Load Impacts Calculations via CalTRACK Hourly Modeling and GRIDmeter % Difference of Differences[47]

A. Establish baseline period for savings calculation

[47] For more information on the % Difference of Differences approach see Chapter 4 of the report *Comparison Groups for the COVID Era and Beyond*

i. For the measurement of event day savings the baseline period consists of the 45-days immediately preceding the event day and the 15 days post. Each event day requires respecification of the baseline period.

ii. All days in the baseline period in which a treatment meter was subject to an event are "blacked out," meaning excluded from the CalTRACK hourly model development.

B. For all treatment group meters calculate CalTRACK 2.0 hourly event day counterfactual[48]

C. Replace event day null hourly observed meter readings with 0

D. Eliminate hours in which Equation 1 (fractional between observed and counterfactual hourly usage) was outside the bounds of -10 to 10.

E. Aggregate hourly treatment group observed meter readings and counterfactuals by summing these quantities over all treatment group meters.

F. Calculate treatment group % Diff as follows:

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

Where the subscript *i* refers to the hour of day.

V. Event Day Comparison Group Load Impacts Calculations

A. Establish baseline period for savings calculation

i. For the measurement of event day savings the baseline period consists of the 45-days immediately preceding the event day and the 15 days post. Each event day requires respecification of the baseline period.

ii. All days in the baseline period in which a treatment meter was subject to an event are "blacked out," meaning excluded from the CalTRACK hourly model development.

B. For all comparison group meters calculate CalTRACK 2.0 Hourly event day counterfactual

C. Aggregate hourly comparison group observed meter readings and counterfactuals by summing these quantities over all comparison group meters.

D. Calculate comparison group % Diff as follows:

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

[48] The meter-level counterfactual is the CalTRACk 2.0 model prediction for event day usage based on the event day time of week and temperature

Where the subscript *i* refers to the hour of day.

VI. Calculate Hourly Event Day Load Impacts as follows:

$$\%Diff\ of\ Diff_{i} = \%Diff_{Treatment,i} - \%Diff_{Comparison,i}$$

$$Load\ Impact_{i} = \%Diff\ of\ Diff_{i} \times Counterfactual_{Treatment,i}$$

VII. Adjust Load Impacts for Full Population Results or Outliers if Needed:

A. Outliers should be isolated and analyzed separately from the population.

i. Additional analysis to specify outlier criteria beyond those detailed above is planned.

B. Scale the hourly savings to account for any customer eliminated due to initial sampling or eligibility and outlier criteria applied.

i. Adjust hourly savings to ensure proportional sampling if initial sampling does not reflect full program participation