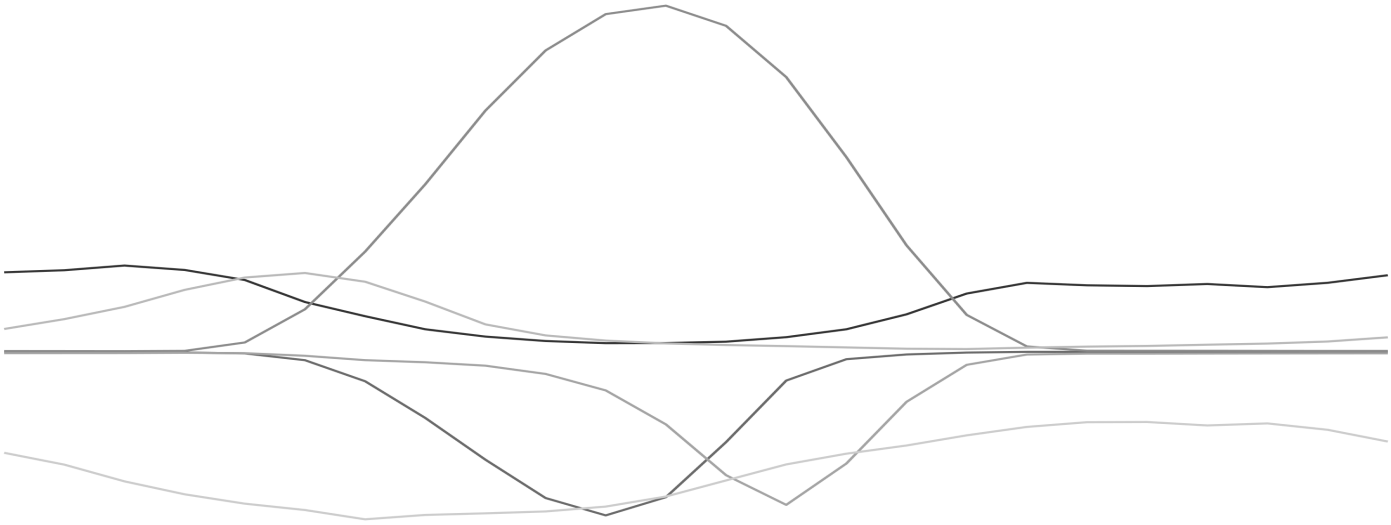
2019 Load Impact Evaluation for Leapfrog Power Final Report

Public Version



May 29, 2020

Prepared for Leapfrog Power, Inc

Prepared by
Sam Borgeson
Phil Price
Convergence Data Analytics, LLC

Stefanie Wayland
Grounded Analytics, LLC

Mary Sutter
Grounded Research and Consulting, LLC

Convergence Data Analytics

CALMAC ID: LPF0001.01

Cover image: Loads and generation from various meters in a residential PV and battery system.

Contents

Figures

Tables

Executive Summary

This report presents results of a load impact analysis of Leapfrog Power, Inc's (Leap's) aggregated Demand Response (DR) capabilities for residential and non-residential sectors. Leap is a DR aggregation company offering a market exchange for grid flexibility services. Specifically, Leap has a software platform that allows their customers to specify the conditions under which their resource is available in the wholesale market to be dispatched and deliver energy. Customers "bid in" their resources (i.e., set the price threshold at which they wish to be dispatched) and Leap triggers the customer DR controls when the threshold is reached.[1] Leap customers are tracked at the SubLAP level and, in 2019, virtually all (98.5%) were dispatched through automated DR controls.

During 2019, the year this report is based on, Leap called 26 test events that took place from April to November 2019 to establish an empirical record of their DR capacity. Their August test events dispatched all customers enrolled at the time at least once (only one test event was called after August and that event only included one large customer). They did not participate in any CAISO called events. This report is based on the data from those load reduction events, combined with Leap's customer enrollment as of the end of 2019 and Leap's enrollment forecast for future years. Leap's resource has been growing quickly. By the end of 2019, the set of meters dispatched during the test events comprised just 52% of enrolled meters - customers Leap enrolled after August were not included in the test events. Leap customers with residential battery storage were also not included in these test events. CDA obtained data from Leap's solar and storage partner to simulate test events for the residential battery loads and provide ex post estimates for use within the ex ante prediction.

In this document we present:

1. Ex post load impact estimates for the year 2019 (PY19)
2. Ex ante prediction of Leap events for program years 2020-2030

Within these analyses, we examined impacts across geography and by customer segments. This included findings for:

- *Local capacity areas (LCAs).* There are ten California Independent System Operator (CAISO) LCAs[2] in California, spanning a great deal of geographic/climatic variability: **PG&E territory**: Greater Bay, Greater Fresno, Humboldt, Kern, North Coast / North Bay, Sierra, Stockton; **SCE territory**: Big Creek / Ventura, LA Basin; **SDG&E territory:** San Diego. The LCA analysis provides insights on the magnitude of available capacity from events in each geographic area.[3] Ex post analysis included all LCAs for both residential and nonresidential sectors except for Kern (where only residential meters were available for ex post analysis) and San Diego (Leap had no customers who participated in events in San Diego in 2019, however, some of the residential battery sample were in SDG&E territory). The ex ante analysis included customers in all LCAs based on the Leap forecast.
- *Customer type:* Leap customers provide resources from various loads in the nonresidential and residential sector. Nonresidential loads include pumping, air conditioning, electric vehicles,

[1] This is a Locational Marginal Price (LMP) price threshold at the Sub Load Aggregation Point (sub-LAP) level.

[2] See http://www.caiso.com/informed/Pages/StakeholderProcesses/LocalCapacityRequirementsProcess.aspx for more details on the CAISO local capacity requirements process.

[3] While the natural geography for DRAM resources is the sub-LAP, there were too few customers per sub-LAP for reliable estimates of ex post impact. Therefore, CDA analyzed data by LCA.

battery and thermal storage and other.[4] Residential loads are from air conditioning and battery storage (connected to PV systems). This report covers all loads within these two customer types.

All of the above categories inform our understanding of ex post results and are explanatory variables in our ex ante model. (The residential ex ante model was trained using sub-LAPs, not LCAs, but they directly correspond with one another. The nonresidential ex ante model was trained by individual meter).[5]

The test events examined for this report each lasted 2 hours, with all event timing within the RA window, which extends from 4 pm to 9pm. In total, data from 26 events, spanning 18 days of 2019, was provided by Leap to CDA for evaluation. Nonresidential test events occurred from April to November while residential events occurred in June and August for air conditioning and simulated test events for residential batteries across May to December. Ex ante results are based on model-predicted event performance across the full RA window.

Ex Post Load Impacts

Leap obtains load impacts from both the nonresidential and the residential sectors. Nonresidential load impacts are from pumping, air conditioning, electric vehicles, battery storage (large and small storage), thermal storage, and other. Residential load impacts occur from air conditioning and battery storage (associated with PV systems). Leap added 48% of the customers present in its end-of-year portfolio after it had concluded its testing for 2019. Leap's full end-of-2019 resource, using per-meter impacts from the 52% of customers included in test events and enrollment as of the end of 2019, provides 30.5 MW. (Table 1)

Table 1. Ex post impacts for Leap's 2019 full resource

Sector Loads	Ex Post Mean Event Temperature (F)	Ex Post Mean Impact (kW)	Ex Post Mean Baseline (kW)	Ex Post Percent Impact (%)	Full Resource Enrollment Count (meters)	Full Resource Total Impact (MW)	Std. Err (MW)	90% Conf. Interval (low, high) (MW)
Nonresidential	82.5	31.3	93.5	34	848	26.55	4.02	(19.94. 33.16)
Residential	86.4	0.5	1.6	33	7,098	3.9	0.16	(3.63, 4.16)
Total	***85.2***	***3.8***	***29.6***	***13***	***7,946***	***30.45***	***1.32***	***(28.27, 32.62)***

Table 1 and many of the other tables and figures in this report include 90% confidence intervals. These confidence intervals are calculated separately for each sector and the total. Confidence intervals are a statistical tool that help to describe uncertainty. If the analysis were repeated many times on new data, then (for a 90% confidence interval) the confidence interval should include the "true" value in 90% of the repetitions. Confidence intervals do not describe the probability that the "true" value is in a particular interval. This limitation of confidence intervals means that while they have substantial value

[4] "Other" sites are light manufacturing and cold storage facilities.

[5] Leap had no customers in the SDG&E territory in 2019 and so is not included in any ex post analyses except for the simulated residential battery estimates. However, Leap has 2020 DRAM obligations and has been enrolling customers in the SDG&E LCA area, so the report includes ex ante analyses for this LCA.

as measures of uncertainty, they can be easily misinterpreted.[6] We include a deeper discussion of the sources of uncertainty in ex post and ex ante estimates in later sections of the report.

One product of the ex post analysis is a 'table generator' spreadsheet that allows the user to select different combinations of events, utility area, LCA and customer type. The ex post table generator includes calculations of "Typical" events. "Typical" estimates are defined by the Load Impact Protocol 8 as averages across events. A "typical" event average is useful when the entire population is dispatched for an event. However, Leap's customers are not dispatched as a full resource (e.g., Leap did not dispatch all customers for an event), so the variation between events is extremely large because entirely different groups of customer meters were dispatched for many test events. Leap dispatched their existing customers in small batches across their test events, but no single event could be said to represent their "Full Resource." For example, some events had 1 participant, while others had over 6,000, but did not feature contributions from all load types. Therefore, the "Typical Event" summary, which is specified and required by the Load Impact Protocols, is not representative of Leap's full resource. For this reason, we added another summarization within the report of Leap's ex post "Full Resource" results. This value is the estimate of what the aggregate impact would have been if all customers were dispatched at once. The "Full Resource" estimate (shown in Table 1) scales average ex post per-meter impacts by the number of meters enrolled at the end of 2019.

Below are evaluation results for 2019 test events by sector with each sector showing average impacts per-customer and by local capacity area (LCA).

Nonresidential

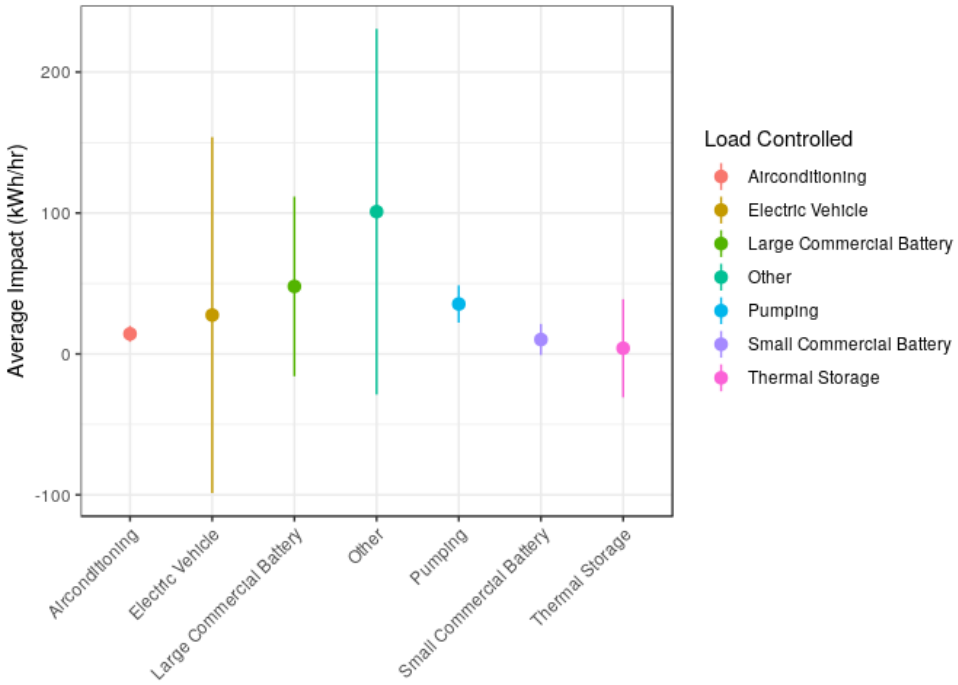
The ex post per-meter impact values that flow into the ex ante impacts (and therefore the qualifying capacity values) are the impacts by load type and location. The majority of nonresidential loads are not temperature dependent, so load type is more of an impact determinant than weather.

The average ex post per-meter impact in 2019 was 31.1 kW with moderate variation across load type. Figure 1 shows *nonresidential* per-meter impacts by load type (regardless of geographic location). The average is shown by a circle and the lines are the 90% confidence levels of the estimated impact. Meters with an "other" load (cold storage and small industrial sites) bring in the highest average impact. Electric vehicle charging shows a wide range of potential impacts, including some negative impacts. During 2019 test events, commercial electric vehicle chargers often dropped out of events after the first hour, which has been the practice under DRAM rules, and is most likely the cause for the confidence interval for two-hour events to cross zero.[7] Leap indicates that they have discussed this issue with their partner and that they are "ensuring that our partners are committed and incentivized to perform across entire test and event windows, and we therefore expect performance to increase for multi-hour tests and events in 2020 and beyond." Air conditioning, pumping, small commercial battery and thermal storage loads all brought in similar average impacts.

[6] Greenland, S., Senn, S.J., Rothman, K.J. et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31, 337–350 (2016). https://doi.org/10.1007/s10654-016-0149-3

[7] Leap indicated that one of their electric vehicle partners optimized for one-hour test events as the DRAM contract for Generic capacity stipulated that performance would be calculated for the highest hour.

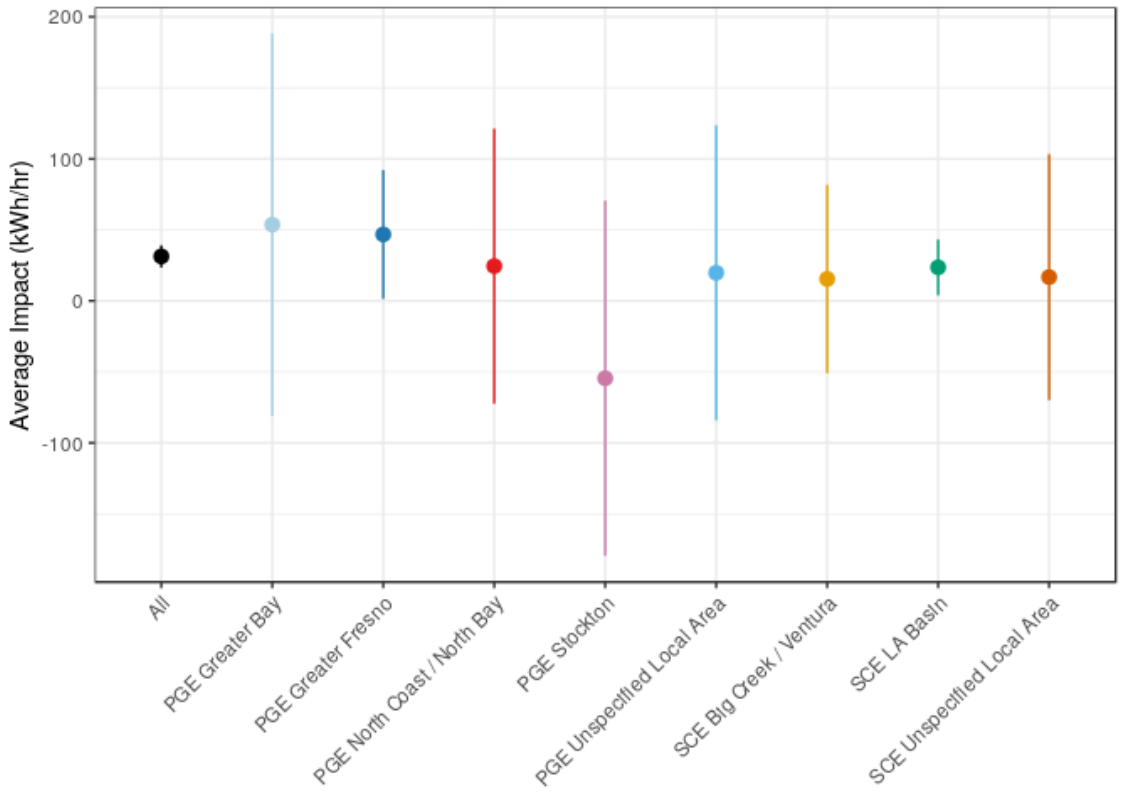
Figure 1. Nonresidential ex post per-meter impacts (average of April-November) by load type



Circles are the average across all participant meters in the monthly event. Lines are the 90% confidence intervals.

Figure 2. shows the average *nonresidential* ex post impacts by LCA. Again, the average is shown by a circle and the lines are the 90% confidence levels of the estimated impact. Impacts are a function of the load type in each area. Stockton, with a negative average load impact, included only electric vehicle loads. Greater Bay, with the highest average impact, has a high percentage of electric vehicle load but also included thermal storage and an "other" load type. The relatively small confidence interval on the "All" LCA category is the result of the evaluated impact estimates becoming more certain as more customers are averaged together.

Figure 2. Nonresidential ex post average impacts by location (Local Capacity Area)

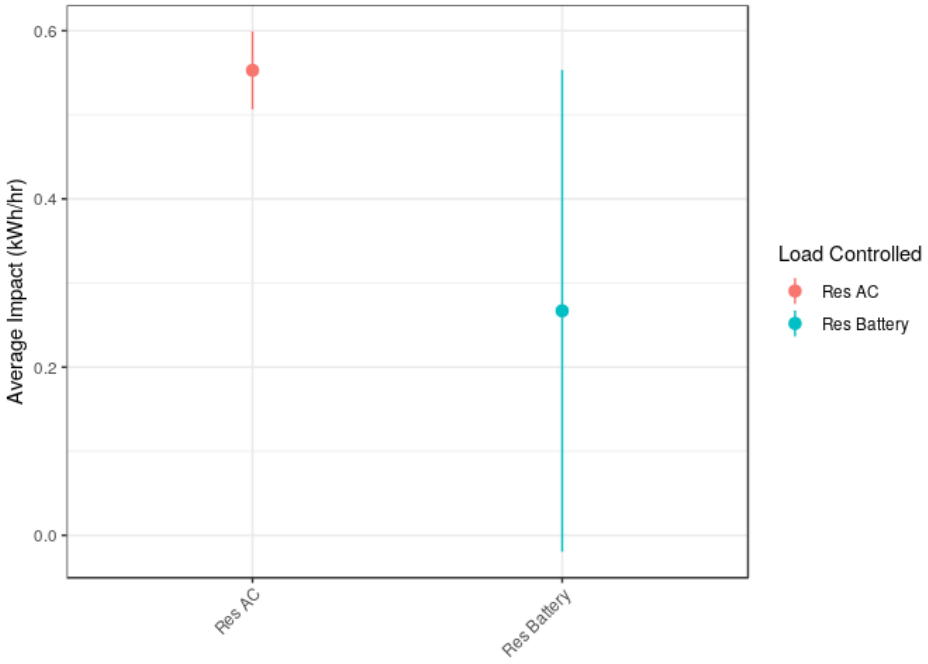


Circles are the average across all meters in the monthly event. Lines are the 90% confidence intervals.

Residential

The ex post per-meter impact values that flow into the ex ante impacts (and therefore the qualifying capacity values) are driven by the impacts by temperature (since most modeled residential loads are from air conditioning) and by location. The average air conditioner load impact (0.55 kWh/hr) is about twice the average residential battery load impact (0.27 kWh/hr). The large confidence interval for residential battery storage is largely a function of uncertainties in baseline load estimates for the small number of customers contributing to the calculation (Figure 3).

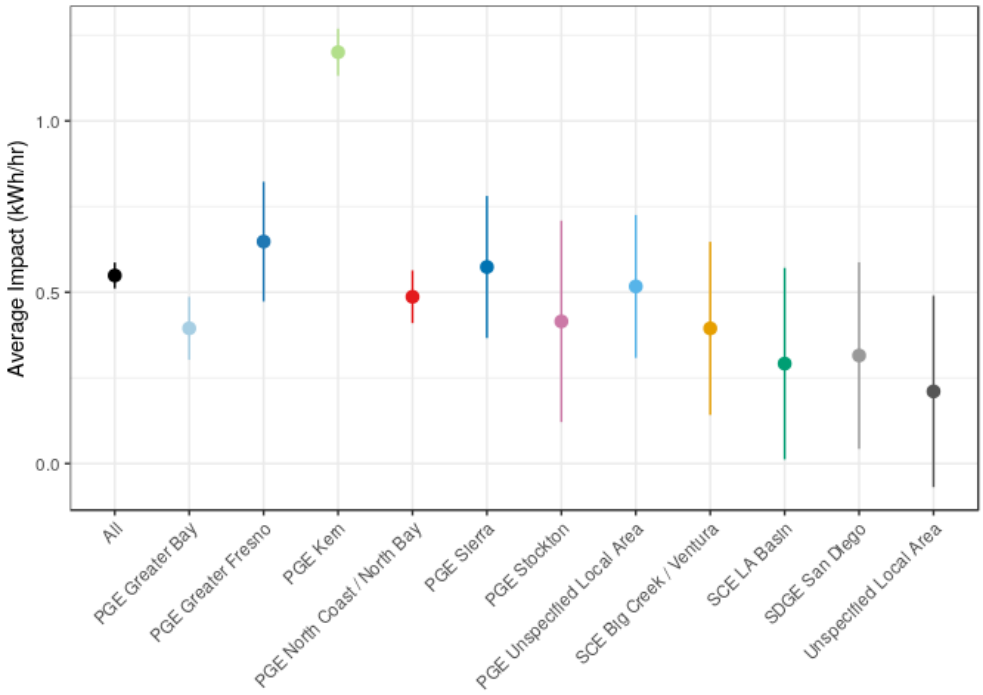
Figure 3. Residential average impacts by load type for all test events



Circles are the average across all meters in the monthly event. Lines are the 90% confidence intervals.

Figure 4 shows the average ex post *residential* impacts by LCA. Because the majority of ex post impacts come from air conditioning, the impacts are a function of temperature. Greater Bay has mild weather and so shows low average impacts. Kern and Greater Fresno, on the other hand, have the two highest average impacts at least partially due to warmer weather. Notably, Kern participants also have much higher baseline usage than those enrolled in the other LCAs, which increases impacts.

Figure 4. Residential ex post average impacts by location (Local Capacity Area)



Circles are the average across all meters in the monthly event. Lines are the 90% confidence intervals.

Ex Ante Load Impacts

The ex post analyses results, shown above, quantify what happened in 2019. This section shows results for the ex ante forecasts and quantifies future DR resources from Leap.

Ex ante load impact forecasts predict the load impact that would occur in standard event times and conditions for a total of four hypothetical weather-years. For this report, predictions are made for the peak day of each month for two standard weather years – corresponding to conditions that are expected to lead to peak electric load in either one out of every two years or one out of every ten years, on average. There is a slight difference in the conditions associated with system-wide peak load for the California Independent System Operator (CAISO) and the peak loads of individual Investor-Owned Utilities (IOUs). Combining the two peaking conditions with the two types of weather year leads to four sets of ex ante weather data.

CDA predicted ex ante impacts through statistical models that predict the load impact per meter for different customers (i.e., different sectors and load types). The models were fit to data from calendar year 2019 and used to predict the load impact per meter for the standard weather conditions mentioned above, for the projected mix of meters by category that is forecast for future years. The predicted load impact per meter multiplied by the projected number of meters equals the projected aggregate load impact. Additionally, because Leap is an organization with plans for significant capacity growth during the forecast period, our evaluation also includes a low, medium, and high scenarios from Leap's participation forecast.

A major product of the ex ante analysis is a 'table generator' – a spreadsheet – that allows the user to select a forecast year and a set of weather conditions and see estimates of load impact by hour, in each LCA separately or the total of all LCAs.

Forecast Growth

Leap expects to grow quickly over the next few years. The projected growth in enrollment varies by SubLAP and is based upon both the data from Leap's enrollments to date, as well as Leap's forward-looking partner pipeline and planned recruitment efforts. The result is a projected rapid increase in aggregate load impact. (Appendix B describes Leap's forecast.)

CDA obtained Leap's forecast in the form of the aggregate load impact during the resource adequacy (RA) time period for a typical weather year and for a low, medium, and high impact forecast. Figure 5 shows Leap's forecasted aggregate load impacts (i.e., the qualifying capacity values) for 2020-2024 for the medium load forecast by load type. As can be seen below, Leap forecasts electric vehicles, residential AC and residential battery impacts to make up a large percent of their impacts each year, with larger absolute values over time as well.

Figure 5 Leap's forecast aggregate load impact, by year and load type (medium forecast)

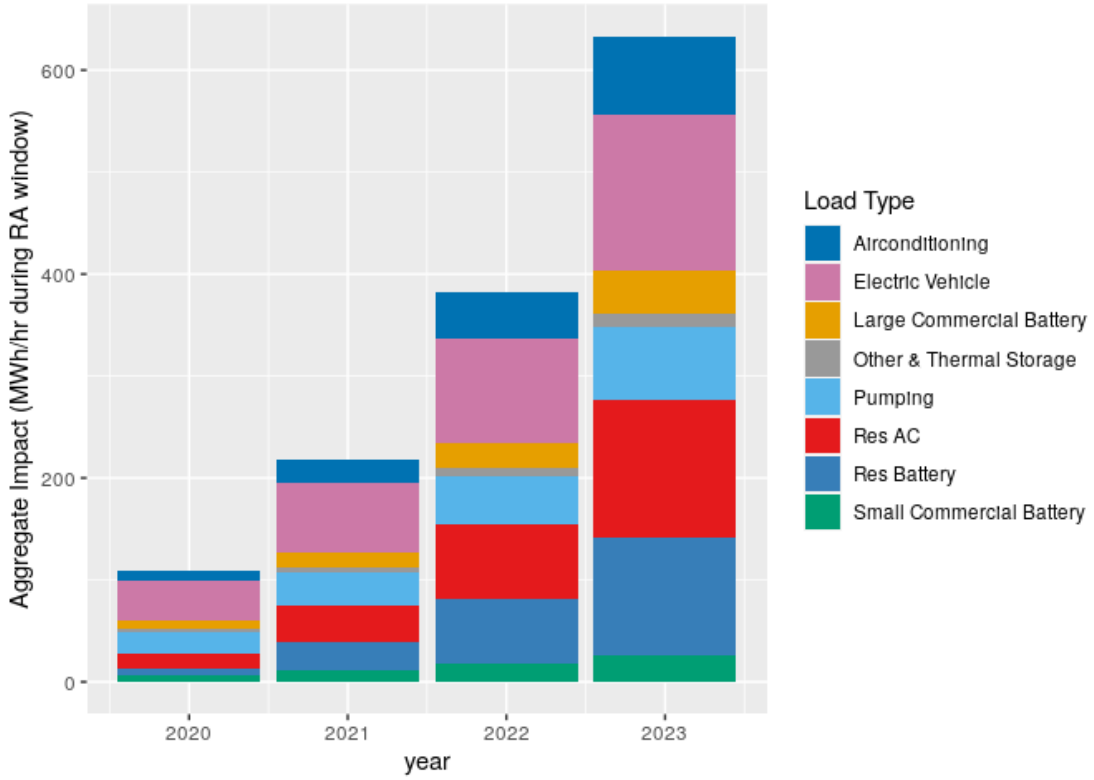
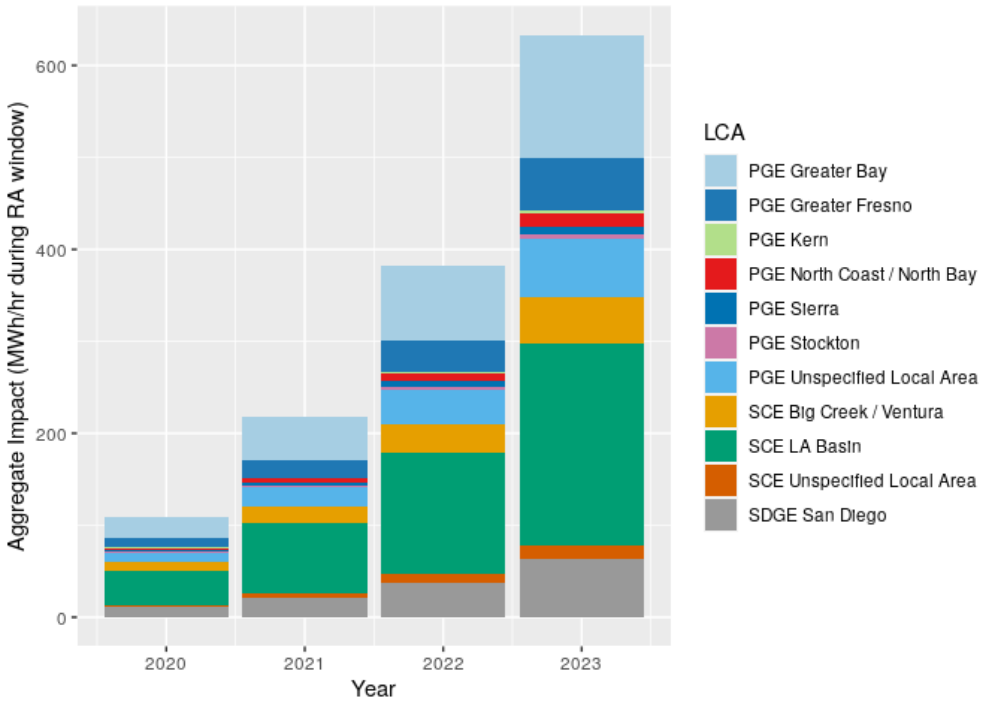


Figure 6 shows the same Leap forecast, but this time by LCA. Leap forecasts much of their impacts to occur in Greater Bay and LA Basin LCAs.

Figure 6 Leap's forecast aggregate load impact, by year and LCA (medium forecast)



Leap Combined Load Impacts

Leap provides a combined DR resource across both nonresidential and residential sectors. Figure 7 and Figure 8 show the combined nonresidential and residential predicted impacts for the medium enrollment forecast (i.e., the full Leap predicted impacts). Figure 7 reflects the growth in capacity that Leap forecasts between 2020 and 2023, with August 2020 obtaining slightly over 100 MW impact and August 2023 obtaining over 600 MW of impact. Figure 8 shows the increasing load impacts by LCA, with PG&E's Bay Area having the highest predicted impacts over time.

Figure 7. Combined sector predicted impacts by month and year (medium forecast)
(nonresidential and residential resources combined)

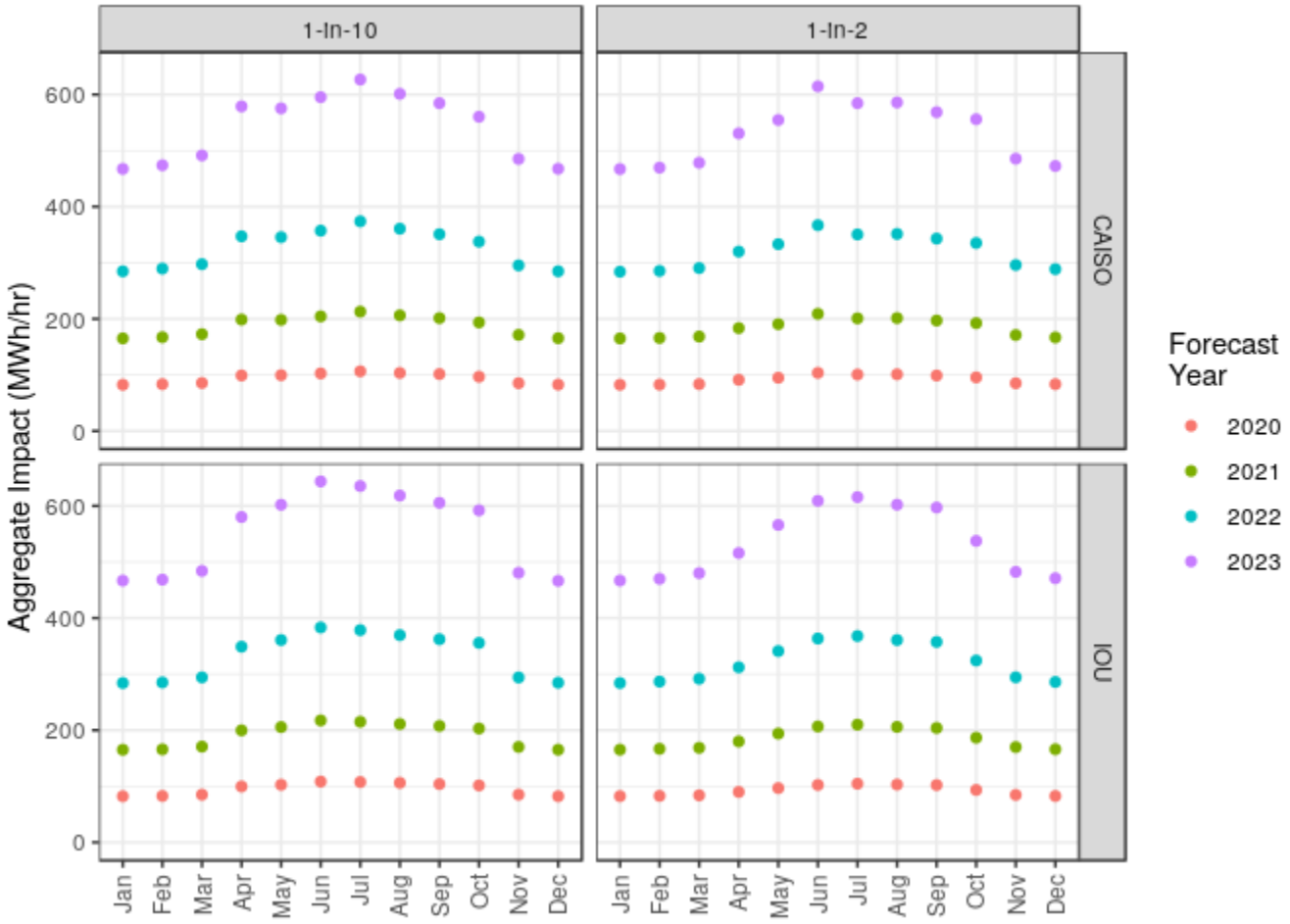
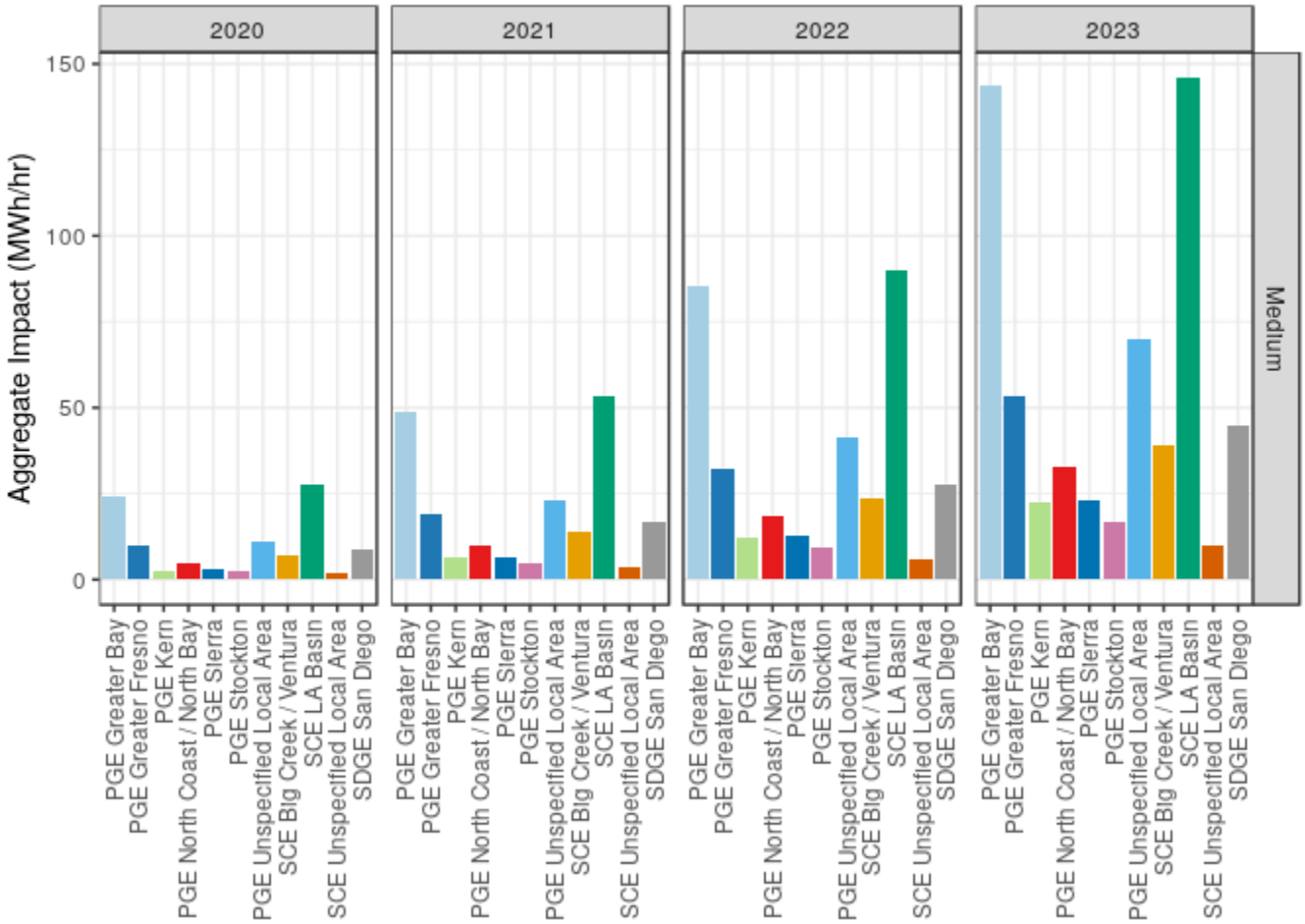


Figure 8. Combined sector predicted impacts by LCA and year (medium forecast)
(nonresidential and residential resources combined)



Recommendations

Based on our evaluation of the 2019 dispatch of Leap's DR resource, we provide the following recommendations to Leap:

- Call some longer-duration and full-resource events that can provide statistical support for full-resource and 4-hour+ RA window events that Qualifying Capacity numbers are based upon.
- Call events during more months of the year to gather information about seasonality and weather influences on event impacts.

Recommendations for future evaluators:

- Investigate baselining and comparison group methodologies for estimating event impacts that best characterize impacts for groups with few participants, varied events, and noisy baselines. This could include Leap dispatching at least a subset of their future events with true randomized controls.
- Study two load types where Leap forecast high future impacts.

    - Closely monitor the EV impacts during any future test events to ensure Leap is obtaining an event response that is consistent across all event hours.
    - Evaluate future test events on residential batteries to determine if the simulated event impacts within the ex post analysis are comparable to actual events. Compare not only the average impacts, but the impacts over the test period.
- Consider how best to apply LIPs so that they align with the needs of third party and emerging program evaluation. Most notably, the "Typical Event" requirement of Protocol 8 is not appropriate to characterize the full resource when all participants are not dispatched for all events. Also consider how best to characterize a resource that is growing and/or changing rapidly.
- Investigate ways to characterize and, where possible, measure sources of uncertainty, such as between customer variation, variation in event participation, and variation in load and customer types.

Introduction to the 2019 Leap Resource

Leap, founded in 2017, won capacity through the California Demand Response Auction Mechanism (DRAM) in 2018, and became an active Scheduling Coordinator and Demand Response Participant in the CAISO system in 2019. Leap delivered Resource Adequacy (RA) to PG&E and SCE in 2019 and has expanded to deliver RA to SDG&E in 2020.

Participants

Leap is continually recruiting new customers and refining its methods and strategies for dispatching events. (Leap recruited 3838 meters after the 2019 test events completed in August, representing 48% of meters). In 2019, Leap obtained load impacts for both the nonresidential and the residential sectors. Nonresidential load impacts were from pumping, air conditioning, electric vehicles, battery storage (large and small), thermal storage, and other (light manufacturing and cold storage facilities). Residential load impacts occur from air conditioning and battery storage (associated with a PV system).

Events

In 2019, Leap called test events only. CDA included 26 test events for nonresidential customers and 16 test events for residential customers in the ex post analyses. Table 2 shows that most events occurred between May and August. [8]

Table 2. Count of events and participants by sector and month

Sector	Month	Event Count	Average Participant Meter Count
Nonresidential	April	1	233
	May	3	123
	June	6	116
	August	15	58
	November	1	1
Residential	June	6	336
	August	10	874

All events lasted two hours and occurred between the hours of 4 to 6 PM, 5 to 7 PM, or 7 to 9 PM. Most events were called from 4 PM to 6 PM. (Table 3)

Table 3. Count of events by time of day and sector

Event Period	Count of Nonresidential Events	Count of Residential Events
4 PM to 6 PM	15	10
5 PM to 7 PM	5	4
7 PM to 9 PM	6	2

[8] Residential battery storage participants were not included in the test events because they were added after the test events. CDA simulated impacts for this category based on data from Leap's solar and storage partner (see section on "Residential Battery Storage" within the ex post impact analysis methods).

LCA

This report presents results in terms of LCAs. There are 10 LCAs, as well as a group of customers that do not fall into any LCAs. Table 4 shows the count of Leap customers and meters by LCA in 2019. Not all customers participated in test events due to being recruited after Leap had called the tests.

Table 4. Count of customers and meters active in 2019 by IOU and LCA

		Nonresidential			Residential	
IOU	**LCA**	Customers	Meters		Customers	Meters
PG&E	Greater Bay	**<16**	66		4,433	4,593
	Greater Fresno	**<15**	176		606	620
	Kern	0	0		201	207
	North Coast / North Bay	**<15**	15		306	318
	Sierra	<15	<15		469	473
	Stockton	**<15**	<15		165	168
	Unspecified Local Area	<15	55		600	617
San Diego	San Diego	0	0		0	0
SCE	Big Creek / Ventura	**<15**	122		**<100**	<100
	LA Basin	24	384		**<100**	<100
	Unspecified Local Area	<15	<15		0	0
Municipal Utility	Unspecified Local Area	<15	<15		<100	<100
All		*86*	*848*		*6,880*	*7,098*

Note: Customer and meter counts are the full resource counts as of the end of 2019.

Key concepts and decisions

This section marshals in one place evaluation-relevant details of Leap's resource and the circumstances of this evaluation that have required careful thought and/or contributed substantially to the outcomes or interpretations presented in this report. The information and caveats presented in this section are necessary to understand our methods and results.

Heterogeneous resource

Leap aggregates across customer types (residential, commercial, agricultural, and light industrial) in a manner that no IOU DR program does. Their resource, and therefore their evaluation is a bit like several more narrowly scoped DR resources rolled into one. For this evaluation, we've drawn upon experience and prior evaluations in residential, commercial, agricultural, and industrial categories. This heterogeneity has been factored into the methods we've employed to measure impacts. It is meaningless to compare across residential AC and agricultural pumping, for example, and they cannot be combined in models. Different categories of programs have widely differing sample sizes and precedent for measurement and reporting, where capacity bidding programs, for example, are often heavily redacted due to the very small number of large customers participating. This heterogeneity is notably on display in the ex ante forecast, which was produced in terms of capacity relative to 2019 rather than enrolled customer counts to provide greater comparability of resource capacity across load types.

Full resource called incrementally

Leap's test events were designed to call virtually all customers who had enrolled by August (52% of their customers participated in test events, because Leap enrolled many customers after August), but they did not attempt a "full resource" test where all customers were called at once. For this reason, the aggregate impacts of individual ex post events (or the average aggregate impact across events) are well below the full capacity of their resource, if it were to be called all at once. Combined with the written requirement in the LIPs that "Typical event" impacts and participation should be reported for ex post as the average across events, the result is a very low reported "Typical" event aggregate performance and participation. To capture something closer to the full resource potential, we define "Full resource event" impacts and participation to be the average across all test event participants multiplied by the number of meters enrolled at the end of 2019. This metric is not corrected for weather, which is done in ex ante, but we note that Leap did not call events on systematically hot days and their average event temperature is slightly below the CAISO 1 in 2 temperature for ex ante.

Day matching baselines vs. control groups

The results in this evaluation are based on the use of an X of 10 baseline. Our methods do not include a matched non-participant comparison group since, as a non-utility, Leap does not have access to utility account data for non-customers.

For residential air conditioning ex post impact estimates, where customers are relatively similar, we adjust the X of 10 baseline using customers who are enrolled but were not dispatched for a given event as a comparison group since those customers are quite similar to those who were dispatched. For nonresidential ex post impacts, using non-participating customers is not helpful because there is so much variation in both electricity consumption and event response between customers. For all residential and nonresidential test day participants, we assess both random error and bias by applying models to non-event days and seeing how well the baseline matches actual electricity consumption (See Ex Post Model sections for graphics that show random error and bias for all tested models).

We are recommending that Leap look into dispatching at least a subset of their future events with true randomized controls. This would help to solve the problem of how to model impacts because the methodology for estimating impacts from randomized control trials is well established statistical practice, and the impact estimates should be unbiased on average.

4-hour and multi-day ex-ante events

To qualify for RA, the Commission requires a DR resource to be able to operate for a minimum of four hours per day during the 5-hour RA window for three consecutive days.[9] In this evaluation, we have mainly 2-hour events to work with. There is no empirical basis to estimate the "decay" of resource performance over events of longer duration. The Leap resource is a combination of load shifts and device dispatch, where the devices include thermostats, batteries, and other controllable loads. Some strategies, like deferring loads, can easily shift loads fully out of the 5-hour RA window, while others, like thermostat setbacks, most likely cannot defer all cooling for a 4-hour event. Because Leap has a highly heterogeneous resource, there is no hard and fast rule or precedent for how their performance by load type is expected to degrade over long events.

[9] See Decision 11-06-022 p53 http://docs.cpuc.ca.gov/PublishedDocs/WORD_PDF/FINAL_DECISION/138375.PDF The RA window is five hours long (i.e., from hour starting 4 PM to hour ending 9 PM)

We have determined that the best course of action for air conditioning loads is to use "hour of event" degradation factors derived from the few DR program evaluations that have documented it. Southern California Edison (SCE) determined the relationship between load impact of a one-hour event and the impact in subsequent hours of a multi-hour event for the Summer Discount Plan and Smart Energy Programs, in both May and August of PY18. For residential air conditioning, we take the average of the SCE May and August results for each event hour and applied this decay curve to the predictions from the ex ante model. By hour, the factors are 1, 0.96, 0.70, 0.63, 0.61. For example, the predicted load impact for the fourth hour of an event is 63% of the prediction for a one-hour event at the same time of day and the same outdoor temperature. Since 2019 events were all two hours long, providing empirical estimates for the first two hours, we shifted the factors to the next hour, scaling a four-hour event using the factors: 1, 1, 0.96, 0.70.

For residential batteries, CDA modeled the full five hours of the RA window to directly estimate impact for each hour. For residential AC, CDA modeled a four-hour event during the first four hours of the five-hour RA window.

In the case of the nonresidential loads other than-air conditioning, most of these are either shifting load or large enough systems that they can support 4- or 5-hour events without degradation. For nonresidential air conditioning, we took results from the 2016 Capacity Bidding Program evaluation which included multiple 4-hour events. We calculated the overall mean empirical degradation across those events to estimate factors for a four-hour event of 1, 0.86, 0.78, 0.74. Since 2019 events were all two hours long, providing empirical estimates for the first two hours, we shifted the factors to the next hour, scaling a four-hour event using the factors: 1, 1, 0.86, 0.78.

For all nonresidential load types, we estimated impacts for a four-hour event in the first four out of the five hours of the RA window.

Key Research Questions and Study Methods

Key Research Questions

The research:

1. Estimates the ex post load impacts for the Leap demand response resources for PY2019
2. Estimates the ex ante predicted load impacts for the Leap demand response resources for years 2020-2030
3. Looks at LCA effects
4. Looks at the impacts of weather and time of year

Limits to our analysis included:

- **Lack of nonparticipant data** – Some LIP calculations that are often part of a utility DR evaluation (such as obtaining a counterfactual estimate using matched nonparticipant data) were not possible because Leap, as a third-party vendor, does not have access to utility nonparticipant data.
- **Lack of past participation data** – The LIPs require comparison to previous years' results, but this is the first year for an evaluation of Leap's resource.

All analysis in this report was performed using R software[10] with data.table[11] and tidyverse[12] packages.

Ex Post Impact Analysis Methods

For the **ex post analysis**, we estimated load impacts and baseline loads (also known as 'reference loads') for participants on event days using 10-of-10 day-matching models for all meters that participated in events during 2019.[13] We simulated impacts from residential battery systems using 15 minute data and information from Leap describing how they operate this load during non-event and event days. Residential and nonresidential meters use electricity and perform quite differently in response to DR events, so we estimate ex post impacts separately and used different models for these groups. Table 5 shows the models that we used for the ex post analysis.

Table 5. Final ex post models

Customer Type	Final Model
Nonresidential	10-of-10 day matching with no adjustment
Residential AC	10-of-10 day matching with Difference-in-Difference adjustment
Residential Battery	Simulation

We tested a variety of models before selecting the final models for this ex post analysis. We wanted to select models that have minimal bias and low variance determined by comparing a variety of day matching and regression models' performance on non-event days. The models we tested are listed in Table 6, described in detail after Table 6, and summarized in Table 8.

Table 6. Ex post models tested

Customer Type	Models Tested
Nonresidential	5-of-10 centered
	10-of-10 with no adjustment
	CAISO 5-of-10 with same day adjustment
	CAISO 10-of-10 with same day adjustment
Residential	5-of-10 centered
	10-of-10 with no adjustment
	10-of-10 with Difference-in-Difference adjustment
	5-of 10 centered with Difference-in-Difference adjustment
	Panel Regression

[10] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[11] Matt Dowle and Arun Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.8. https://CRAN.R-project.org/package=data.table

[12] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

[13] All analyses were performed at the meter level, not customer level, as Leap can control dispatch at the meter level.

Model Descriptions

Centered 5-of-10 and Unadjusted 10-of-10

A standard industry approach for estimating baselines is the 'X-of-10' method, often '10-in-10' or '5-in-10'. A standard X-of-10 model assumes the baseline load for a customer is the mean (average) load of the X highest-load days out of the 10 non-holiday weekdays prior to the event day (X being equal to 5 or 10). More generally, this approach can sample from any set of days (before or after the event) understood to potentially share similar conditions with event days. Centered X-of-10 uses the X days where load is closest to the event day during a set of non-event hours.

To calculate the 5-of-10 and 10-of-10 day-matching models, we first removed all weekends, NERC holidays and Public Safety Power Shutoff (PSPS) days from the data. Then, separately for each meter on each day, we selected the 10 days prior to the event day. For 10-of-10, the mean load for each hour on the selected non-event days is the unadjusted 10-of-10 baseline. For 5-of-10, we selected 5 of the 10 days where the average load from 11 am to 3 pm is closest to the event day load in the same time period. The average hourly load across the selected 5 days is the centered 5-of-10 baseline.

CAISO 5-of-10 and 10-of-10 with same day adjustments

For the CAISO 5-of-10 and 10-of-10 day-matching models, we followed the CAISO day matching examples from CAISO's website.[14] These models first remove all weekends, NERC holidays and PSPS days, then select the 10 days prior to the event day. For 10-of-10, the mean load for each hour on the selected non-event days is the unadjusted 10-of-10 baseline. For 5-of-10, we selected 5 of the 10 days where the average load from 11 am to 3 pm is closest to the event day load in the same time period. The average hourly load across the selected 5 days is the unadjusted 5-of-10 baseline. Next, we adjust each of the baselines using average load on selected non-event hours. For the CAISO 10-of-10, these are the three hours starting four hours before the event, and for CAISO 5-of-10, these are the two hours starting four hours before the event and the two hours starting two hours after the event. Finally, we multiply the unadjusted X-of-10 by the ratio of the mean load during the adjustment hours on the non-event days, and the mean load during the adjustment hours on the event day. If the ratio is outside of the interval [0.8, 1.2] for 10-of-10 or [0.71, 1.4] for the 5-of-10, we use the limit, rather than the actual ratio in the adjustment.

Centered 5-of-10 and 10-of-10 with Difference-in-Difference Adjustment

We use Difference-in-Difference (DID) models for residential AC customers because they draw on meter data from nonparticipating in addition to participating meter data. These models take advantage of the fact that only a subset of meters were dispatched in each event, so we are able to use information from event and non-event day electricity consumption from all customers, not just those who were dispatched for a given event. In this way, they use substantially more available information than most same-day adjustment models that only use data from the customer being measured.

The DID models work on aggregated data. Instead of a same day adjustment, we use a DID calculation that calculates event day impacts using both nonparticipating meter loads and the selected 5 or 10 days. To calculate the centered 5-of-10 and 10-of-10 DID models, we use nonparticipating load data during event days to separately adjust each hour. For a given event in a given SubLAP, we averaged the load for each hour for participants on the baseline days and on the event day, and for non-participants on the baseline days and on the event day, yielding 4 averages for each hour. For each hour of each event day,

[14] CAISO day-matching example workbook: http://www.caiso.com/Documents/Example-DayMatchWorkbook.xlsx

we calculate the impact as the difference of participating meter event day average and the participating meter non-event day average and then subtract the difference of nonparticipating meter event day average and nonparticipating meter non-event day average.

Panel Regression

Panel regression uses meter data for all enrolled meters on event days and comparison days to estimate impacts for residential sites. We do not use panel regression models for nonresidential sites because electricity consumption across nonresidential meters is much more variable than for residential meters.

The CDA team identified a set of 'comparison days' for each event with weather conditions comparable to each event day by matching hourly temperatures from non-event to event days using Euclidean distance matching. These days are distinct from the X-of-10 days selected by the day matching models. We use some of the comparison days in the regression models and others for calculating errors, as described in the non-event day model section.

Regression-based impact estimates and reference loads use three different models, one for each of the event start times (16:00, 17:00, and 19:00). The input data (the same for each of the three models) was designed to be run as a panel regression with data for the sub-group of customers whose load impact is to be estimated. These variables inform the regression model:

Table 7. Data columns involved in panel estimation of hourly load impacts

Data column	Explanation
kWh	The total kWh consumption for the hour. This is the variable on the left-hand side of the regression equation to be explained by all the other factors (i.e., it is the dependent variable while all other data shown below are the independent variables).
date	Date of the electricity consumption. The dates in one panel will include all event days with events starting the same hour (16:00, 17:00, and 19:00) and all comparison days.
meter_id	The anonymized unique identifier of the meter each reading belongs to. The meter_ids in a panel will include all the event participants (the cases) and all of their controls.
hour	Hour of day of the electricity consumption indexed to 1 through 24, with 1 spanning midnight to 1am and so on. The panel will have been filtered to a single hour of day prior to the estimation of the load impacts for that hour.
event	Indicator that is 1 if the reading is from an event day or 0 if the date is a comparison day.
cdh70	"Cooling Degree-hours". The total number of degrees by which the average temperature for the date and hour exceeds 70 F at the nearest weather station to each customer. This is used to quantify the air conditioning (and other temperature sensitive load) contribution to the load data. The common choice is 65 F, but we tested 65, 70 and 75, and found that 70 leads to substantially improved model fit.
late_eve_load	Average electricity consumption from 10pm to midnight, added to the model to allow it to perform a same day correction that recalibrates impacts to near zero late in the evening.
morning_load	Average electricity consumption from 7am to noon. We include this in the model to help adjust for day to day differences in energy consumption that can occur due to several hot days in a row.
night_temp	Early morning average outdoor temperature from midnight to 6am, added to the model to help adjust for differences in overnight temperatures.

We also tested several other variables, including maximum temperature, day of week, and early afternoon load, but found that the addition of these variables did not improve model fit, so they are not included in the regression model.

In R's formula notation, all three panel regression models all have the same formula:

```
Equation 1: kWh ~ event + hour + cdh70 + late_eve_load + morning_load
+ night_temp + event:hour + event:cdh70 + hour:cdh70 +
event:hour:cdh70 + late_eve_load:cdh70 + morning_load:cdh70 | meter_id
```

This means that the energy consumed for a given hour (kWh) is explained by a regression model that first converts each meter id to variation around its mean electricity consumption, then estimates coefficients for each of the listed variables and interactions (as well as an intercept term, which is implicit in R's notation). Because our panel data set spans readings from many customers days, the fits apply to the average outcome across all modeled customers.

With this regression model, we are effectively implementing a baseline calculation that can adjust for outside temperature and make corrections to ensure that event impacts start near zero prior to each event. The load impact term is the event coefficient and interaction that include the event variable, corresponding to the difference in consumption seen during event days, compared to what would have happened if there were no event, after adjusting for weather and overnight temperature, morning load, and late evening load.

From this model, the baseline load on an event day is determined by using the model to make a prediction using the event-day values of all variables but setting event = 0.

The table on the next page summarizes each model based on the number of days in the baseline, choice of baseline days, hours used to choose baseline and any adjustments.

Model Summary

Table 8. Summary of models tested in the Leap ex post analysis

Model	Days in baseline load	Choice of baseline days	Hours used to choose baseline days	Adjustment
5-of-10 centered	Average load of the 5 closest load days out of 10 non-holiday or PSPS weekdays prior to the event	Has similar conditions with event days and uses the closest 5 days of the 10 before the event day	11 AM to 3 PM	None
5-of-10 centered with DID adjustment				Adjust using DID on selected non-event days and for non-participants
10-of-10 with no adjustment	Average load of the 10 non-holiday or PSPS weekdays prior to the event	Has similar conditions with event days	NA	None
10-of-10 with DID adjustment				Adjust using DID on selected non-event days and for non-participants
CAISO 5-of-10 with same day adjustment	Average load for 5 highest of the 10 non-holiday or PSPS weekdays prior to the event	The 5 days out of the 10 day before the event with similar conditions to event days	11 AM to 3 PM	Adjust using the average load on non-event hours using average estimated from 4 hours (2 hours before and after the event) with a 2-hour gap around the event, limited to 0.71 or 1.4 (See Figure 9)
CAISO 10-of-10 with same day adjustment	Average load for the 10 non-holiday or PSPS weekdays prior to the event	The 10 days before the event	NA	Adjust using the average load on non-event hours. Average estimated from 3 hours before the event with a 1-hour gap before the event, limited to 0.8 or 1.2 (See Figure 10)
Panel Regression	Same days as event, uses weather matched comparison days	Days with matching temperature profiles	Comparison days are selected based on temperature profile	No adjustment needed. Impacts based on difference in meters that were participating in an event and those that were not

Figure 9. Hours used to adjust baseline for CAISO 5-of-10 Model shown in green

Figure 10. Hours used to adjust baseline for CAISO 10-of-10 Model shown in green

Errors and Validity

X-of-10 baselines do not automatically generate uncertainty (aka error) estimates for their values. Regression models applied to time series data like that used in this analysis do report error metrics, but these are often under-estimates since regression models use only the data provided to estimate both regression parameters and errors. This is an important limitation, so we estimated errors for all approaches using out-of-sample non-event day models.

There are three important sources of errors between a model and the phenomena it is trying to capture: intrinsic variation, model error, and prediction error.

Intrinsic variation comes from the fact that data, especially whole building electricity data in this work, can vary for reasons unrelated to the driving forces that determine event savings. This additional variability is called 'statistical noise', and it can interfere with the ability to quantify the effect of a demand response event. We address this error in our models by including non-event day information in the models and adjusting for weather through comparison day selection and regression models. The wide range of load reduction capacities among Leap's nonresidential customers is one major source of intrinsic variation that leads to wider confidence intervals than for their more homogenous group of residential customers.

Model errors have to do with the structure of the model and the choices and assumptions of the modeler. For instance, the regression equation discussed above linearly adjusts mean electric load at a given time of day for outdoor temperature (above a reference temperature such as 65 F) and other factors, but in fact perfect linearity is very unlikely.

Prediction errors come from the fact that it is impossible to perfectly predict the future or a counterfactual (something that did not actually happen but that we use to measure impact). Baselines are a type of counterfactual, and it is only possible to compare the baselines generated by a model against meters that were not dispatched or against the dispatched meters' non-event day usage. On average, if the non-dispatched meters use energy similarly to dispatched meters, checking against non-dispatched meters' energy usage and dispatched meters' usage on non-event days should reveal the random error and bias in the baseline prediction. This is the reason we use difference-in-differences to help reduce, but not remove, bias and random error.

Validity is generally broken into discussions of internal and external validity.

- An analysis is internally valid if the estimates are representative for the specific group being studied, we evaluate this by assessing bias. There are discussions of bias throughout this report, and we have endeavored to keep bias to a minimum. (See Figure 11 and Figure 12 for graphics of the bias we considered across multiple ex post models.)
- An analysis is externally valid if the estimates are representative of a larger outside population. In this case, external validity only applies to the ex ante analysis. The current participants are not a random sample of future participants, so it is not possible to be sure that the results are perfectly representative. We have used best evaluation practice, Bayesian models and careful assessments of error in the ex ante analysis to ameliorate this issue as best as possible.

Non-event-day models

To determine the statistical distribution of errors – also known as bias and variance – we take advantage of the fact that the load impact is zero (by definition) on days when there is no event. There can be no

load impact if there is no event.[15] Our approach to error estimates is to run the event models on non-event (comparison days) days from the same participants used to evaluate event days. By definition, all deviations from zero impact on comparison days are errors so this gives us the errors for every hour of the day, for each comparison day. We thus obtain a statistical distribution of errors for each hour of the day and calculate the bias (as deviation from 0) and variation (as RMSE) and apply these to the corresponding event day.

Nonresidential Ex Post Impact Analysis

For the nonresidential **ex post analysis**, we estimated load impacts and reference loads for participants on event days compared to similar non-event days (comparison days) using centered 5-of-10, unadjusted 10-of-10, CAISO 5-of-10 and CAISO 10-of-10. For each of these models, we estimated load impacts for each event, and for sub-groups, including IOU and LCA.[16]

We estimated average reference loads and load impacts (both with uncertainties) and tabulated meter count weighted temperatures for each hour of each event day for every sub-group modeled to report in quantitative deliverables.

These are the steps we took to arrive at our ex post estimates for nonresidential participants:

(1) **Identify impact model input data**: Assemble participant meter data associated with the sub-group(s) of customers whose impacts are to be modeled for event days and comparison days.
(2) **Run event models**: Run the X-of-10 calculations to estimate the load impact on the event days. Also estimate the load impact on 15 comparison days prior to the event.
(3) **Estimate errors**: Estimate model errors by running our event models on all comparison days (days without events). The correct answer for these non-events is zero impact, so any deviations from zero are taken as empirical model errors.
(4) **Run and store estimates for every customer sub-group**: Repeat the basic prescription of steps 1-3 over and over for every combination of customer attributes defining each sub-group, and for event days and comparison days (i.e. to compute the errors), with 24 hourly estimates of reference loads and load impacts returned with empirical errors alongside of participant counts and population weighted hourly average temperatures.
(5) **Use the best-performing model:** The load impact on comparison days is zero by definition, so the model that predicts the comparison-day load impacts to be closest to zero (on average) is the one that predicts baselines with the least bias. We assessed model variance (random error) using the same comparison-day impacts and calculated the overall variability in the deviation from zero.

[15] We are aware that DR can and very likely does have spillover effects on non-event days, but our job as evaluators of ex post impacts is to assess the impacts of calling an event vs. not calling an event because that is the dispatchable resource.

[16] We did not perform regression modeling for non-residential customers because of very high variability in electricity consumption and controlled loads across non-residential sites, and for some sites, such as those controlling EV charging, very high internal variability. The timing of loads such as electric vehicle charging can't be effectively predicted based on explanatory variables available to us (such as temperature, hour of day, and day of the week).

Ex Post Model Selection

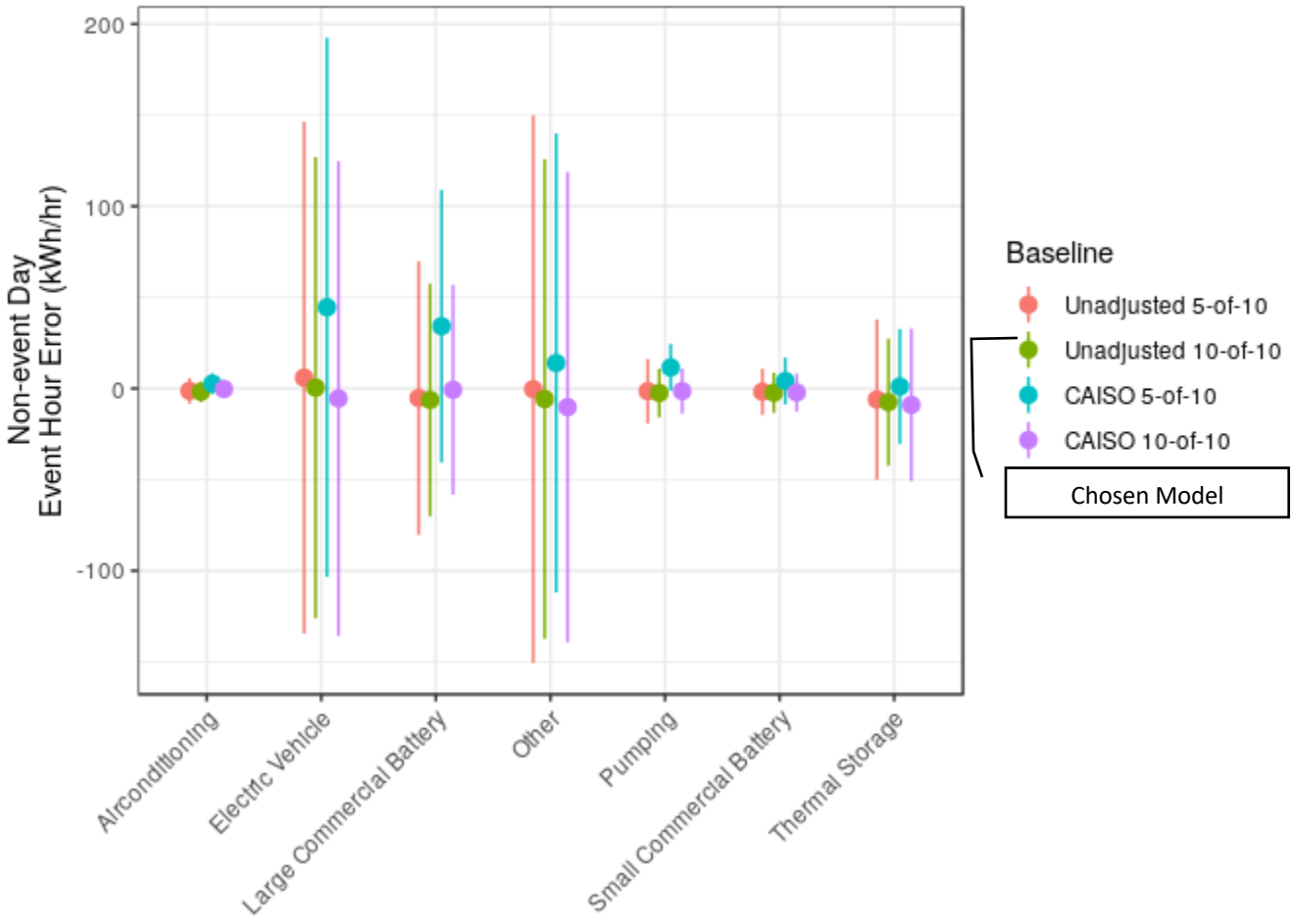
We compared the four model's performance against each other, looking for bias and random errors. Since these are non-event days, the mean impact should be zero.

- If the mean is not at zero, this is evidence of bias.
- Random error is shown by the length of the vertical lines in Figure 11, so models with less random error have shorter lines.

Figure 11 shows non-event day mean (colored circles) impact and random error (vertical lines) for each baseline model for non-residential meters controlling different loads. This plot shows that we often see substantial bias (visible here as deviation from zero impact) in the CAISO 5-of-10 model, and very little bias in the centered 5-of-10 and 10-of-10 models.

We selected the unadjusted 10-of-10 model because it has somewhat less random error than the unadjusted 5-of-10 model for most of the load types, while still having little to no bias.

Figure 11. Nonresidential plot of bias and random error by load controlled for four baseline models



Residential Ex Post Analysis

For the residential **ex post analysis**, we estimated load impacts and reference loads for participants on event days compared to similar non-event days (comparison days) using panel regression, centered 5-of-

10, unadjusted 10-of-10, centered 5-of-10 with DID, and 10-of-10 with DID. For each of these models, we estimated load impacts for each event, and for sub-groups, including IOU and LCA.

We estimated average reference loads and load impacts (both with uncertainties) and tabulated meter count weighted temperatures for each hour of each event day for every sub-group modeled to report in quantitative deliverables.

We used six steps to arrive at our ex post estimates for residential participants:

(1) **Identify comparison days for regression models and error estimates**: Match temperature shapes for event days to similar non-event days using weather data.
(2) **Identify impact model input data**: Assemble participant meter data and the local weather data associated with the sub-group(s) of customers whose impacts are to be modeled for event days and comparison days. The weather data is from NOAA's Integrated Surface Database, with each meter matched to the nearest reliable weather station.
(3) **Run event models**: Run the panel regression and the X-of-10 calculations to estimate the load impact on the event days. Also estimate the load impact on the comparison days; in the case of the panel regression this is done by leaving out one comparison day at a time and fitting the model using the other comparison days.
(4) **Estimate errors**: Estimate model errors by running our event models on comparison days (days without events). The correct answer for these non-events is zero impact, so deviations from zero are taken as empirical model errors.[17]
(5) **Run and store estimates for every customer sub-group**: Repeat the basic prescription of steps 1-4 over and over for every combination of utility, LCA, and SubLAP, and for event days and comparison days (to compute the errors), with 24 hourly estimates of reference loads and load impacts returned with empirical errors alongside of participant counts and population weighted hourly average temperatures.
(6) **Use the best-performing model:** The load impact on comparison days is zero by definition, so the model that predicts the comparison-day load impacts to be closest to zero (on average) is the one that predicts baselines with the least bias. We assessed model variance using the same comparison-day impacts and calculated the overall variability in the deviation from zero.

Ex Post Model Selection

We tested five modeling approaches for residential customers and compared the bias and random errors found during non-event days. (Non-event day impacts should be zero.)

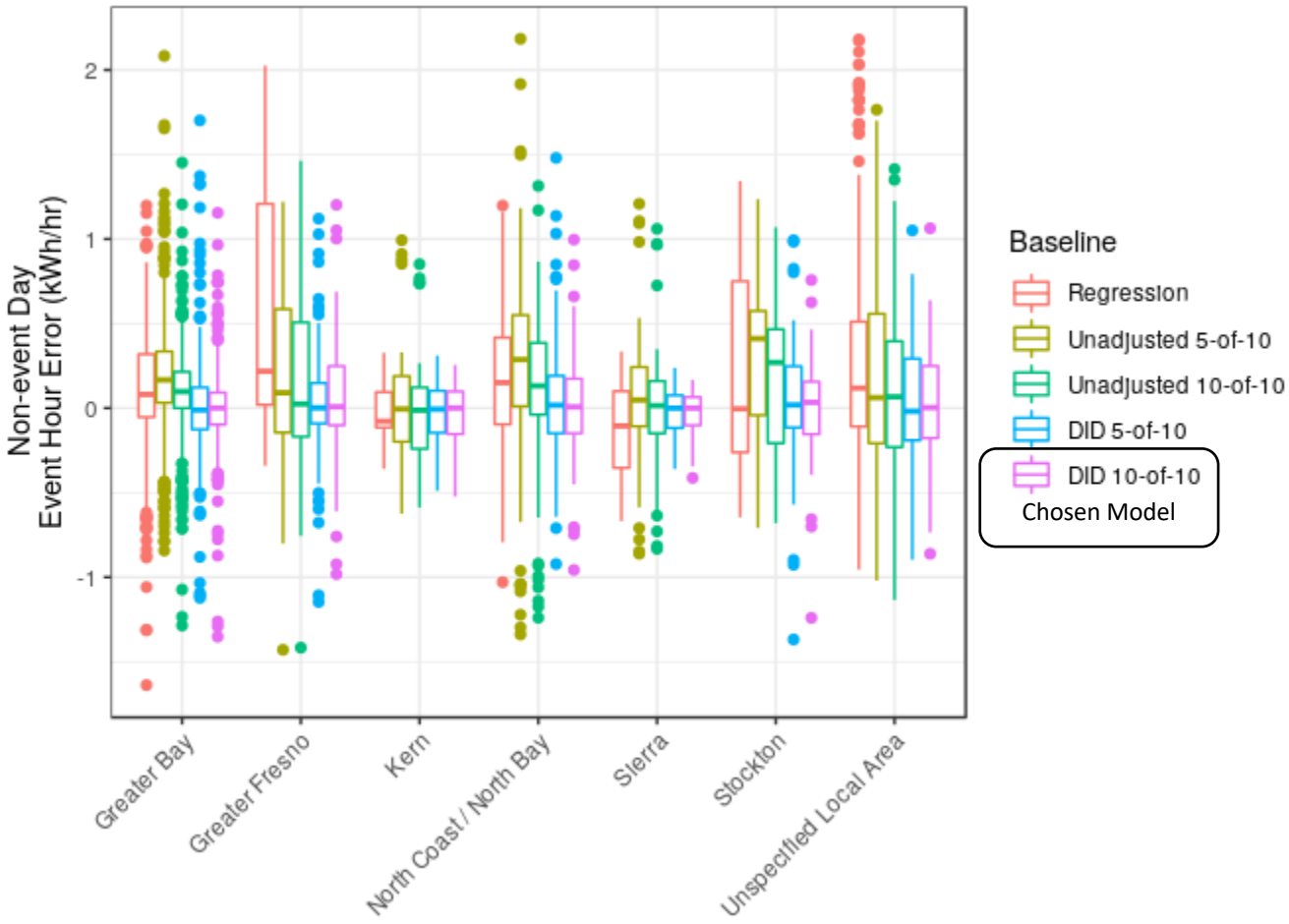
- If the average is not at zero, this is evidence of bias.
- Random error is shown by the length of the vertical lines in Figure 12. Models with less random error have shorter lines.

Figure 12 shows boxplots for each baseline model for residential meters in different LCAs. This plot shows that we often see substantial bias in the regression model, and very little bias in the DID 5-of-10

[17] The error metrics reported by the regression models are often under-estimates since regression models use only the data provided to estimate both repression parameters and errors. In order to estimate out-of-sample error (counterfactual baselines are out-of-sample because they are unmeasurable) we use out-of-sample comparison days to provide error estimates.

and DID 10-of-10 models. We selected the DID 10-of-10 model because it has somewhat less random error than the DID 5-of-10 model for most of the LCAs, while still having very little bias.

Figure 12. Residential boxplots of bias and random error by LCA for five baseline models



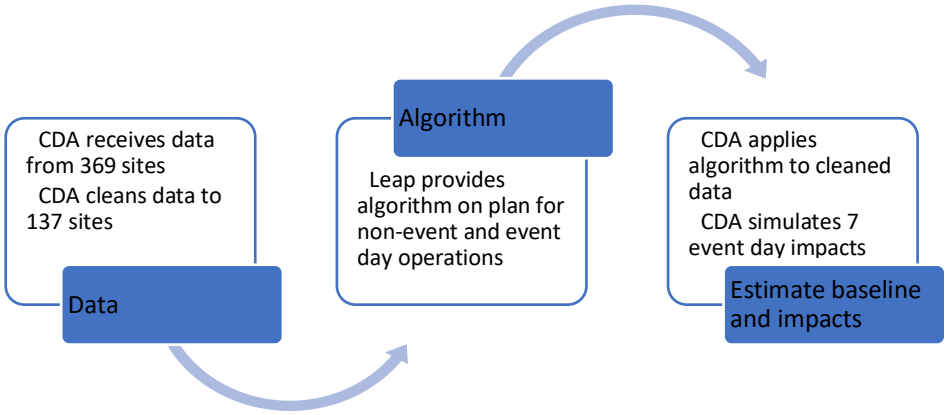
Note: Box and whiskers summarize non-event day impacts by Leap (where the expected result for the median would be zero). Boxes extend from the 25th to the 75th percentile, with the median, or 50th percentile marked with a horizontal line. Outliers are shown as unmarked points.

Residential Battery Storage

CDA simulated residential battery storage event impacts from data provided by Leap.[18] Leap provided inverter and battery data for 369 residential sites across California that have solar PV and battery systems. The data includes 15-minute interval measurements for a large portion of 2019: energy imported and exported from the utility, PV generation, battery charge level, battery charge energy and battery discharge energy. This provides a full picture of how the household is consuming energy and how the battery + PV system is working. Appendix H: Data Cleaning and Analysis for Residential Batteries describes the data cleaning and modeling steps in detail.

[18] Leap started to enroll residential battery customers late in 2019, after the summer event season, so these customers were not included in any test events.

Figure 13. Residential Battery Storage Inputs for Impacts



Leap also provided an in-depth description of the algorithm they plan to use for non-event and event day operation of enrolled residential battery systems. CDA applied this algorithm to the data to calculate baselines and impacts that simulate events using actual PV and battery performance, and actual site electricity consumption. We performed these calculations for seven (7) representative event days, one for each month where there was sufficient data (May, June, July, August, October, November, and December 2019).

Leap's algorithm clearly defines non-event and event day operation of the battery systems. Effectively, on non-event days the batteries charge with energy from the PV and then export at full capacity starting at 4pm until the battery is discharged to its reserve level (generally between 10 and 20% of charge). On event days, the batteries charge from the PV until the event start time, and then the batteries fulfill the household electricity demand (zeroing out imports, up to the maximum discharge rate of the battery system) until discharged to the reserve level.

We calculated the site electricity consumption for each 15-minute period using the meter data:

$$site\ demand = utility\ import + PV\ generation + battery\ discharge - utility\ export - battery\ charge$$

The **baseline** is the combination of site demand with the PV and battery operation calculated for non-event days as it would be visible at the utility meter, with the baseline set to zero during periods of net export.

- CDA calculated baseline on the selected event days by applying the measured PV generation to charging the battery until 4 pm, then assuming that the battery discharges at its maximum discharge rate from 4 pm until the battery discharges to its reserve level (usually around 5:30).
- After the battery is discharged, the PV generation is used to provide energy for site consumption or exported.

The **impact** is the baseline minus the event day consumption.

- CDA calculated electricity consumption on event days also using the Leap algorithm, which states that the systems will charge the batteries from PV generation until they are at full charge or the event starts.
- At the start of the event, the PV generation and battery discharge provide enough energy to zero out load visible at the utility meter up to the battery and/or inverter maximum load. This

discharge continues this until the end of the event period or when the battery is discharged to the reserve level.

- The event day consumption is set to zero during periods of net export.

CDA estimated error by running the same baseline and event-day calculations on a set of 10-20 days from each month. The random error is the standard error of average baseline and impacts for those non-event days. Since we simulated the non-event day baselines, it is not possible to estimate bias.

Ex Ante Load Impact Analysis Methods

As opposed to ex post analyses, which quantify what has happened in the past, ex ante predictions attempt to quantify the future. Ex ante load impact forecasts predict the load impact that would occur in standard event times and conditions for a total of four hypothetical weather-years.[19] For this report, predictions are made for the peak day of each month for two standard weather years – corresponding to conditions that are expected to lead to peak electric load in either one out of every two years, or one out of every ten years on average. There is a slight difference in the conditions that cause peak load for California Independent System Operator (CAISO) and for Investor-Owned Utilities (IOU). Combining the two peaking conditions with the two types of weather year leads to four sets of standard weather data.

The load impact capacity in future years is, of course, strongly dependent on the number, type, and scale of customers who are enrolled. Leap has provided a forecast of future enrollment, in the form of the future load that they anticipate being able to control in future years. They provided this forecast for each subLAP, for each load type, for each year. Taken at face value, these forecasts would themselves constitute the ex ante predictions that are needed, but such a prediction would be completely disconnected from the ex post results.

Instead, we treat the forecasts as a scaling factor: given the observed impact per meter in 2019, in each load type, how many more meters would need to be in the program in order to meet Leap's forecast of their load impact capacity? This converts Leap's impact capacity forecast to an enrollment forecast.

We then fit statistical models to the ex post data to quantify the performance per meter. The models serve three roles:

1. They quantify the extent to which the actual impact per meter is uncertain even in the ex post data, thus leading to uncertainty in future capacity even if the program grows as expected.
2. The models quantify the amount of variation in load shed per meter from LCA to LCA, thus allowing quantification of the uncertainty if the program expands into LCAs in which there are currently no participants, as they are forecast to do.
3. The models determine the temperature-dependence of the load impact for residential and commercial air conditioning and for residential batteries – the only three load types for which there is assumed to be temperature-dependence -- so that the load impact can be adjusted to the ex ante weather conditions.

The models were fit to data from calendar year 2019 and used to predict the load impact per meter for the standard weather conditions mentioned above, for the projected mix of customers by load type and LCA that is forecasted for future years. The predicted load impact per participant multiplied by the projected number of participants equals the projected aggregate load impact.

[19] For purposes of this report, we describe the impacts from the ex ante analysis as *predictions* and impacts received from Leap of their enrollments as *forecasts.*

By construction, the central estimate of the aggregate load shed in each category and each future year is fairly close to Leap's forecast, but even the central estimate is not identical to Leap's forecast for reasons described in items 1 and 2 above. Additionally, the model predictions include uncertainties, which are available in the table generator spreadsheet.

Finally, we applied 'fatigue' --- degradation of the results for long events – to the model predictions for both commercial and residential air conditioning. The air conditioning ex post events that were used to train the models were two hours long. The models predict the load shed per meter in a given LCA and in a given hour based only on which LCA it is, and on the outdoor air temperature in that hour, making no distinction between whether the hour is, for example, part of a two-hour event and a four-hour event. But customers who will tolerate higher air temperatures for an hour or two will not necessarily tolerate them for longer periods, so one might expect the amount of load impact to decay or degrade as the hours pass. This effect is included by decreasing the amount of load shed after the second hour, compared to what is predicted from the models. There is a section in this report that discusses where we obtained the degradation factors.

All predictions discussed in this report are made for four-hour events in the five-hour Resource Adequacy (RA) time window that runs from 4 p.m. through 9 p.m., unless specifically noted.

Ex Ante Methods

The CDA team estimated the ex ante load impacts for monthly typical and peak days for 2020-2030, separately for nonresidential and residential customers. Ex ante modeling uses ex post results and models the effect of various external determinants of performance, like location, load types, outside temperature, etc. on load impacts. We then apply forecasts for the external determinants of performance with the model to make predictions for future load impacts.

We fit the statistical models to the ex-post data, to develop predictions of load impact as a function of LCA and temperature for each load type. The modeling approach, called 'Bayesian Hierarchical Modeling', is described in Appendix E. The main reason for choosing this approach is that it provides reasonable estimates of uncertainty, especially for LCAs where Leap plans to expand and there is no test data to include in the ex post analysis. It would be wrong to assume that the unobserved LCA will provide exactly the same load impact per customer as is observed in the LCAs for which we do have data, but we also expect that the unobserved LCA is unlikely to have wildly different load impact per customer from the LCAs for which we do have data. The Bayesian analysis provides a statistical method that captures this expected behavior.

As part of the ex ante analysis, the CDA team estimated ex ante impacts for all areas covered by Leap customers in 2019, and additionally for SDG&E territory, where Leap is currently recruiting and registering customers.

Ex Ante Modeling

The ex ante load impact predictions require extrapolating the ex post impacts to future years:

1. For each hour of the day, each LCA, and separately for residential and nonresidential customers, find the ex post load impact for each event using the approach described above in the section on the 'Ex-Post Impact Analysis'. This yields one number per hour of the day, in each LCA, and load type for each event.
2. Fit a Bayesian hierarchical model – a type of regression model -- to predict ex post load impact from the outdoor air temperature at that hour, for that event. The model also generates uncertainties in the form of confidence intervals.

3. Use the model to make predictions for the required weather scenarios, including typical event days and monthly system peak days for the IOUs 1-in-2 (i.e., normal) and 1-in-10 (i.e., extreme) and CAISO 1-in-2 (i.e., normal) and 1-in-10 (i.e., extreme). These scenarios use the current resource adequacy 4-9 PM window, extending events to cover the entire window.
4. Baseline loads (reference loads) are needed in order to provide estimates of the relative (percentage) load shed due to Leap events for the weather scenarios. We fit models to predict hourly baseline load as a function of weather and season. The resulting models were used to predict the baseline loads for the weather scenarios and customer type.

Three of the ex ante models use outdoor air temperature as a predictive variable; these are the residential and commercial air conditioning models, and the residential battery model. Use of temperature as a predictive variable for the air conditioning model does not need explanation. We use temperature in the battery model because most of the residential batteries are charged from photovoltaic systems, and thus have a state of charge dependent on the amount of solar energy stored on the event day. Temperature serves as a proxy for both solar access – days are longer and the sun is higher in the sky in summer than in winter – and for the sunniness of the day, since sunny days tend to be warmer.

Ex Post Results

Our ex post results are by sector, with nonresidential sector first followed by the residential sector. Leap called a total of 26 two-hour test events, with each event calling only a portion of their available capacity. Throughout most of this section, the reported numbers are estimates of impacts for participating meters and are not Leap's full resource. We provide the ex post full resource estimate for Leap's 2019 resource at the end of the nonresidential and residential sections.

Throughout this report, averages represent the best estimate of measured impacts, and 90% confidence intervals provide a measure of uncertainty. There are always uncertainties in impacts. For Leap, ex post uncertainties are due to three sources:

1. Uncertainty in modeled baselines
2. High variability in impacts across participating meters due to large differences in baselines
3. Variable participation in events due to different customers being dispatched

These uncertainties show up as wider or smaller confidence intervals associated with the test events. Additionally, as results are "rolled up" across more events and more meters, these uncertainties shrink.

Nonresidential

Number of Events

For the nonresidential sector, CDA identified and estimated ex post impacts for 26 discrete Leap DR events in 2019. All events were test events. Table 9 below presents the nonresidential test event information. Most test events were called in August, but more participants were included in the months of April to June. Aggregate impacts ranged from 11% to 27% of the baseline values with the largest aggregate load shed (10.1 MW) occurring during the month of August, the month with the largest number of events.

Test events always included a subset of the total customer portfolio, so Leap's aggregate impacts across their entire portfolio in a given month and in 2019 exceeded any of the individual aggregate impacts noted in the table below.

Table 9. Nonresidential summary of 2019 events, temperature, baselines, and impacts

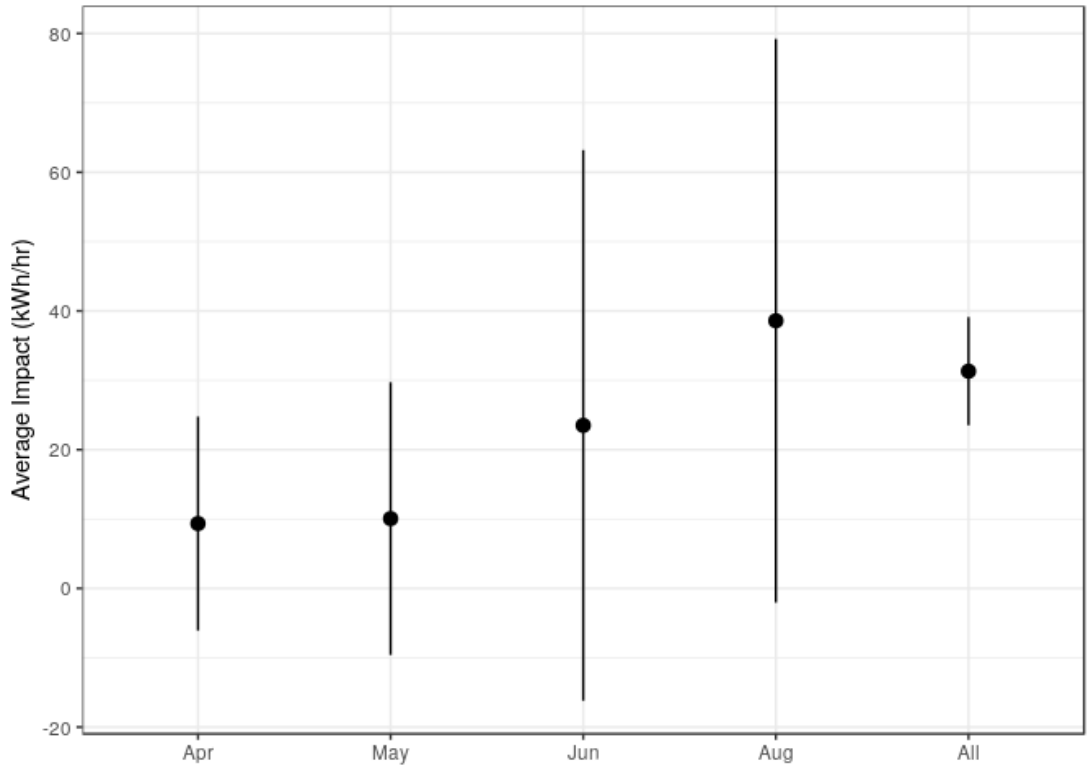
month	# of events	average # of participant meters	average temperature (F)	average baseline (kW)	average impact (kW)	aggregate baseline (MW)	aggregate impact (MW)	impact %
April	1	233	78	86	9	20.0	2.2	11%
May	3	123	82	67	10	14.6	1.2	15%
June	6	116	78	118	23	24.6	4.6	20%
August	15	58	87	142	39	32.1	10.1	27%

Average Ex Post Load Impacts by Test Event Months

Figure 14 depicts the range of per-meter impacts by month for events called in 2019. Within this figure, August has the highest average impact (39 kW) with the majority of the impact coming from air conditioning and pumping loads. April, with the lowest average impact (9 kW) also had the majority of impact from air conditioning and pumping loads, but with the impact coming from more meters and a slightly cooler average temperature.

The highest average per-meter impact is 162 kW during November, which is not included in the table above or figure below because it is for one event with one participating meter. August has the widest range of average impacts due to a smaller number of participating meters and greater proportion of EV charging in each event (EVs impacts were found to be highly variable, as discussed in the next section). April, with the highest number of meters per event and no EV charging, has the smallest range. The overall, full resource estimate at the far right shows an estimate of end-of-2019 enrolled capacity which has lower uncertainty due to combining data from all events and participating meters across months (including November).

Figure 14. Nonresidential ex post monthly average event impacts



Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

Average Load Impacts by Customer Type and LCA

This section presents average load impacts for customer type and LCA.

Customer Type Results

Among the multiple customers nonresidential load impacts (customer types), the "other" has the highest per-meter impact (100 kW).[20] Air conditioning, small commercial battery and thermal storage loads all brought in similar loads (between 4 and 14 kW) per meter.

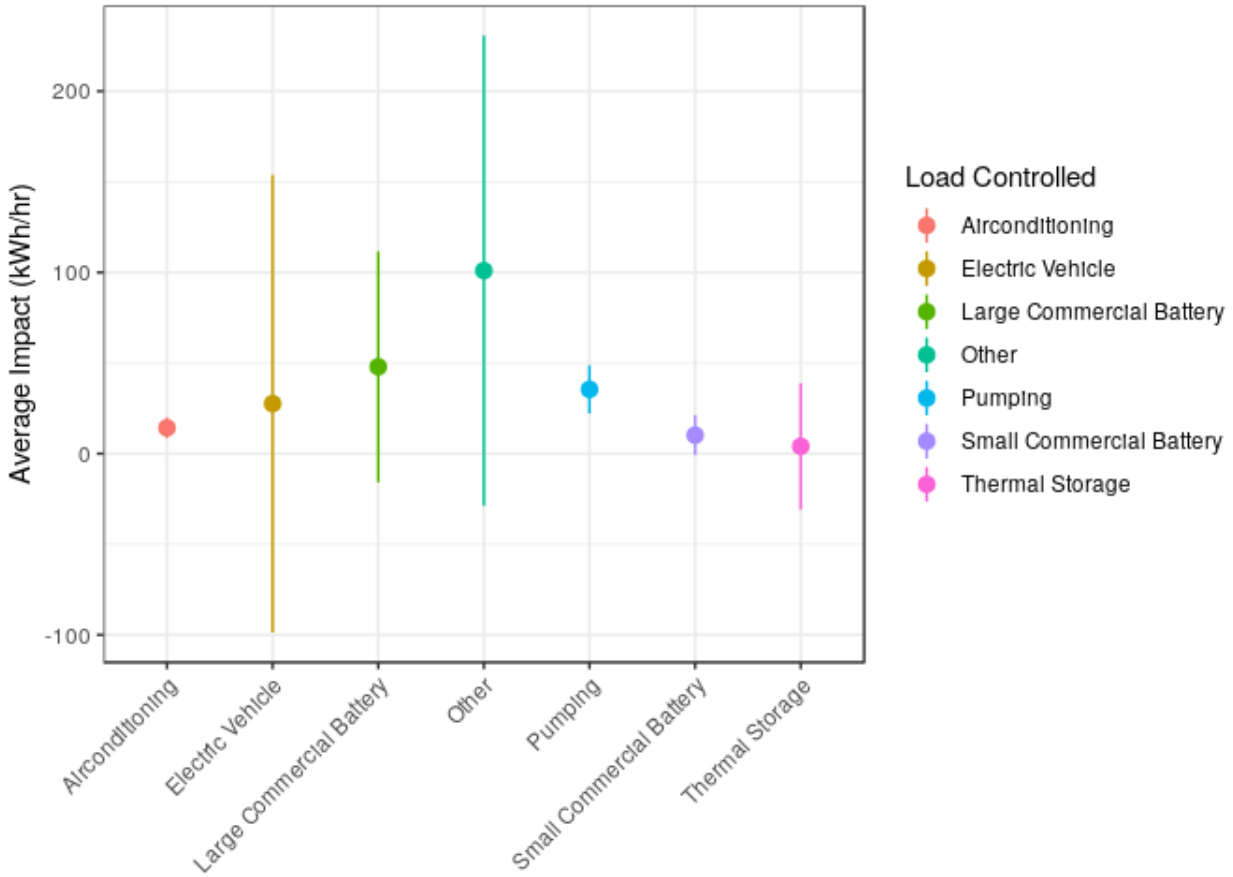
Electric vehicles, large commercial batteries, other, and thermal storage impacts have confidence intervals that cross zero (confidence intervals that cross zero do not mean that the estimate was not

[20] "Other" loads are cold storage and light manufacturing facilities.

measurable, but rather that there is just more uncertainty than if it didn't[21]). The "other" group is especially variable in terms of both customer demand and impacts, which results in higher uncertainty than if the customers were homogenous.

During 2019 test events, commercial electric vehicle charging often dropped out after the first hour (this appears to be a continuation of their practice under DRAM rules that draw upon the best hour of event impacts), which is most likely the cause for the confidence interval to cross zero.[22] The event drop-out was especially visible in the specific LCAs where one of Leap's partners operates. For example, in one LCA, the first hour had an impact of 31% of baseline and the second hour had a -1% impact. Leap indicates that they have discussed this issue with their partner and that they are "ensuring that our partners are committed and incentivized to perform across entire test and event windows, and we therefore expect performance to increase for multi-hour tests and events in 2020 and beyond."

Figure 15. Nonresidential ex post per-meter impacts (average of April - November) by load type



Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

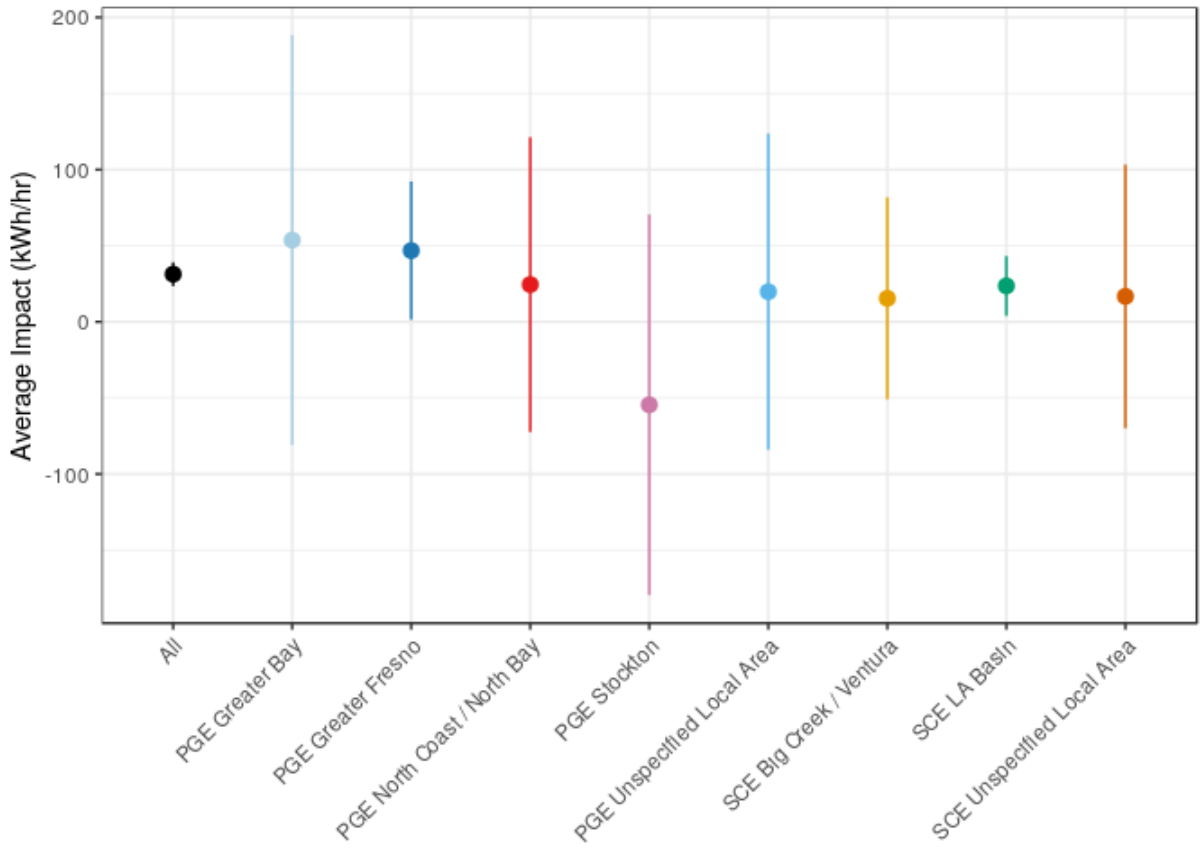
[21] The American Statistical Association statement on statistical significance states, "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold." Wasserstein, Ronald L., and Nicole A. Lazar. *The ASA Statement on P-Values: Context, Process, and Purpose*. Taylor & Francis, 2016.

[22] Leap indicated that one of their electric vehicle partners optimized for one hour test events as the DRAM contract for Generic capacity stipulated that performance would be calculated for the highest hour.

LCA-specific Results

Figure 16. shows the average ex post impacts grouped by LCA. Impacts are a function of the load type in each area more than by temperature. Stockton, with a negative load impact, included only one participating meter. Greater Bay has the highest average impact (54 kW) and the widest variation (-51 to 158 kW). The high variation in Greater Bay is due to high uncertainty in EV impact estimates as most of the participating meters in Greater Bay are for EV charging loads.

Figure 16. Nonresidential ex post average impacts (average of April - November) by location (Local Capacity Area)



Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

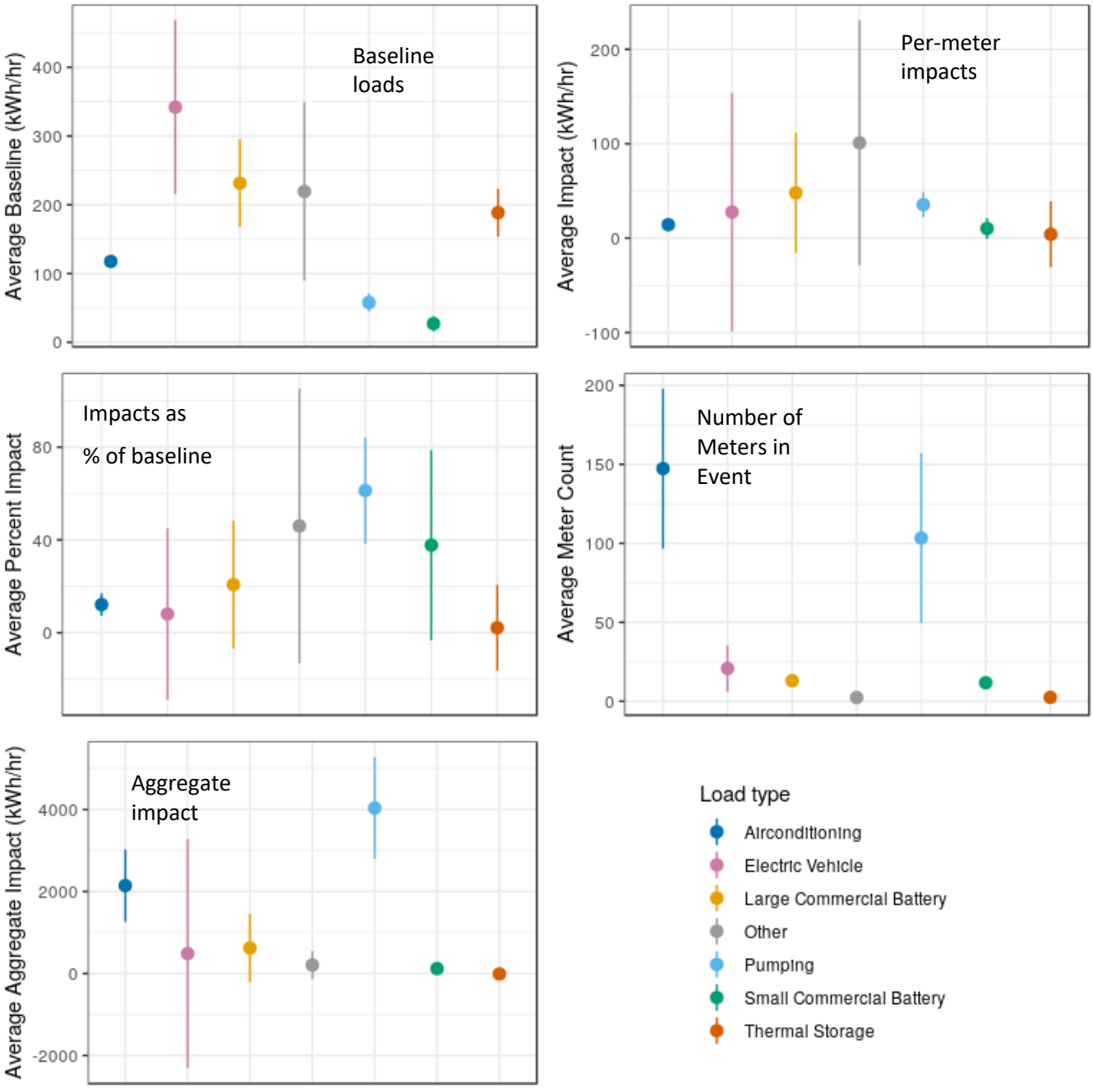
Determinants of aggregate results

Aggregate impacts are population weighted per-meter impacts for meters included in test events.[23] Per-meter impacts are a function of baseline loads and type of load. For the nonresidential impacts, only the air conditioning loads are driven by temperature. Figure 17 presents ex post summaries of per-meter baseline loads, per-meter impacts, impacts as a percent of baseline, and average event meters, followed by the aggregate impacts. From this comparison, one can clearly see the relationship between the aggregate impact and per-meter impact and number of meters. For example, pumping has the highest

[23] The aggregate results do not include all meters enrolled with Leap as of the end of 2019. Results are from those meters that were dispatched during test events (See Table 9 for a count of meters in the test events by LCA and type).

aggregate impact (~4 MWh/hr), but has a moderate average impact and a high number of meters. While air conditioning has the highest number of meters, the low per-meter impact brings down the aggregate impact.

Figure 17. Nonresidential ex post monthly average baseline loads, per-meter impacts, impacts as a percentage of baseline loads, meter counts for each load type, and aggregate impacts (for meters in test events)



Note that each graph has a different y-axis range.
Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

Aggregate impacts within an LCA are a function of the specific load type. Table 10 shows the variation in load type and number of meters included in ex post analysis for the load type across the LCAs.

Table 10. Nonresidential summary of 2019 meters included in test events by LCA and load type

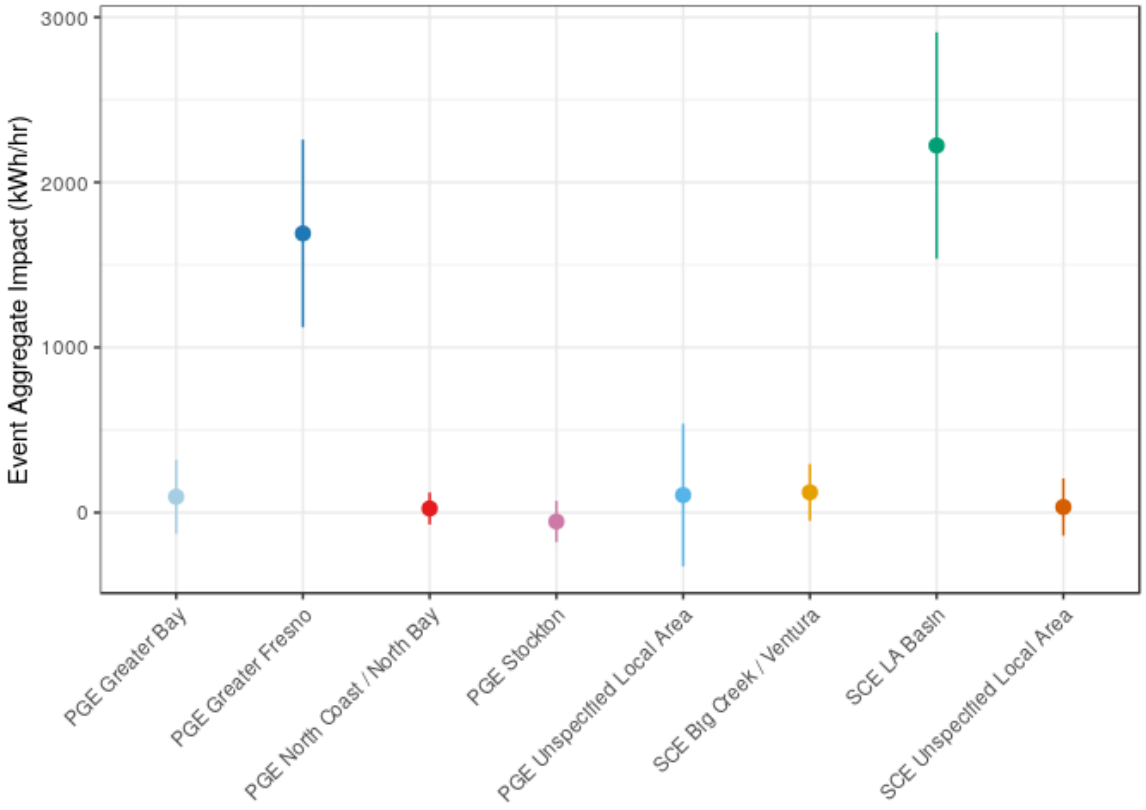
LCA	Air conditioning	Electric Vehicle Charging	Large Commercial Battery	Other	Pumping	Small Commercial Battery	Thermal Storage
Big Creek / Ventura	28	<15	<15	<15	<15	<15	
Greater Bay		16		<15			<15
Greater Fresno		<15	<15	<15	55		
LA Basin	135	15	<15	<15	138	<15	
North Coast / North Bay		<15				<15	<15
Stockton		<15					
Unspecified Local Area	<15	<15			<15	<15	

Note: Leap has nonresidential meters located in Sierra (3 meters) that were not part of test events and so are not modeled in the ex post analysis.

Aggregate Ex Post Summary

Among the LCAs, the LA Basin and Greater Fresno have the top-two participating meter counts during the ex post events which overcomes low per-meter impacts placing them toward the top of aggregate impacts. By comparison, Big Creek/Ventura has a higher average per-meter count during the events than Greater Bay, but a lower average impact which leads both Big Creek/Ventura and Greater Bay to have comparable aggregate impacts. Stockton, on the other hand, shows an aggregate impact that is below zero because impacts are just from electric vehicle charging stations in the LCA and most of them participated for only half the test event period (see discussion above under Customer Type Results).

Figure 18. Nonresidential ex post aggregate load shed
(overall, by LCA for meters in test events)



Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

Nonresidential Ex Post Results for Leap's Full Resource

Leap added customers throughout 2019, with many being added after test events were called. As shown in Table 11, when extrapolating the per-meter ex post results described earlier in this section to the full resource available as of the end of 2019, Leap's nonresidential sector provides an impact of 26.6 MW.

Table 11. Nonresidential ex post impacts for Leap's full resource

ik	Ex Post Mean Event Temperature (F)	Ex Post Mean Impact (kW)	Ex Post Mean Baseline (kW)	Ex Post Percent Impact (%)	Full Resource Enrollment Count	Full Resource Total Impact (MW)	Full Resource Std.Err. (MW)	90% Confidence Interval (MW)
Airconditioning	80.3	14.2	117.6	12	203	2.89	0.59	(1.93, 3.85)
Electric Vehicle	81.6	27.6	342.0	8	139	3.83	8.89	(-10.79, 18.46)
Lg Comm Battery	85.6	47.9	231.5	21	21	1.01	0.68	(-0.11,2.12)
Other	75.8	101.0	219.4	46	64	6.46	4.21	(-0.46, 13.38)
Pumping	84.7	35.5	57.9	61	320	11.34	2.14	(7.82, 14.87)
Sm Comm Battery	79.7	10.2	27.1	38	98	1.00	0.55	(0.10, 1.91)
Thermal Storage	86.2	4.0	188.5	2	3	0.01	0.05	(-0.08, 0.10)

All	***82.5***	***31.3***	***93.5***	***34***	***848***	***26.55***	***4.02***	***(19.94. 33.16)***

Residential

Number of events

For the residential sector, CDA identified and estimated ex post impacts for 16 discrete Leap events in 2019 for air conditioning loads. All events were test events. As described in the methods section, CDA synthesized events for the residential batteries for the months of May, June, July, August, October, November, and December 2019 using 15-minute data from rate arbitrage dispatch.

Table 12 below presents the residential test event information. August has the most test events and a higher number of participants, on average. For the two months that included air conditioning loads (June and August), aggregate impacts were about a third of baseline values.

Test events always included a subset of the total customer portfolio, so Leap's aggregate impacts across their entire portfolio in a given month and in 2019 exceeded any of the individual aggregate impacts noted in the table below.

Table 12. Residential summary of 2019 events, conditions, enrollment, and impacts

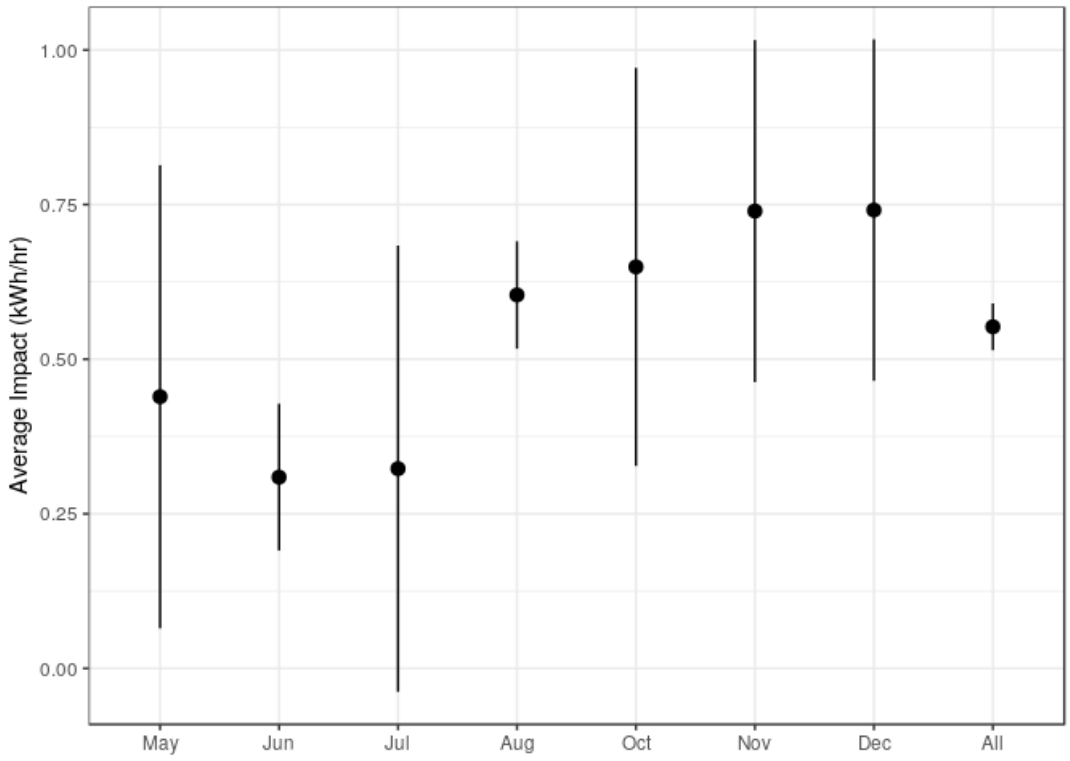
month	# of events	average # of participant meters	average temperature (F)	baseline load (kW)	per-meter impact (kW)	aggregate baseline load (kW)	aggregate impact (kW)	impact %
May	1	89	71	0.65	0.44	58	39	68%
Jun*	7	264	76	0.96	0.31	398	136	32%
Jul	1	102	81	1.30	0.32	132	33	25%
Aug*	11	720	89	1.82	0.60	3226	1027	33%
Oct	1	101	78	1.07	0.65	108	66	61%
Nov	1	99	60	0.96	0.74	95	73	77%
Dec	1	101	53	1.12	0.74	113	75	66%

Note: Participant meter counts for residential batteries in ex post events from Leap solar partner.
**Includes air conditioning and storage battery meters in events. All other months are impacts for storage batteries only.*

Average Ex Post Load Impacts by Test Event Months

Figure 14 depicts the range of per-participant impacts by month for all events called in 2019. The highest monthly weighted average per-participant impact are November and December with 0.74 kW while June is less than half that value (at ~0.31 kW). Months with only the residential battery impacts (May, July, October, November, and December) have larger confidence intervals because these are based on just over 100 meters, while residential AC has thousands of dispatchable customers and several hundred that participated in each event.

Figure 19. Residential ex post monthly average event impacts

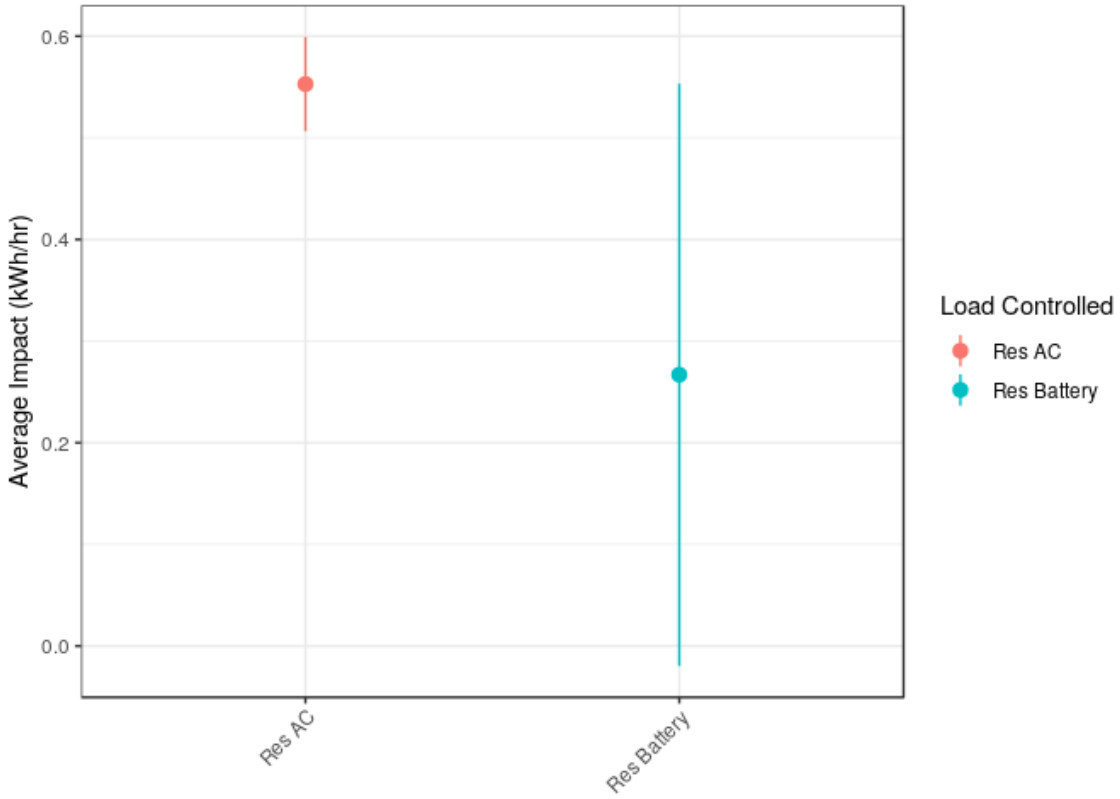


Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

Average Load Impact by Load Type

Figure 20 shows the two residential loads controlled by Leap. The average air conditioner load impact (0.55 kWh/hr) is about twice the average residential battery load impact (0.27 kWh/hr). The large confidence interval for batteries is primarily due to variation in their baseline loads given their small sample size of just over 100 meters.

Figure 20. Residential ex post average impacts by load type for all test events

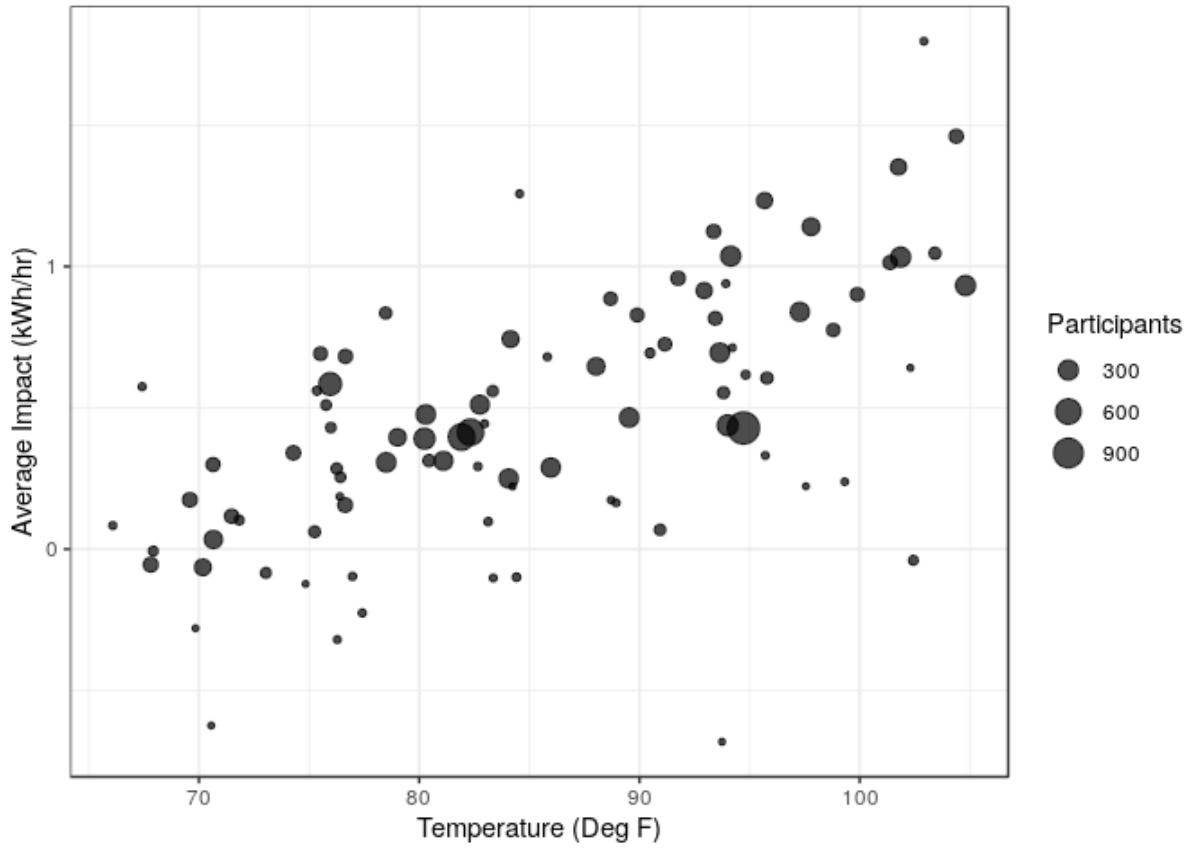


Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

Average Load Impacts by Temperature and Time of Day

In Figure 21, we see the relationship between air conditioning per-meter event impact and temperature based on monthly event roll-ups. The figure depicts the average impacts per-meter (y-axis) for June and August events called in 2019 vs. the population weighted outside temperature during the events (x-axis), with dot sizes roughly corresponding to participant counts within the LCAs. The data confirms a rough correlation between impacts and outside temperature, with plenty of variability caused by other factors.

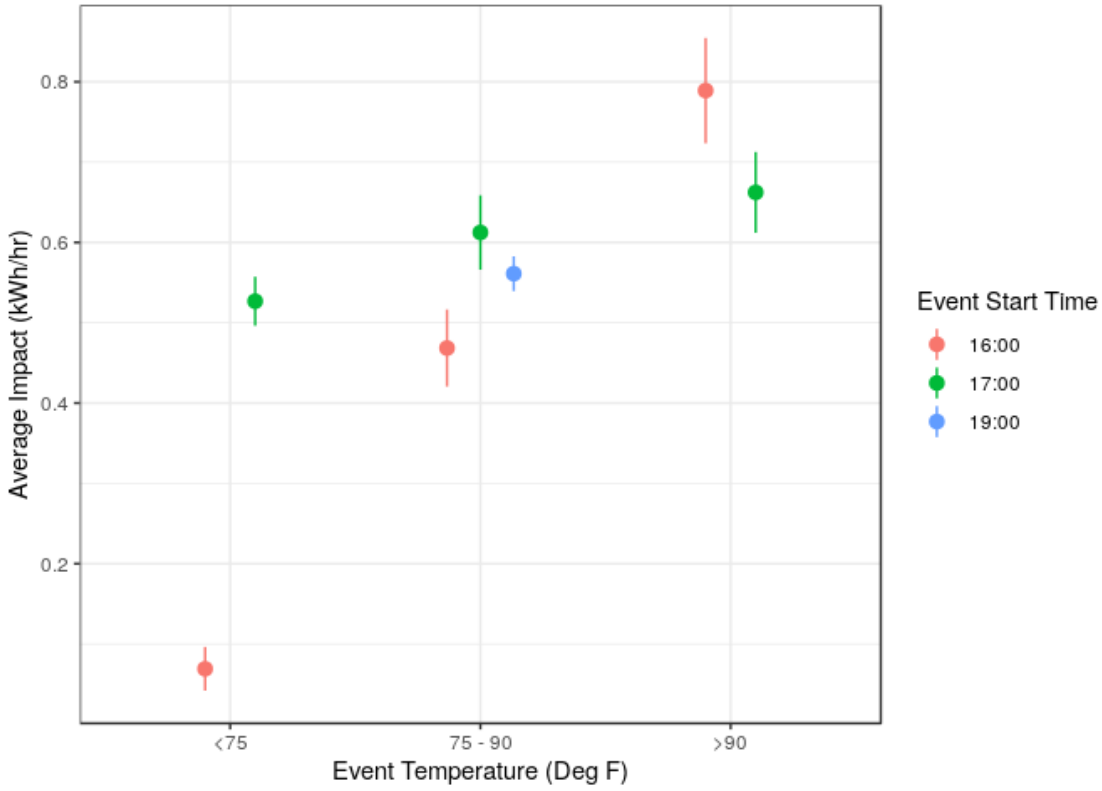
Figure 21. Residential ex post scatter plot of monthly average impacts vs. outside temperature (air conditioning loads only)



Each dot in the figure represents AC participants within a specific LCA during an event. Dot size indicates number of participants.

Event timing can impact outcomes, but for the residential air conditioning impacts, temperature effects were larger. Figure 22 provides a view of the average per-meter impacts for each event. The event start time is shown by different colors and the hourly temperature ranges during the events are the x-axis. Events that began when it was less than 75 degrees (start times of 4 PM and 5 PM) have the lowest impacts while events with temperatures over 90 degrees (also start times of 4 PM and 5 PM) had the highest impacts.

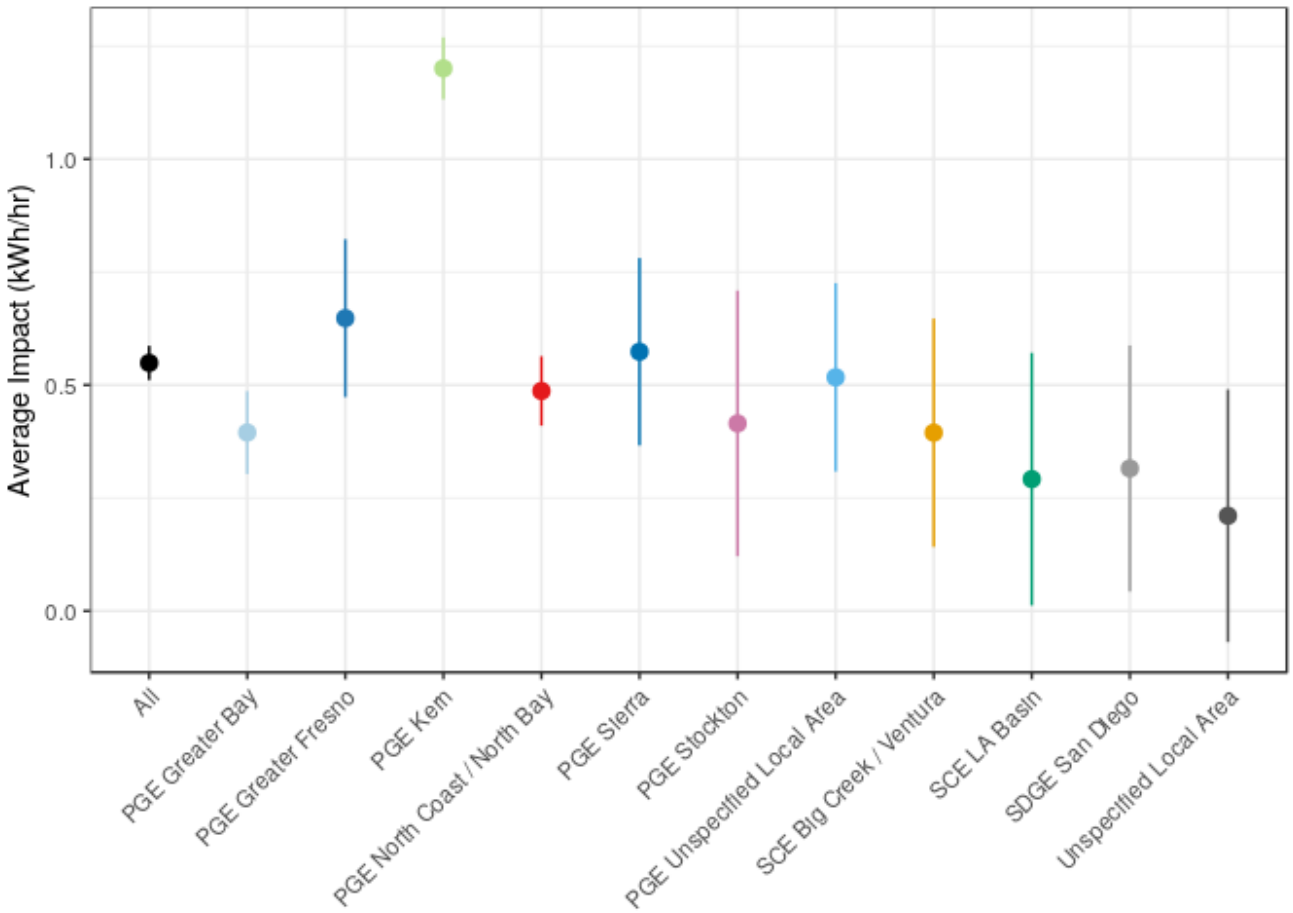
Figure 22. Residential ex post event average impact by time of day
(air conditioning load only)



Note: Event start time 16:00 is 4 PM, 17:00 is 5 PM, and 19:00 is 7 PM
Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

Among the LCAs, Kern, Greater Fresno, and Sierra have the highest temperatures during the test events and therefore also have the higher per-meter impacts for air conditioning, which makes up the majority of the residential load. Notably, Kern has much higher average baseline than the other LCAs, which helps to explain the higher impacts. San Diego and LA Basin, with only battery storage impacts, have among the lowest per-meter impacts.

Figure 23. Residential ex post average per-meter impacts for all events grouped by LCA



Circles are the average across all participant meters in the monthly event. Lines are the 90% confidence intervals.

Determinants of aggregate results

Aggregate impacts are population weighted per-meter impacts. Much of the residential Leap impacts are from air conditioning, which is driven by seasonal factors. Residential batteries, on the other hand, show little seasonal difference (i.e. there is sufficient insolation even in the winter to fill the battery).

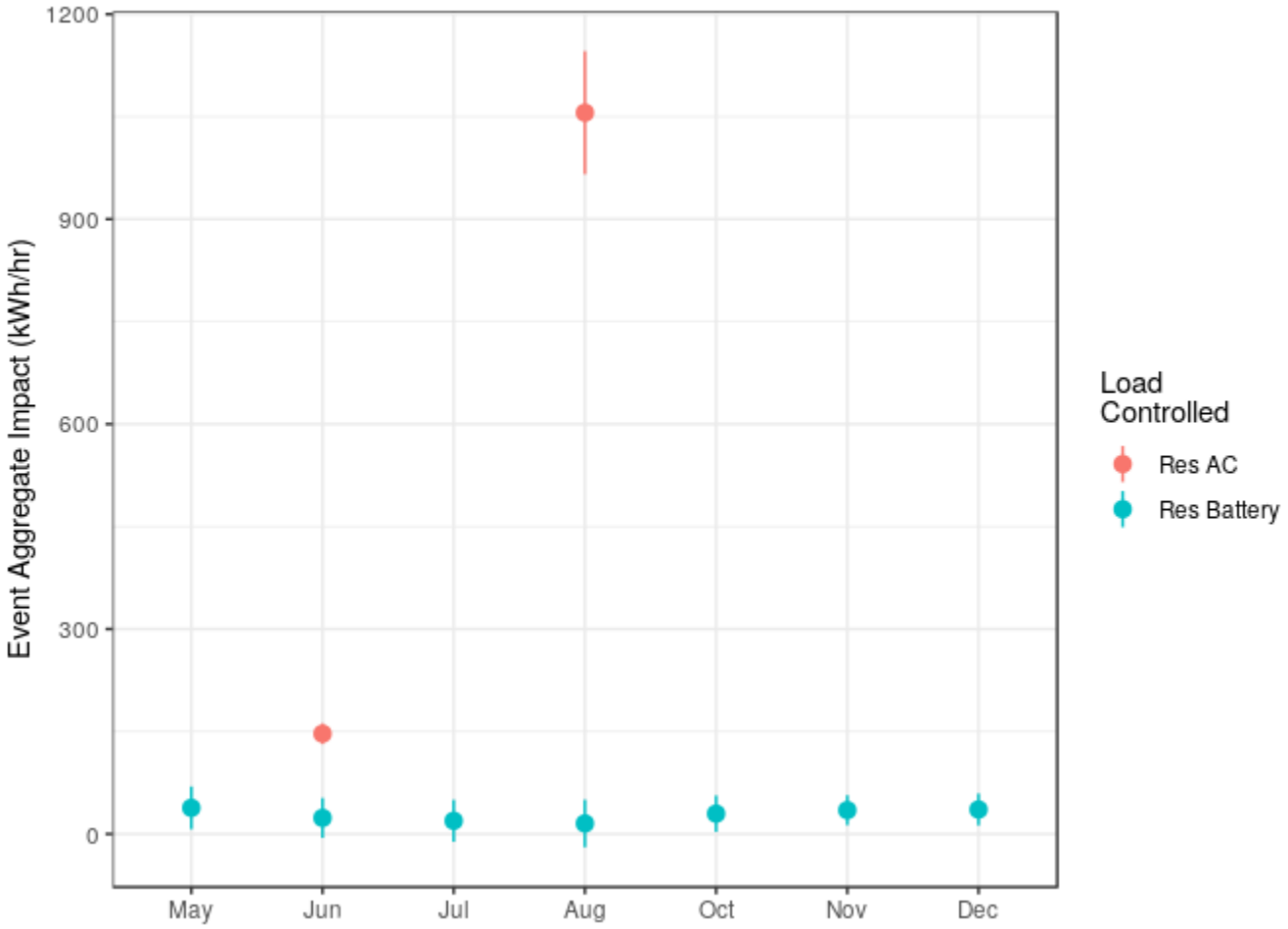
Aggregate Ex Post Summary

Aggregate impacts are population weighted per-meter impacts for meters included in Leap test events (residential air conditioning) or CDA simulated test events (residential batteries).[24]

Figure 24 depicts the monthly roll-up event aggregate impacts by load type. With over an order of magnitude more participants, air conditioning has a higher impact than residential batteries in the two months when there were test events, topping out at about 1 MWh/hr in August.

[24] The aggregate results do not include all meters enrolled with Leap as of the end of 2019. It is just those meters that were dispatched during test events (See Table 9 for a count of meters in the test events by LCA and type).

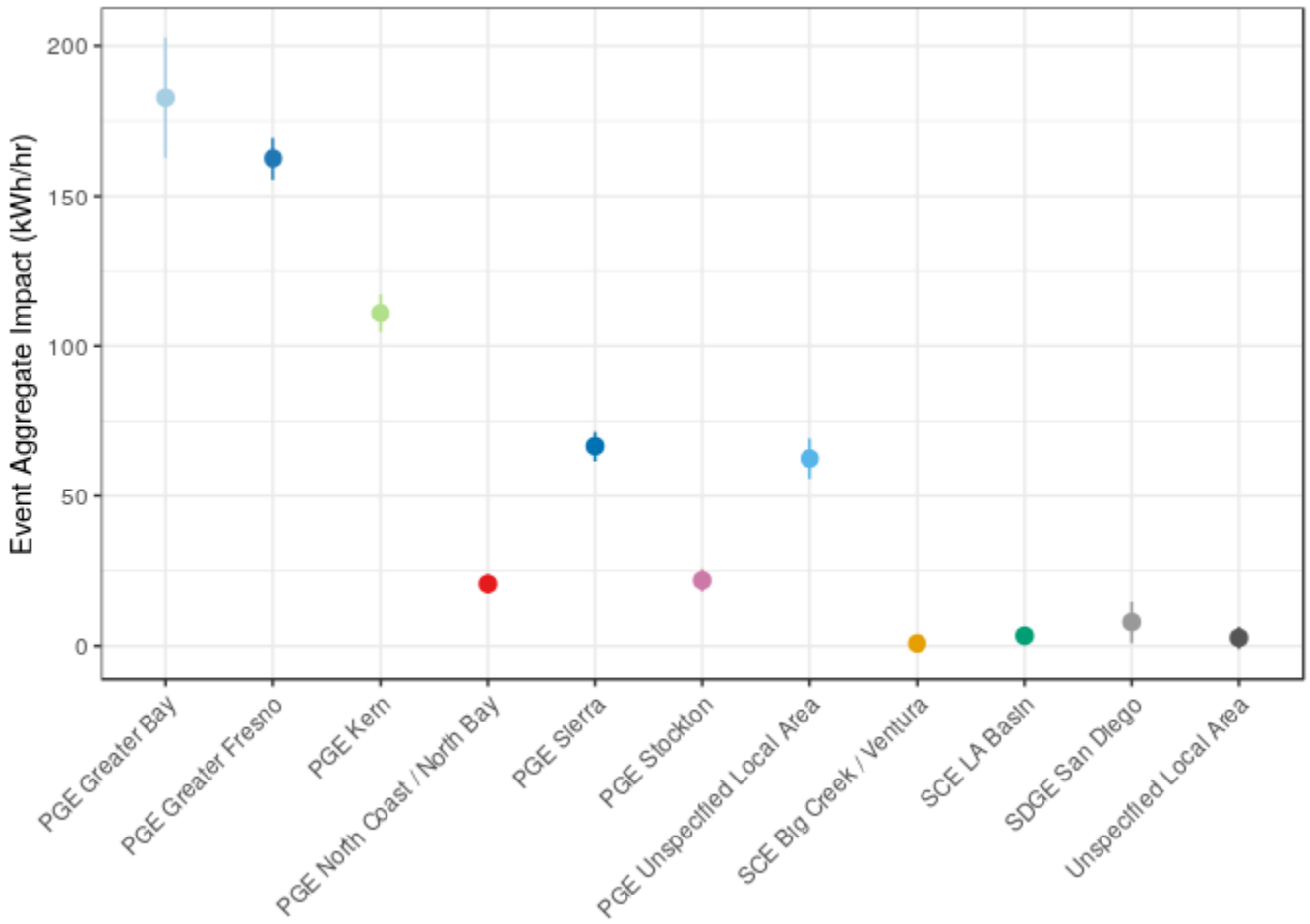
Figure 24. Residential ex post aggregate load impact
(overall, by month and load type for meters in test events)



Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

Greater Bay and Greater Fresno have the two highest meter counts and the two highest aggregate impacts. While Greater Bay has lower per-meter impacts than Greater Fresno (shown in Figure 23), there are close to twice as many meters in the Greater Bay area, thus bringing in more aggregate load impacts. Big Creek/Ventura has a per-meter impact comparable to other LCAs but has lowest aggregate impact. This is because there are very few meters in this LCA.

Figure 25. Residential ex post aggregate impact across all event grouped by LCA
(for meters in test events)



Circles are the average across all participants in the monthly event. Lines are the 90% confidence intervals.

Residential Ex Post Results for Leap's Full Resource

Leap added customers throughout 2019, with many being added after test events were called in 2019. When applying the per-meter ex post results described earlier in this section to the full resource available as of the end of 2019, Leap's nonresidential sector provides 3.9 MW.

Table 13. Residential ex post impacts for Leap's full resource

Load Type	Ex Post Mean Event Temperature (F)	Ex Post Mean Impact (kW)	Ex Post Mean Baseline (kW)	Ex Post Percent Impact (%)	Full Resource Enrollment Count	Full Resource Total Impact (MW)	Std. Err (MW)	90% Conf. Interval (low, high) (MW)
Air Conditioning	86.9	0.6	1.7	33	6990	3.86	0.16	(3.59, 4.13)
Battery	71.4	0.3	0.8	34	108	0.03	0.02	(0.00, 0.06)
Total	***86.4***	***0.5***	***1.6***	***33***	***7098***	***3.89***	***0.16***	***(3.63, 4.16)***

Enrollment Forecast

To make ex ante load forecasts, a forecast for changes from the ex post values is needed by SubLAP and customer category. CDA used forecasts provided by Leap for this purpose. Leap had significant growth in 2019 and projects increasing impacts for the next several years.

In most program evaluations with which we are familiar, the company that runs the program provides estimates of the projected number of enrollees in the future, and this is then multiplied by the load shed per customer as derived from historical data in order to forecast the future load impact. Given the heterogeneity of their load types and their internal business focus on capacity (including contracts with partners with deliverables in terms of capacity), Leap's forecasting alters this process by directly forecasting the load impact they expect to provide in the future. If we (CDA) were to accept Leap's forecasts at face value, then there is no connection between the ex post and ex ante results, and we could simply summarize Leap's forecasts and present them as ex ante results. Instead of following that hollow procedure, we interpret Leap's forecasts as expectations of how much their customer base will scale up from its current level, which is how enrollment forecasts can be interpreted as well. We note that Leap's forecasting approach, though unconventional, is not inherently more subject to error than the approach of forecasting by the number of new customers.

For example, if Leap gives their Residential AC impact as 6 MW in 2019 and projects 15 MW in 2020, this is an increase by a factor of 2.5. For purposes of the evaluation we accept that Leap will attain the number of meters in the category that would give them their forecasted load impact if the impact per meter is the same that it was during 2019, but we do not assume that the impact per meter will in fact remain constant:

- Leap plans to expand into LCAs where they do not currently operate and where they therefore have little data, or in some cases no data, available for the ex post analysis. For instance, several LCAs had no residential air conditioning meters in 2019.
- 2019 events took place only in certain months, but we need to predict capacity for all months of the year. For example, residential air conditioning events took place only in June and August in 2019.

CDA used a Bayesian analysis that gives a way to quantify how big those two effects are likely to be. That is, based on what we know, how likely is it that the un-observed LCAs are very different from the ones we know? The model also gives us reasonable uncertainty bounds on the aggregate load shed predictions.

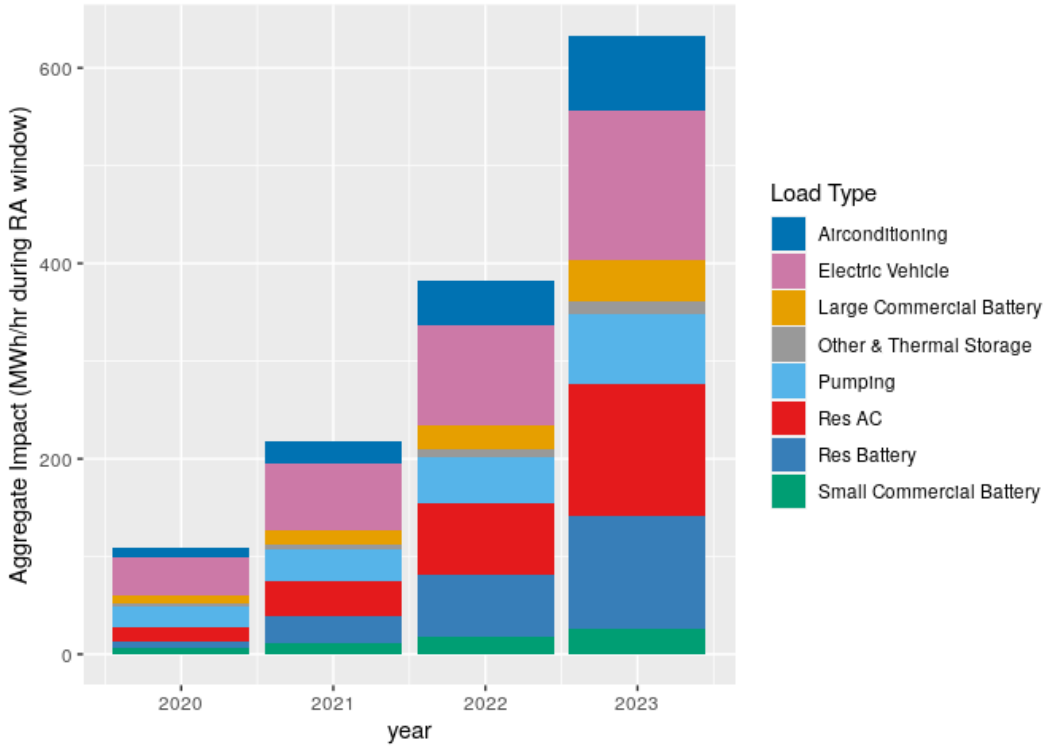
Enrollment Forecast Rationale

CDA did not produce the enrollment forecast used for the ex ante modeling. The forecast was provided by Leap based on their internal growth models used for budgeting and planning and CDA made no changes. Leap's forecast rationale is included in Appendix B: Leap enrollment forecast rationale (note that this rationale is not available in the public version for confidentiality reasons).

Forecasted Enrollment

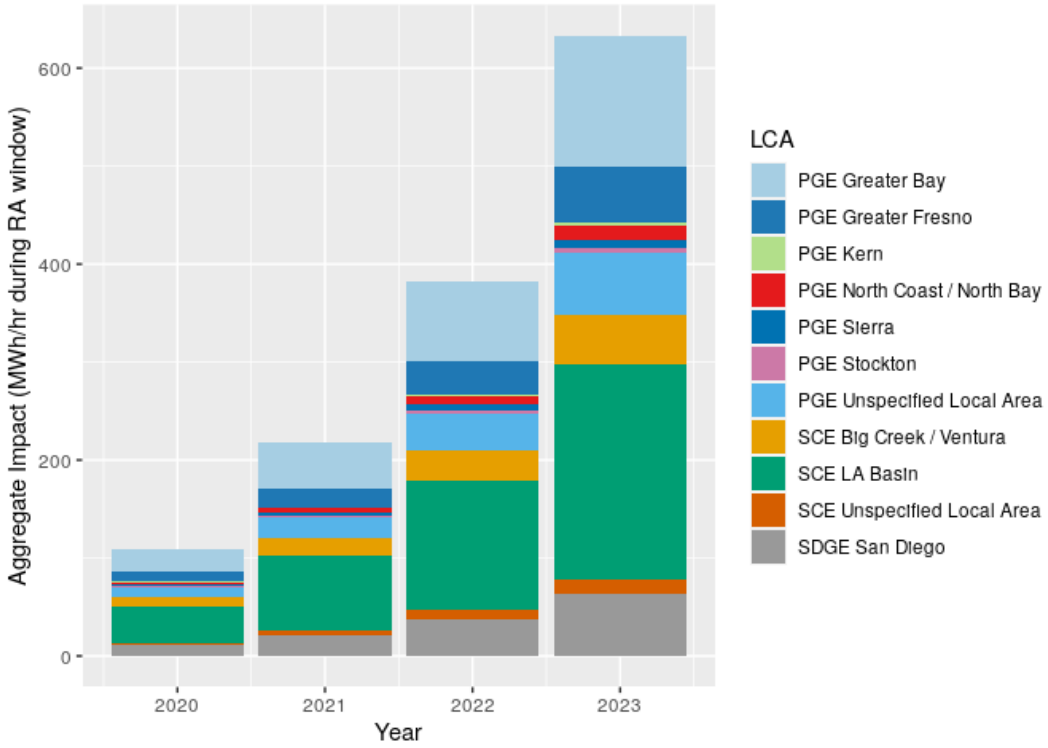
Leap projects the mix of customers among LCAs to change substantially. The projected growth in enrollment varies by SubLAP and is based upon both the data from Leap's enrollments to date, as well as Leap's forward-looking partner pipeline and planned recruitment efforts. Figure 26 shows the forecast impacts within the RA window for the medium forecast by load type.

Figure 26. Leap forecast of aggregate impacts for medium enrollment by load type during RA window



Leap forecasts the following aggregate impacts for medium enrollment forecast by LCA across both sectors.

Figure 27. Leap forecast of aggregate impacts for medium enrollment by LCA during RA window



Ex-Ante Results

Ex ante load impacts are based on predictions for standard event times and conditions. Predictions are made for two standard weather years – that is, two sets of monthly peak temperatures – corresponding to conditions that are expected to lead to peak electric load in one out of every two years (1-in-2) and one out of every ten years (1-in-10) on average. There is a slight difference in the conditions that cause peak load for the statewide California Independent System Operator (CAISO) and Investor-Owned Utilities (IOU) territories. Accounting for the two peaking conditions with the two weather years leads to four sets of standard weather data per territory.

Unless explicitly stated, all predictions discussed in this report are made for events in the Resource Adequacy (RA) time window that runs from 4 p.m. through 9 p.m. and for the medium enrollment forecast. (See at Figure 41 the end of the combined section to see predicted impacts for all three enrollment forecasts.)

We present the ex ante results for the nonresidential sector followed by the results for the residential sector. Lastly, we combine the two sectors to show the Leap resource across both sectors.

Nonresidential

Figure 33 shows the predicted load impact per meter for each of the four sets of weather conditions described above. All four sets of conditions produce similar predictions and have minimal variation by month (as much of the impacts are not weather sensitive). For these four ex ante weather-years the highest load impact per customer occurs in June, July and August, at about 25 kW per meter; the lowest occurs in five months (January, February, March, November and December) at about 22.5 kW per customer.

Figure 28. Nonresidential predicted per-meter load impact by month,
mean over RA hours, for four standard sets of weather conditions (medium forecast)

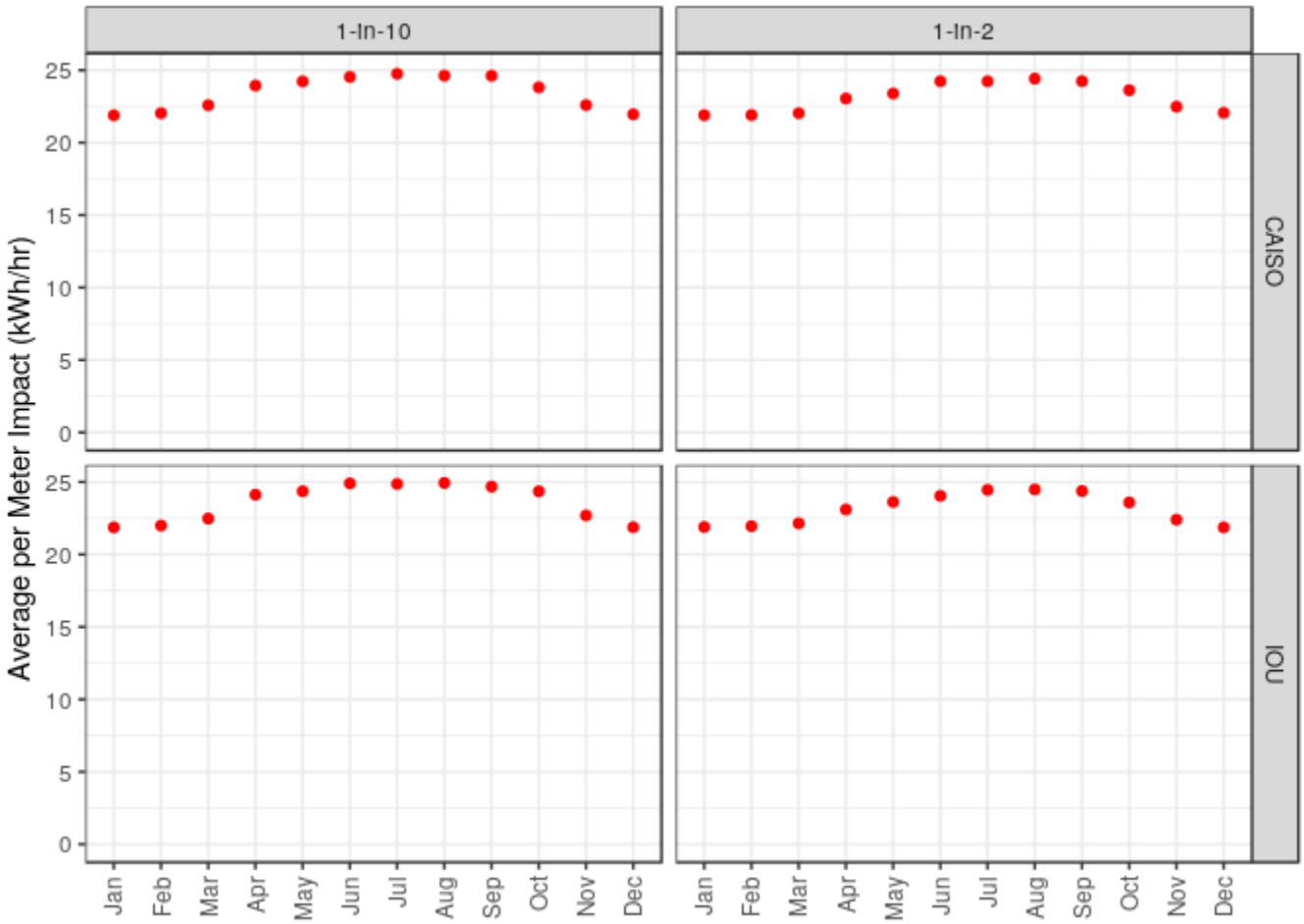


Figure 35 shows the predicted aggregate load impact for 2020; this is the product of Leap forecast enrollment in each month times the predicted load shed per meter in each month, again for the standard weather conditions. The maximum is around 83 MWh/hr statewide, in August.

Figure 29. Nonresidential predicted aggregate load impact by month,
mean over RA hours, for four standard sets of weather conditions in 2020 (medium forecast)

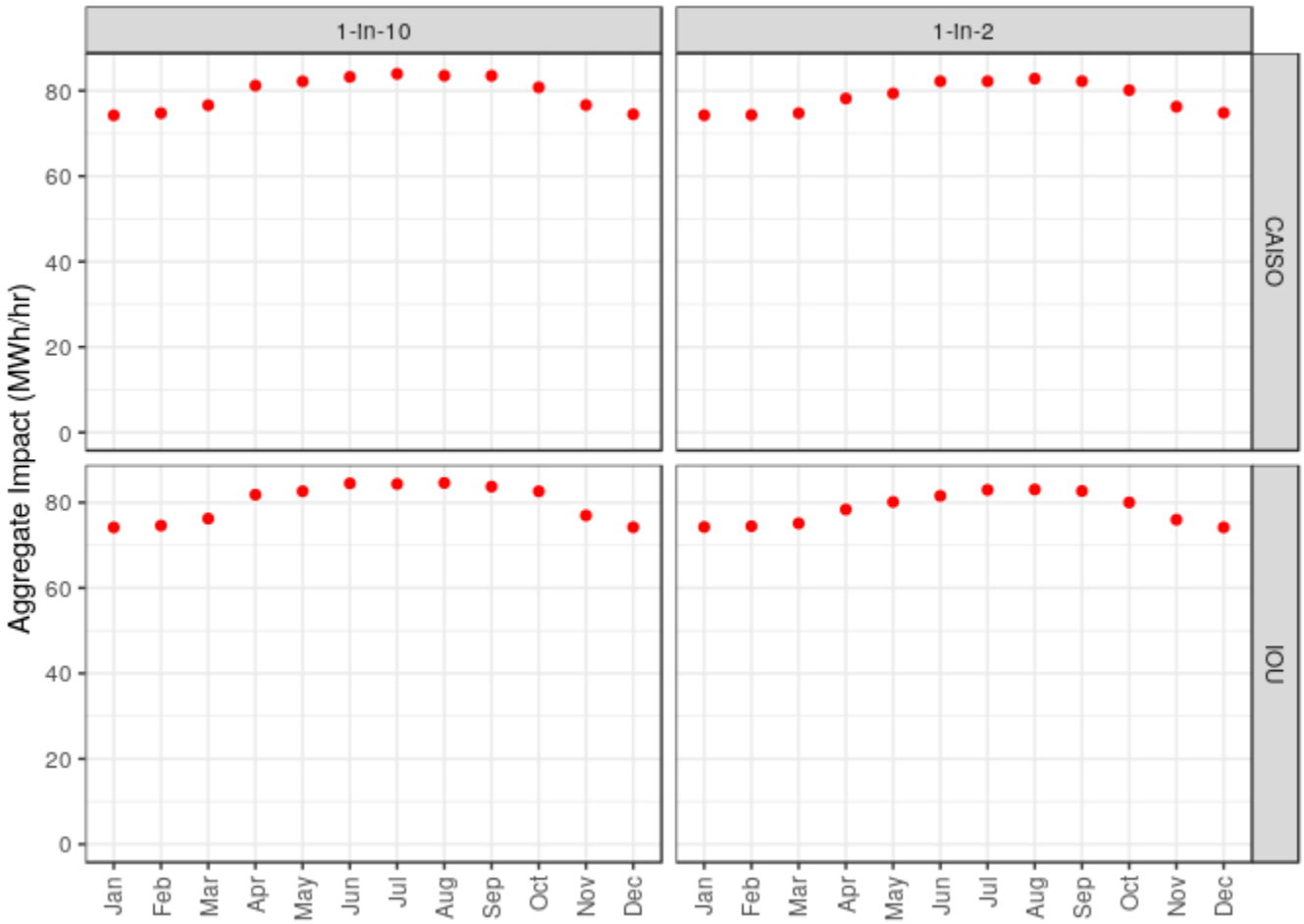


Figure 36 shows the same sort of information, this time for the next several years. The projected year-over-year increase in customers (and thus event participants) leads to a very rapid year-over-year increase in load impact.

Figure 30. Nonresidential predicted aggregate load impact, by month and year (medium forecast) mean over RA hours

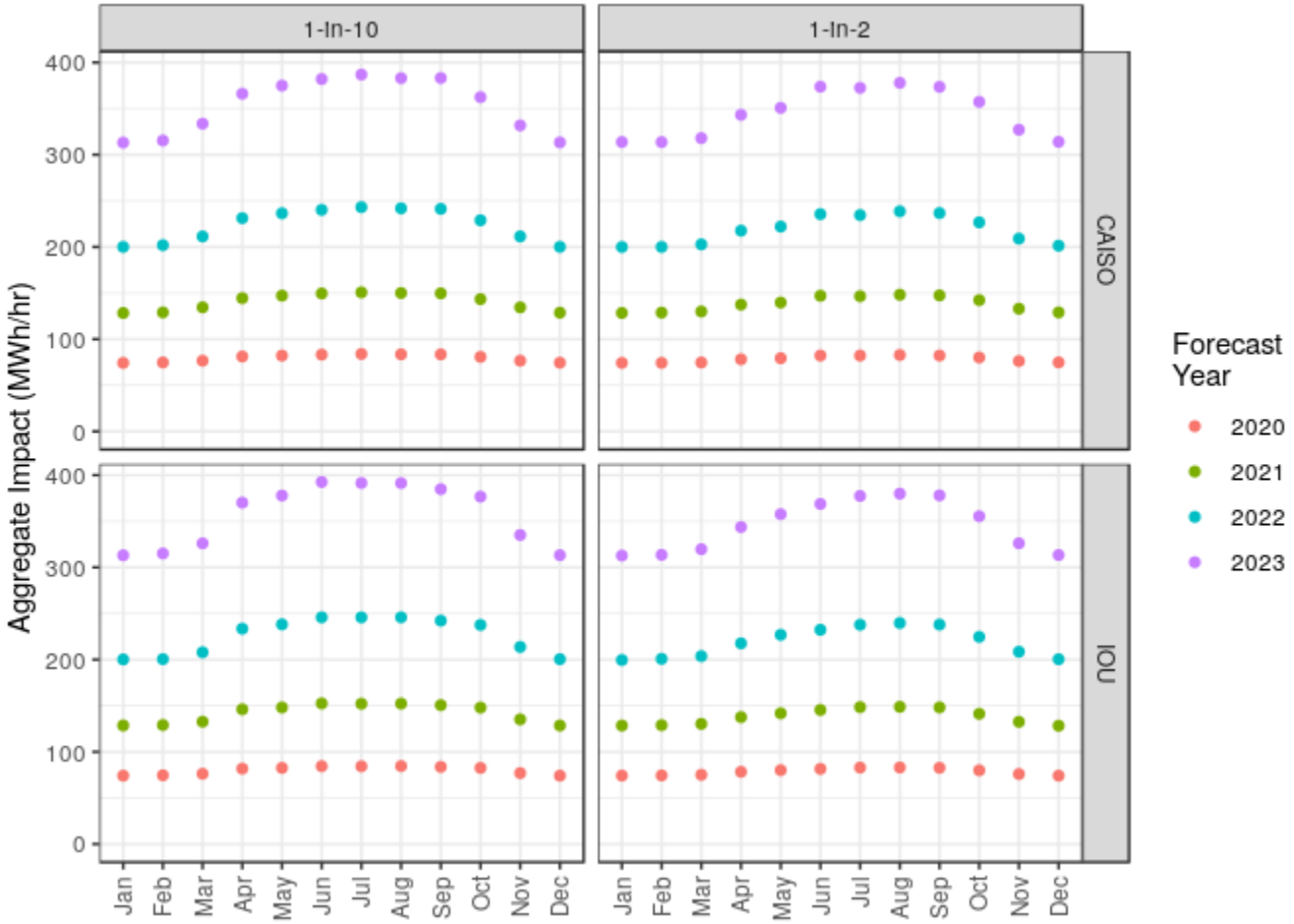


Table 15 summarizes the predicted load impact by year for the August 1-in-2 monthly CAISO peak day and the IOU 1-in-2 peak days. The forecast increase over the next three years brings about a four-fold increase in impacts.

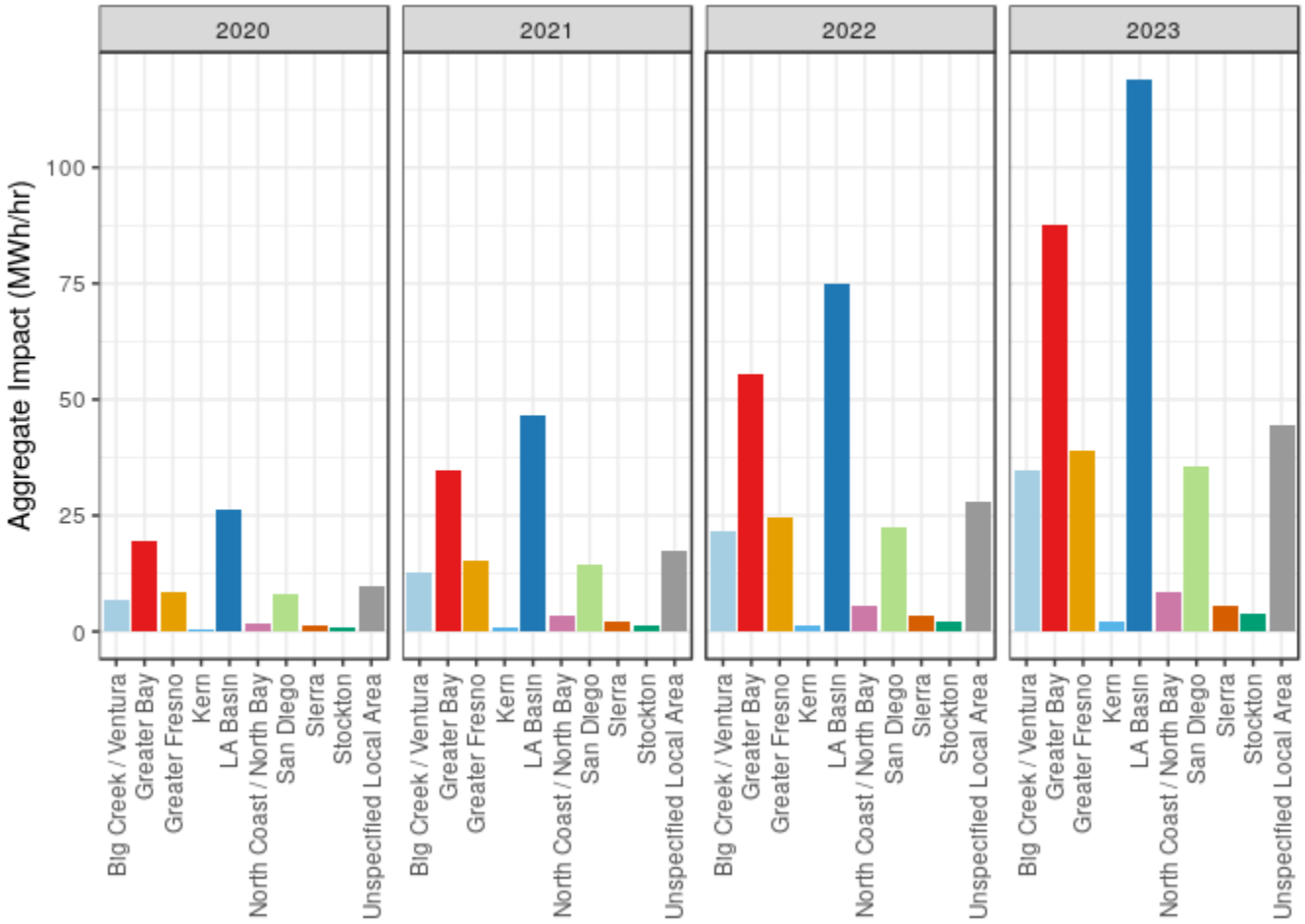
Table 14. Nonresidential predicted aggregate load impact for August CAISO 1-in-2 day and IOU 1-in-2 day

Year	CAISO 1-in-2 day		IOU 1-in-2 day	
	Temp (F)	Aggregate Impact (MW)	Temp (F)	Aggregate Impact (MW)
2020	85.0	82.8	85.6	83.1
2021	85.0	148.1	85.6	148.8
2022	85.0	238.6	85.6	239.7

2023	85.0	377.9	85.6	379.9

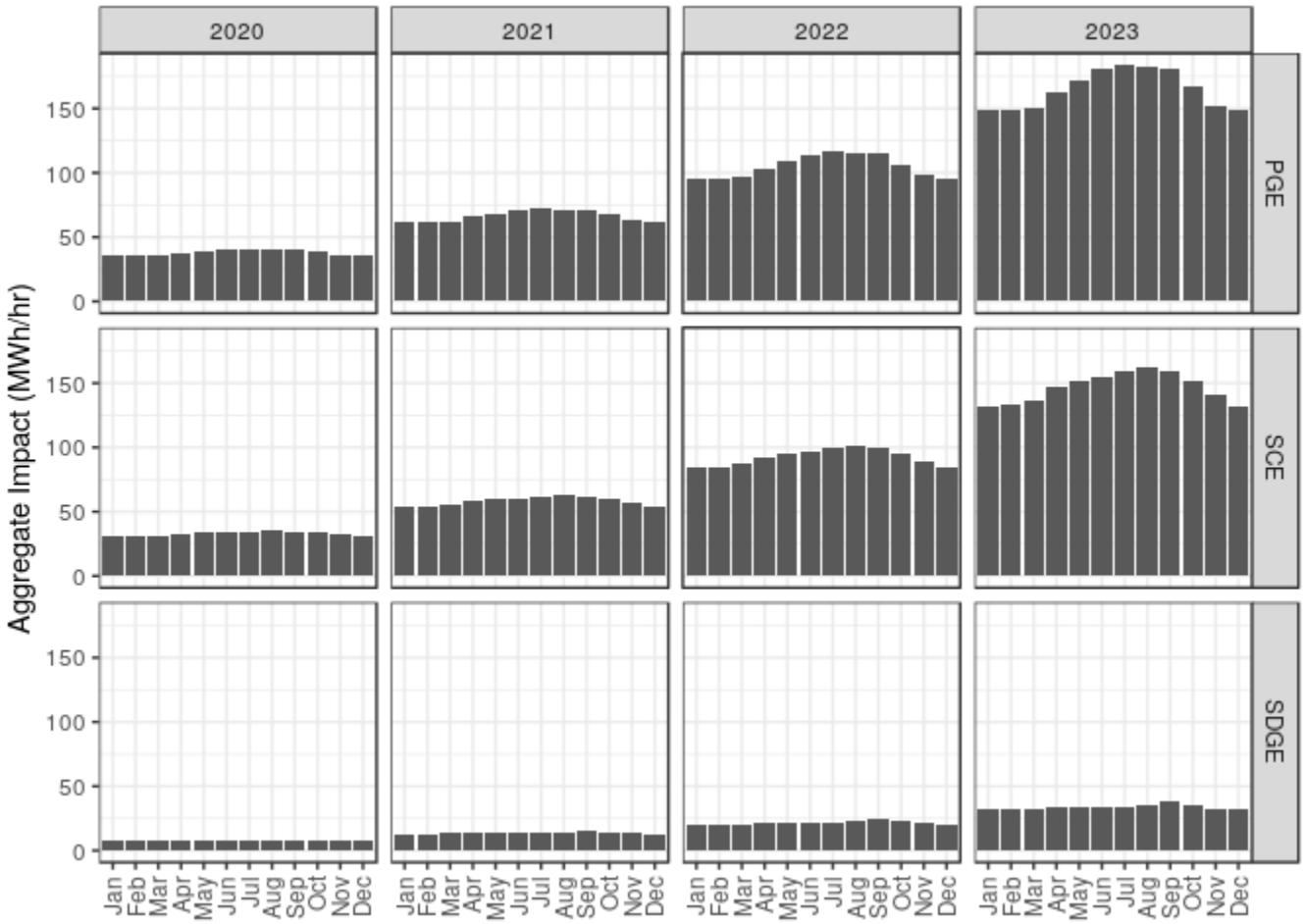
The ex ante predictions were made for each LCA. Figure 37 is an example: it shows the predicted aggregate impact for August, for the IOU 1-in-2 weather year for 2020-2023. Forecast enrollment for each LCA is growing over these years, with LA Basin and Greater Bay providing the highest impacts each year.

Figure 31. Nonresidential predicted aggregate impact by Local Capacity Area, mean over RA hours, for August of different years, IOU 1-in-2 weather data (medium forecast)



Aggregate impacts by month for the IOU 1-in-2 weather year are shown in Figure 38, this time separately for PG&E, SCE, and SDG&E and by year. PG&E brings in the highest aggregate impacts across all months.

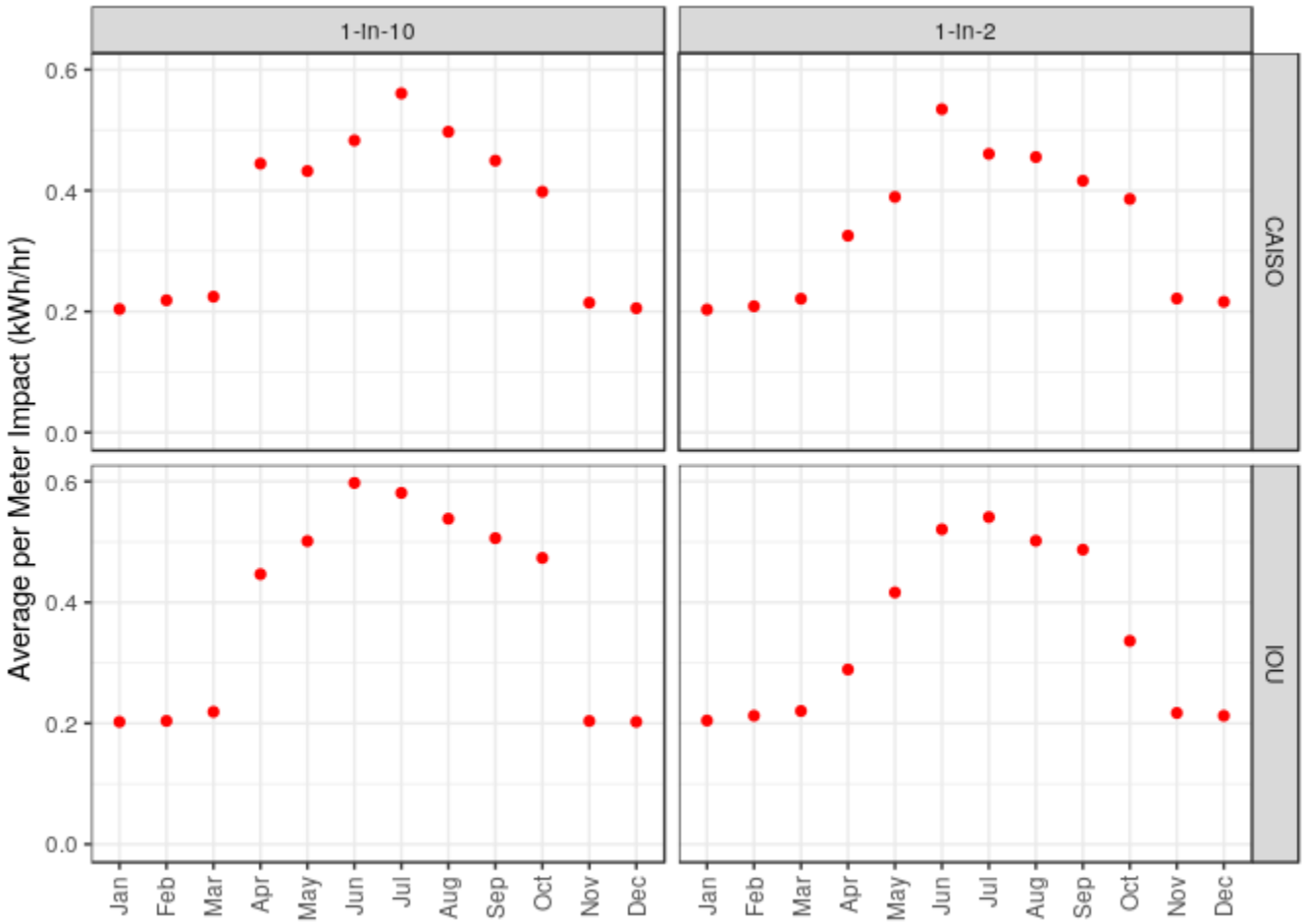
Figure 32. Nonresidential predicted aggregate impact by month, mean over RA hours, IOU 1-in-2 weather data, separately by utility (medium forecast)



Residential

Figure 33 shows the predicted load impact per meter for each of the four sets of weather conditions described above. All four sets of conditions produce similar predictions. For these four ex ante weather-years the highest load impact per customer occurs in June and July, at about 0.60 kW per customer; the lowest occurs in November through March, at about 0.20 kW per customer.

Figure 33. Residential predicted per-meter load impact by month, for four standard sets of weather conditions (medium forecast)



The load impacts vary by month and time of day. Figure 34 shows that air conditioning in July has some of the highest predicted load impact and that it falls across the RA window. Predicted impacts from batteries, on the other hand, are lower than air conditioning, have their highest impacts in the non-summer months, and increase across the RA window.

Figure 34 Residential predicted load impact, for IOU 1-in-2 weather year and by month and hour

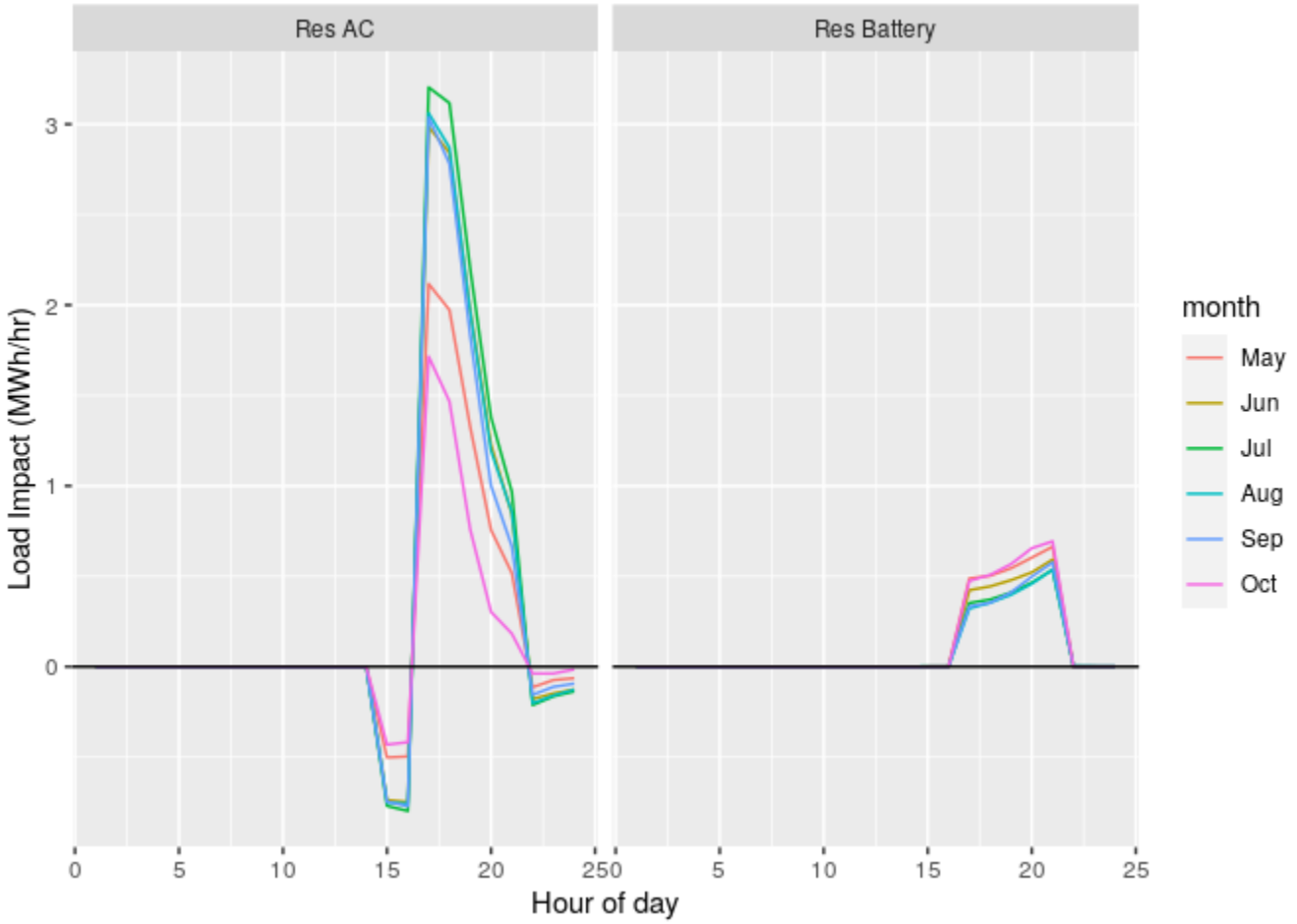


Figure 35 shows the predicted aggregate load impact for 2020; this is the product of Leap's forecast enrollment in each month multiplied by the predicted load shed per customer in each month, again for the standard weather conditions. The maximum is around 24 MW statewide, in June or July (depending on the specific weather condition. also note that this represents aggressive growth over the ex post full resource value for 2019).

Figure 35. Residential predicted aggregate load impact by month
for four standard sets of weather conditions in 2020 (medium forecast)

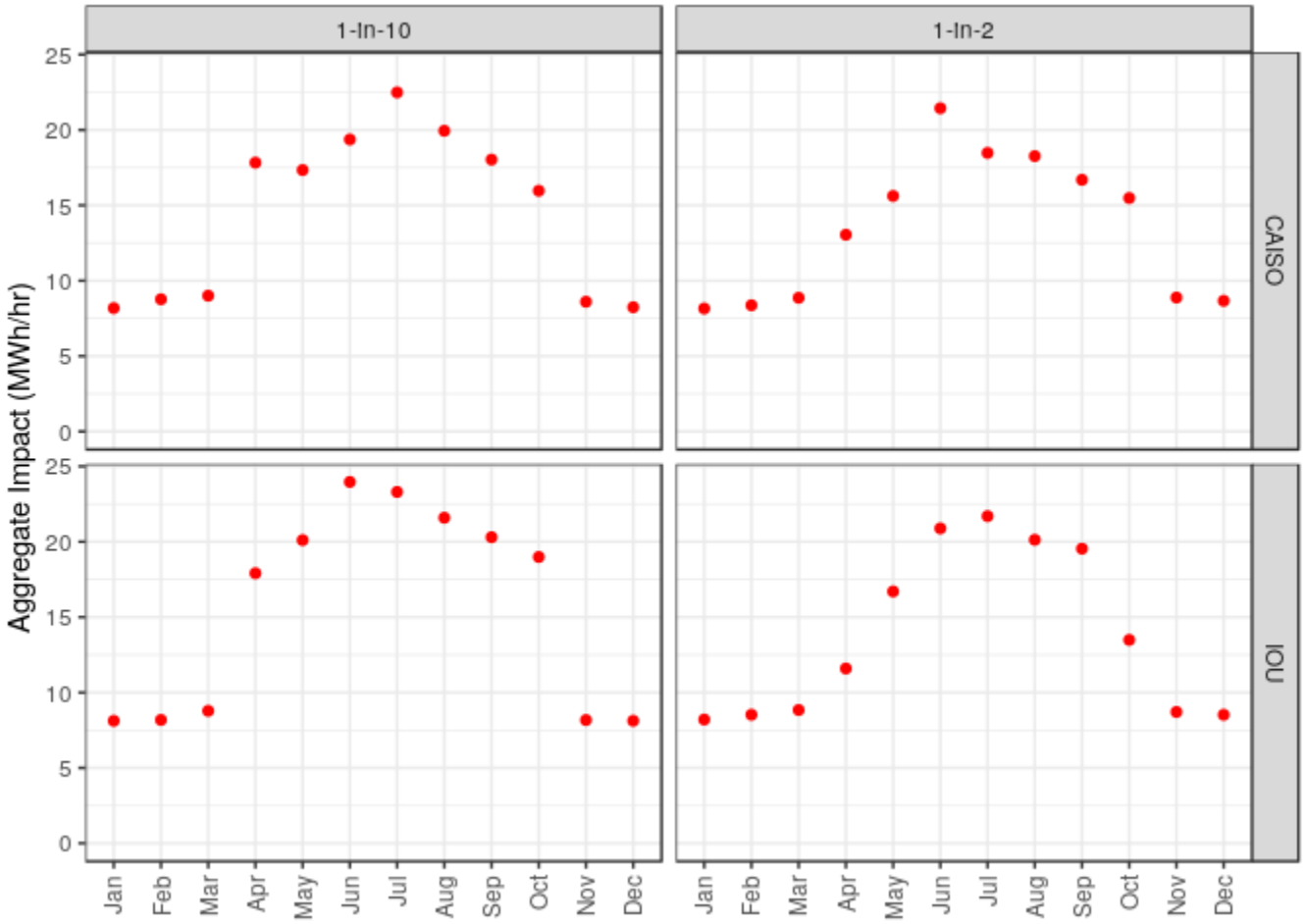


Figure 36 shows the same sort of information, this time for the next several years. The projected year-over-year increase in customers (and thus event participants) leads to a very rapid year-over-year increase in load impact. For example, August impacts for the IOU 1-in-2 weather year in 2020 are around 20 MWh/hr while August impacts in 2023 are about 200 MWh/hr higher.

Figure 36. Residential predicted aggregate load impact, by month and year, mean over RA hours (medium forecast)

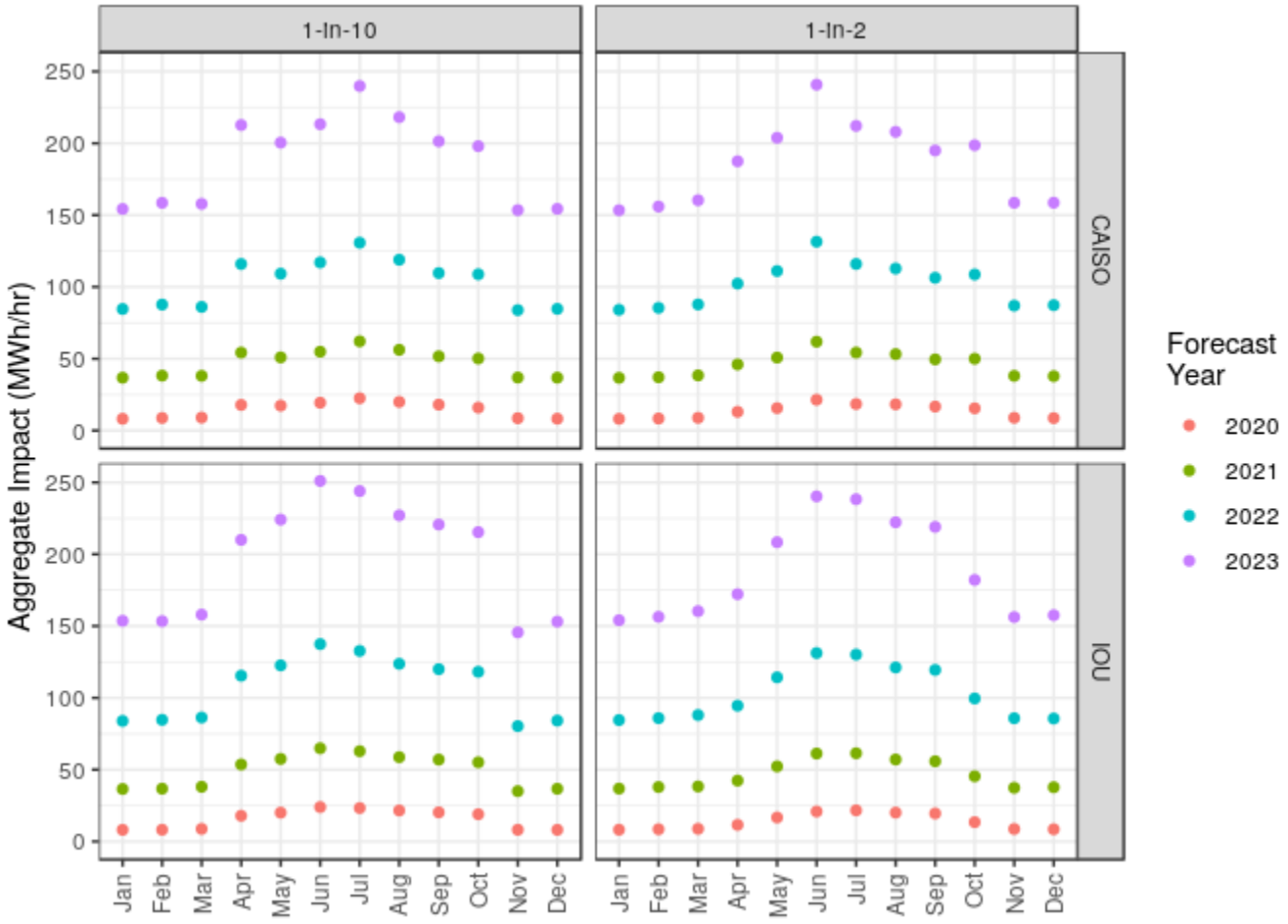


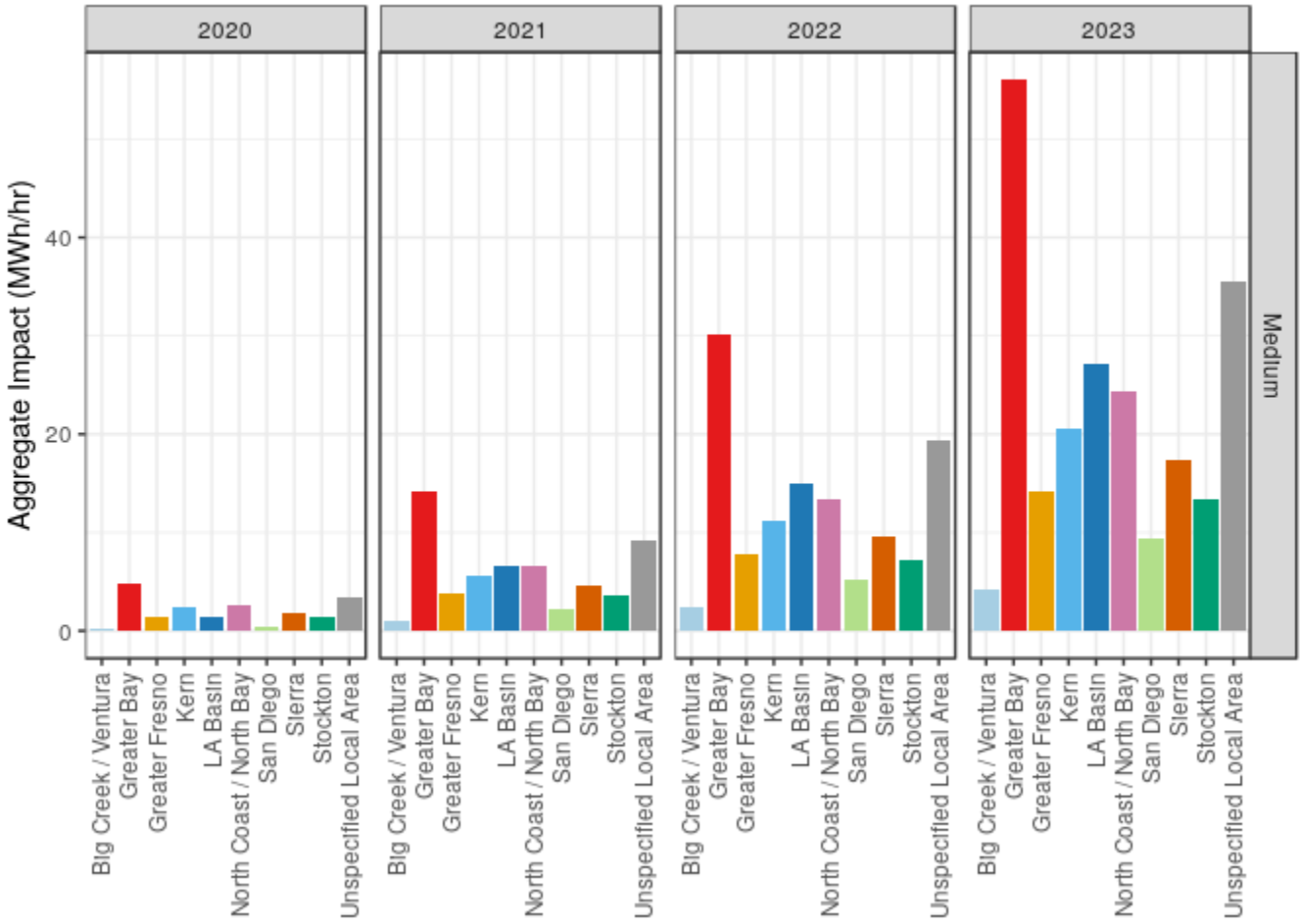
Table 15 shows the predicted load impact by year for the August 1-in-2 monthly CAISO peak day and the IOU 1-in-2 peak days. The large predicted impact increase over the next two years is due to the forecast enrollment increase.

Table 15. Residential predicted aggregate load impact for August CAISO 1-in-2 day and IOU 1-in-2 day

Year	CAISO 1-in-2 day		IOU 1-in-2 day	
	Temp (F)	Aggregate Impact (MW)	Temp (F)	Aggregate Impact (MW)
2020	86.8	18.3	88.8	20.1
2021	86.4	53.0	88.1	57.1
2022	86.3	112.8	88.0	121.2
2023	86.3	208.0	88.0	222.3

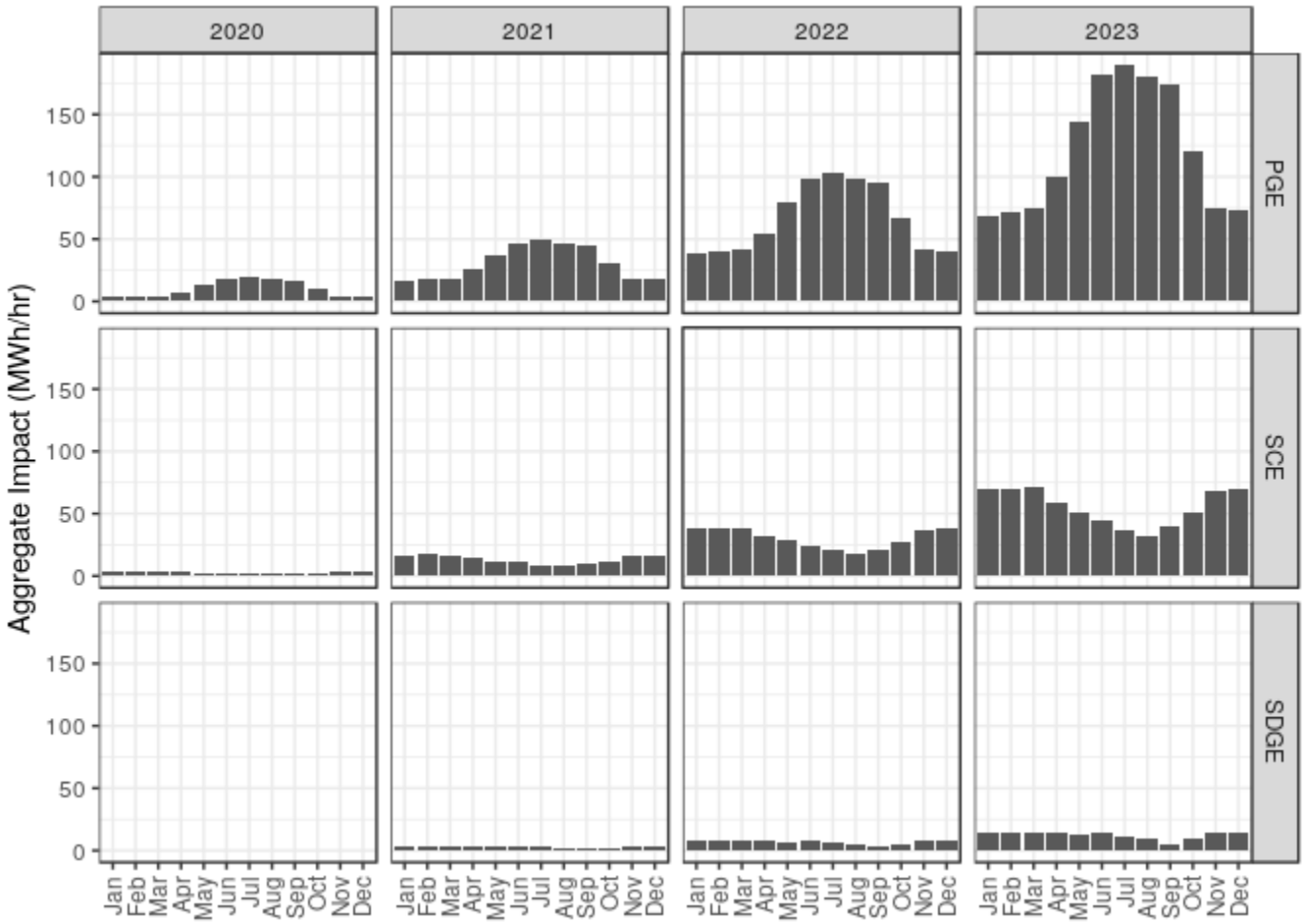
The ex ante predictions were made for each LCA level. Figure 37 shows the predicted aggregate impact for August, for the IOU 1-in-2 weather year by LCA. Greater Bay and the Unspecified Local Areas have the highest aggregate impacts across the years. While Leap forecasts impacts from the Unspecified Local Areas that are lower than LA Basin (for both sectors), Leap forecasts a large increase in impacts from residential air conditioning. Because Unspecified Local Area for residential air conditioning has a higher predicted impact than the LA Basin, more aggregate impacts occur in this area.

Figure 37. Residential predicted aggregate impact by Local Capacity Area, mean over RA hours, IOU 1-in-2 weather data, and for August of different years (medium forecast)



Aggregate impacts by month are shown in Figure 38, separately for PG&E, SCE, and SDG&E and by year.

Figure 38. Residential predicted aggregate impact by month, mean over RA hours, IOU 1-in-2 weather data, separately by utility (medium forecast)



SCE and SDG&E have lower impacts in the summer because summer impacts from residential batteries are lower than shoulder season due to batteries discharging to capacity from higher loads in the summer

Combined Nonresidential and Residential Forecast

Leap provides a combined resource across both nonresidential and residential sectors. The previous sections provided ex ante results by sector for clarity. This section combines the results and provides the total ex ante Leap resource.

Figure 39 and Figure 40 show the combined nonresidential and residential predicted impacts for the medium enrollment forecast. Conditional on the enrollment forecast, CDA projects that Leap resources could provide about 100 MWh/hr in August 2020 and increasing over the years to slightly over 600 MWh/hr in August 2023. As shown below, June or July have the highest monthly impacts (depending on the weather scenario).

Figure 39. Combined sector predicted impacts by month and year, mean over RA hours, (medium forecast) (nonresidential and residential resources combined)

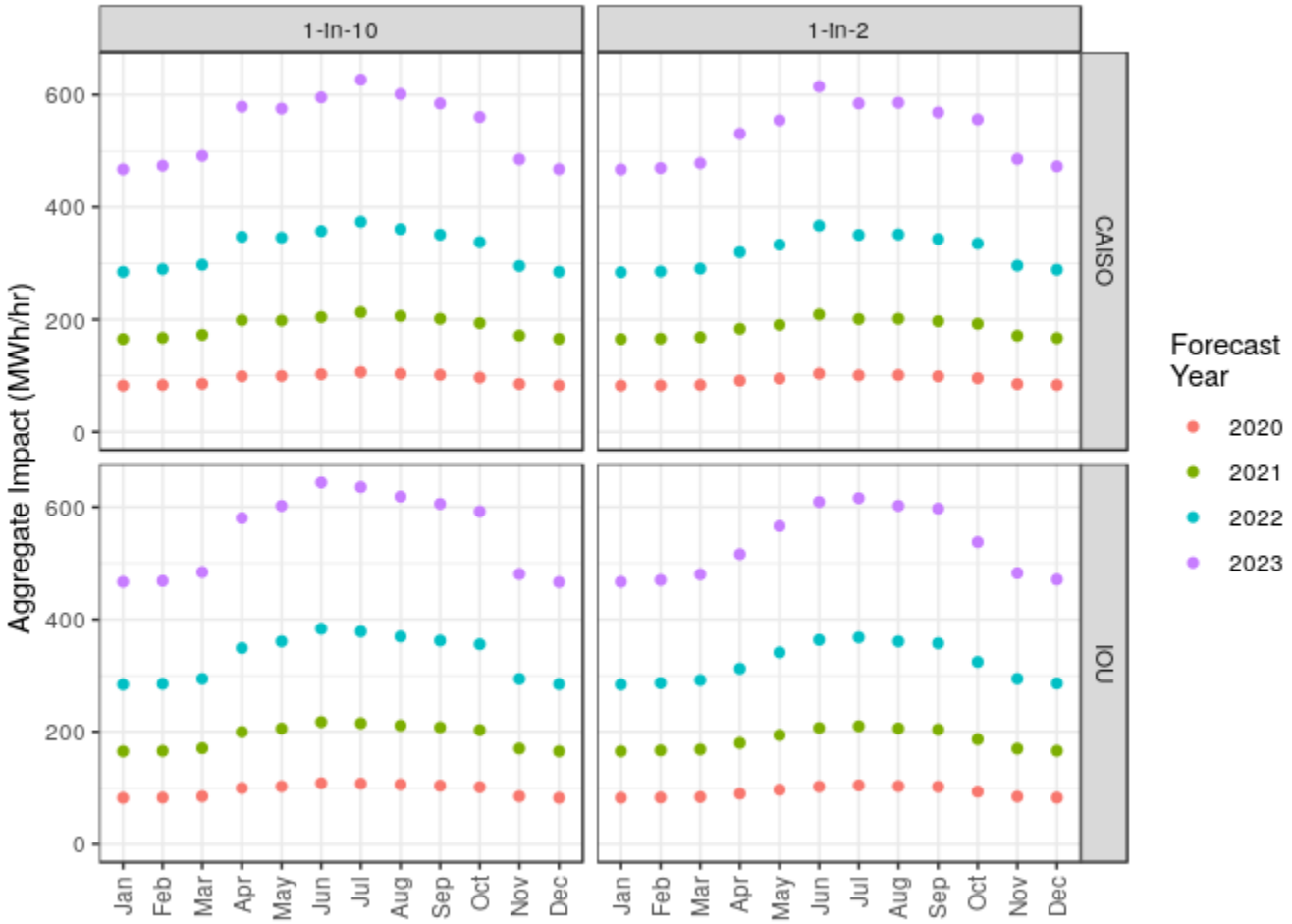
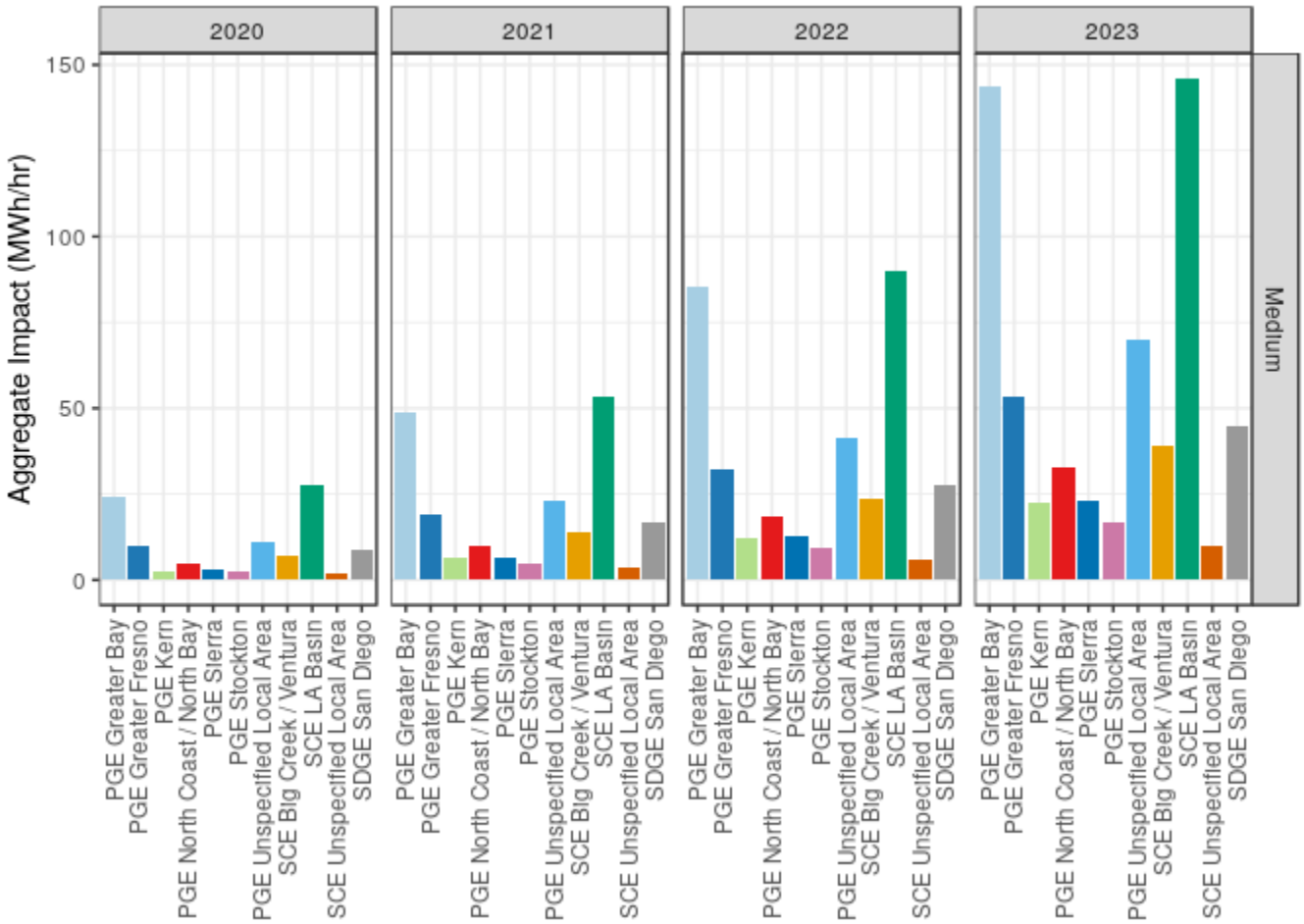


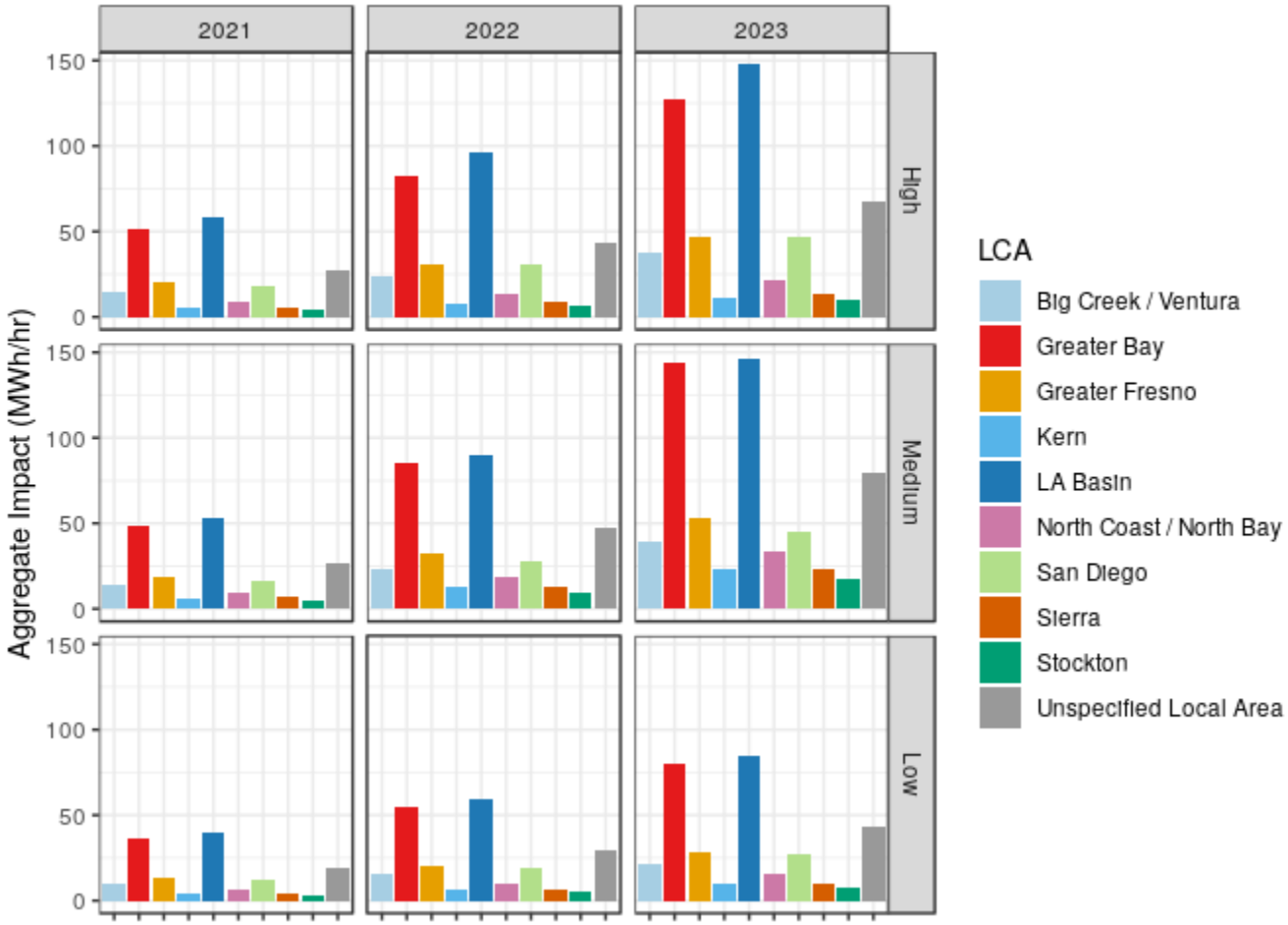
Figure 40 shows that Greater Bay and LA Basin have the highest predicted impacts across the years, with about two-thirds of their impacts from the nonresidential sector. The increased impacts for Kern and Stockton over the years are driven more by residential impacts than nonresidential impacts.

Figure 40. Combined sector predicted impacts by LCA and year IOU 1-in-2 weather data, mean over RA hours, August (medium forecast) (nonresidential and residential resources combined)



At the request of the CPUC, Leap provided CDA with low, medium, and high forecasts for 2020-2023 by aggregated impact for each load type. Figure 41 shows the combined sector predicted aggregate load impact for low, medium, and high forecast scenarios. The Greater Bay and LA Basin have the highest aggregate impacts in each forecast scenario.

Figure 41. Combined sector predicted aggregate load impact for low, medium, and high forecast scenarios
mean over RA hours, IOU 1-in-2 weather data, and for August of different years



Comparing Current and Prior Estimates

This report represents the first time Leap's DR resource has been evaluated using the DR Load Impact Protocols. For this reason, certain LIP required tables are not included. (Shown in the table below)

Table 16. Comparisons of Ex Post to Ex Ante Included in or Excluded from Report

Comparison	Included in Report	Excluded from Report
Current Ex Post to Prior Ex Post		✓
Current Ex Post to Current Ex Ante	✓	
Prior Ex Post to Prior Ex Ante		✓

Current Ex Ante to Prior Ex Ante		✓

Comparison of ex ante model predictions to ex post observations

In order to compare ex ante results to ex post results, we fit ex ante models to the ex post data from 2019 and used them to forecast the ex post load impact per meter for the ex ante weather conditions. The same models can be used with the actual 2019 conditions as inputs. Results are shown in Table 17.

Table 17. Comparison of 2019 ex post to 2019 ex ante

Load Type	Ex Post load impact per meter	Ex Ante load impact per meter
Airconditioning	14.2	14.4
Electric Vehicle	27.6	36.3
Large Commercial Battery	47.9	42.2
Other & Thermal Storage	76.0	70.5
Pumping	35.5	35.5
Small Commercial Battery	10.2	10.2
Residential Airconditioning	0.6	0.5
Residential Battery	0.3	0.5

We expect fairly close agreement between observations and predictions because we are comparing the model to the data on which it was trained. However, the Bayesian ex ante models are constrained so that the actual load impact of an event cannot be negative (although the estimated load impact can be): we do not think there is any realistic mechanism by which customers would increase their load in response to an event. This introduces some asymmetry to the predicted load impacts in some situations: although the estimated aggregate event impact is very low or even slightly negative for some load types in some LCAs, with error bars extending below zero, the ex ante models effectively redistribute the portion of the probability distribution that is below zero, moving it in to low positive numbers. On the other hand, the model also adjusts for the fact that it is possible to have a very high load impact estimate by combining high actual impact with positive random variation in the estimated load impact. High estimated load impacts in an LCA – compared to other LCAs – are pulled towards the overall mean, by an amount that depends on the strength of the evidence from test events that the load shed is actually very high. For both of these reasons, the ex ante load impact should not and do not perfectly match the ex post load impact.

Errors and Uncertainties

In this section we discuss the reasons the forecasts of the future (and current) load impact capacity are likely to differ from the actual capacity, and the ways we have tried to characterize the likely magnitude of the difference. We begin by discussing the most important sources of potential error, starting from the ex post results and working our way to the ex ante forecasts. We then introduce the key concepts of the Bayesian models that help these models quantify the uncertainty. The detailed models are presented in Appendix E.

"It's tough to make predictions, especially about the future" – Yogi Berra (attributed)

Any prediction can be wrong. In this report we predict what would happen if Leap were to call their entire resource to provide demand response load shed on a future date on which there are specific weather conditions. Here are some factors that affect the accuracy and precision of those predictions:

1. We do not know the exact ex post load shed. Yogi was right that predictions about the future are especially hard, but in this case even predicting the past is not easy; we do not know exactly how much load shed Leap's customers provided in their 2019 events. Load shed cannot be directly measured: the meter can only tell us how much energy the customer used, not how much they would have used in the absence of a DR event. To quantify the performance of their customers in 2019, we have to predict the amount of energy they would have used if there were no event, and subtract this from the amount they actually used. This prediction can be wrong. Indeed, it is essentially impossible that it is perfectly correct. As discussed in the Ex Post section, we use an empirical approach to quantify the statistical distribution of errors that are expected to occur. For any given event we don't know the exact value of the error, of course (if we did we would remove it) but the distribution of possible values of baseline error produces a distribution of possible values of load impact; the width of that distribution is what we call the 'uncertainty' in the load impact. For any individual event, the uncertainty in load impact can be a fairly large percentage of the impact, but the uncertainty in the average over many events is much lower.

2. There may be systematic change in customer performance. Even if we knew the exact amount of load shed provided by each customer for each event in 2019, we would not know exactly how much the same customers would provide in 2020. Changes in operational practices, economic conditions, and other factors can lead to an overall shift that would affect both existing customers and new recruits. Agricultural groundwater pumping can increase or decrease due to year-to-year differences in regional rainfall or crop type using the water. Electric vehicle charging will change depending on the market penetration of electric vehicles and on the behavior of EV owners (for instance, as battery capacities increase, the average EV trip distance may go up, leading to higher charging requirements and thus potentially increased DR load impact). Additionally, Leap indicates the possibility of upward performance potential as they work with partners over multiple years, leading to increased performance and an improved testing regime that more accurately characterizes resources' availability.

3. Spotty temporal coverage in the ex post data leads to inability to fully capture seasonal variability. Leap called DR events for only parts of 2019. Residential air conditioning events were called only in June and August, to give the most extreme example. All of Leap's resources were exercised in at least some summer months, so we do have data from the most important part of the year, but we have no way to estimate systematic seasonal variation for most of the load types, other than variation associated with temperature.

4. Spotty spatial coverage in the ex post data leads to inability to fully capture spatial variability. Leap is planning to expand their program into areas where they currently have few or no customers. There is spatial variability in load impacts, and we have insufficient data to quantify this, or to know whether the new areas into which they expand will provide more or less load shed per customer than the ones represented in the 2019 data.

5. There is a lot of uncertainty in Leap's future enrollment, whether in terms of enrolled customers, or load impact capacity, or any other metric. Leap has grown quickly and expects to

continue growing in the future, but nobody can be sure of an accurate prediction of how quickly their capacity will grow, either overall or in specific LCAs and load types.

We cannot eliminate the sources of uncertainty listed above. The spatial coverage is spotty because Leap does not yet have customers in certain areas of the state; the temporal coverage is spotty because they did not have customers in some months; systematic changes in customer behavior cannot be estimated for the past or predicted for the future based on data available to us.

For a given load type, we would not expect to be wildly wrong if we assume the load shed *per customer* in the future will be roughly what it was in 2019 in the same month of the year – or the load shed as a percent of aggregate baseline load -- even if many new customers are added in different locales. But we also would not expect the load shed per customer, or the load shed as a percent of baseline load, to be unchanged as new customers are added in different places. Part of the goal of our modeling is to try to quantify how different things can be, even if Leap is successful in the sense of adding enough customers that they have reasonable expectation of meeting their forecast of future capacity.

The modeling gives us a way to quantify the uncertainties that are due to effects 1-4. As discussed in the modeling appendix there are choices in the models that could lead to higher or lower uncertainties, but at least there is a well-defined way of estimating those uncertainties. But the largest source of uncertainty is how much Leap will grow.

Leap has provided high, medium, and low growth scenarios. Leap has forecasted how much load shed capacity they will have in each year. This contrasts the usual approach with which we are familiar, which is to forecast a number of customers and then model the load shed per customer, but we see no reason to believe it would be easier or more accurate to forecast the growth in the number of customers than the growth in load shed capacity. However, it does lay bare the extent to which the outcome of the ex ante forecasts is driven by the growth forecast.

As discussed earlier in the report, even though Leap has provided forecasts of load impact we have opted not to take those at face value. Instead we have chosen to interpret Leap's forecasts in terms of growth relative to their 2019 ex post impacts. This yields a growth curve in customer counts based on the assumption per-customer load impacts remain the same as in 2019, but there is potential for higher or lower impacts depending on the character of new customers. Due to small sample sizes, and in some cases small numbers of events, even the LCAs in which Leap currently has customers may not be well-characterized in terms of the load shed they can offer.

Recommendations

Based on our evaluation of the 2019 dispatch of Leap's DR resource, we provide the following recommendations to Leap:

- Call some longer-duration and full-resource events that can provide statistical support for full-resource and 4-hour+ RA window events that Qualifying Capacity numbers are based upon.
- Call events during more months of the year to gather information about seasonality and weather influences on event impacts.

Recommendations for future evaluators:

- Investigate baselining and comparison group methodologies for estimating event impacts that best characterize impacts for groups with few participants, varied events, and noisy baselines.

This could include Leap dispatching at least a subset of their future events with true randomized controls.

- Study two load types where Leap forecast high future impacts.
    - Closely monitor the EV impacts during any future test events to ensure Leap is obtaining an event response that is consistent across all event hours.
    - Evaluate future test events on residential batteries to determine if the simulated event impacts within the ex post analysis are comparable to actual events. Compare not only the average impacts, but the impacts over the test period.
- Consider how best to apply LIPs so that they align with the needs of third party and emerging program evaluation. Most notably, the "Typical Event" requirement of Protocol 8 is not appropriate to characterize the full resource when all participants are not dispatched for all events. Also consider how best to characterize a resource that is growing and/or changing rapidly.
- Investigate ways to characterize and, where possible, measure sources of uncertainty, such as between customer variation, variation in event participation, and variation in load and customer types.

Appendix A: About Leap

CDA notes: The content in this section is provided for context on the resource being evaluated. Leap wrote it all and CDA has not evaluated any claims made or altered it in any way.

Leap enables real-time automated trading on energy markets. Leap's marketplace for grid flexibility grants energy resources including battery energy storage, electric vehicles, HVAC systems, pumping loads, and more access to global demand response programs, wholesale markets, and real-time pricing through a single API.

Leap's open, hardware-agnostic platform turns the operators of energy resources of any size and type into responsive sources of grid flexibility, providing revenue to participants while unlocking the benefits of a truly resilient and transactive grid.

Leap's ability to aggregate multiple load types and sizes into larger blocks creates higher value grid resources and helps our partners to realize the full value of their automated resources in the wholesale market. Our business model helps Leap serve smaller loads cost-effectively and brings new Demand Response participants into the market. Leap is a privately held company with offices in San Francisco and the Netherlands.

History

Leap was founded in 2017, won capacity through the California Demand Response Auction Mechanism (DRAM) in 2018, and became an active Scheduling Coordinator and Demand Response Participant in the CAISO system in 2019. Leap has forged partnerships with a number of leading distributed energy solution (DER) solution providers, expanding in both customer base and geographic reach within and beyond California. Leap delivered Resource Adequacy to PG&E and SCE in 2019 and will be expanding to deliver RA to SDG&E as well in 2020. Leap has received recognition as a clean energy technology leader, including being selected to join Elemental Excelerator's 8th cohort in December 2019 to help California's agricultural customers to earn revenue in California's electricity markets. Leap was also recognized in January 2020 as Cleantech Group's Early Stage Company of the Year as part of the 2020 Global Cleantech 100.

Appendix B: Leap enrollment forecast rationale [removed from public version]

Appendix C: Leap discussion on controlling load by load type [removed from public version]

Appendix D: Known 2019 Event Failures and Solar and Storage Event Operation [Removed from public version]

Appendix E: Ex Ante Models

This appendix describes CDA's approach to the ex ante modeling as well as providing an example of modeling code for one load type (residential air conditioning).

The goal of the ex ante models is to allow prediction of the aggregate load impact, by LCA, if Leap's entire resource is dispatched in standardized weather conditions.

Leap provided CDA a forecast of their available capacity, by SubLAP, for each of the next ten years, which CDA rolled up to an LCA for modeling. Taken at face value, their forecast directly answers the question for ex ante. But of course, even if Leap knew how many meters they would have for each load type in each LCA in each future year, they would not know what impact capacity those meters would actually be able to provide. To quantify uncertainties in Leap's forecast, we use their forecast numbers (of capacity in each SubLAP, by load type) as the *expected* load impact ('expected' in the statistical sense), but use a model that takes into account many of the ways the true value could vary from the expected value.

Leap's 2019 events provide an estimated load shed per meter in each load type. An estimate of the load impact capacity they will have in the future can be obtained by multiplying the estimated load shed per meter from the 2019 data by the expected number of meters in the future. We applied this formula in reverse: given Leap's forecast of their load impact capacity in the future, we divide by the estimated load impact per meter in 2019 to get a forecast for the number of meters. We then ask: if this number of meters is obtained, what might the load impact actually be?

We used several different models, with the choice depending on whether the load impact is expected to be temperature-dependent: Residential and nonresidential air conditioning, and residential batteries, were assumed to have temperature-dependent loads. All other loads were assumed non-temperature-dependent.

Residential battery systems deserve some discussion to explain their observed (and expected) load impact characteristics: The temperature-dependence of the residential battery model is intended to help capture the seasonality of insolation and to serve as a proxy for cloudiness (temperatures higher on sunny days). The batteries are usually charged fully before the 4 pm start of the RA window, but on hot days, their capacity can be exhausted before the end of the event due to high AC loads. When the battery is fully discharged, it cannot reduce load. Therefore, the DR performance of a group of residential batteries tends to be worse on sunny summer days than on cloudy days.

Whether temperature-dependent or not, all of the models share the same basic structure:

1. The estimated load shed for an ex post event in a given LCA is assumed to be a random sample from a distribution centered on a 'true load shed' for that event in that LCA, with the width of the distribution determined by the uncertainty of the estimate.
2. The true impact for the event is assumed to be a random sample from a distribution centered on an overall mean. This mean is either a constant (but unknown) value associated with the LCA, or, in the temperature-dependent models, a constant LCA -specific value plus the product of an LCA-specific regression coefficient times the number of degrees by which the temperature exceeds 70 F

A model in which the observed values are assumed to be drawn from a distribution around a true value (such as the LCA mean), and the true values are assumed to be drawn from a distribution about yet

another value (such as the mean of all LCAs), is called a 'Bayesian hierarchical model' or a 'multi-level model.' Describing such models in detail is far beyond the scope of this report. An excellent freely available resource is Bayesian Data Analysis, by Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin, which is available at http://www.stat.columbia.edu/~gelman/book/BDA3.pdf Our non-temperature-dependent models are very similar to the classic '8 schools' model of Rubin, as described starting on p. 119 of that book; the temperature-dependent part is a logical extension.

We strongly recommend reading the portion of the book described above if you are not familiar with Bayesian models, but we will say a few words about them here in the context of sources of uncertainty and error that were discussed earlier. As previously noted, the ex post load impact for an event is not known exactly: it is an estimate that is subject to baseline error. There is some 'true' event impact – the value that would have been obtained if the measurement had no error, but we do not know what it is. Additionally, when only a few DR events are conducted there is always the possibility, indeed the near certainty, that the average event impact will be higher or lower than what would have been seen if many more events had been conducted. There is some 'true' event mean, but we do not know what it is. And finally, the ex ante predictions for this program require predictions for some load types to be made in LCAs where Leap does not yet have any data; the mean impact in those LCAs will differ from the LCAs for which there is data --- there is surely some variation from LCA to LCA – but we don't know whether those new LCAs will be higher or lower, or by how much..

A Bayesian model cannot tell us these unknown values – you cannot squeeze blood from a stone -- but it can help constrain them. For instance, all of the LCAs for which we *do* have data have similar temperature coefficients – similar to each other – then it is likely that the new LCAs will not fall far from those values. But if all of the LCAs for which we have data appear to be very different from each other in terms of load impact, then the new LCAs could fall anywhere within a wide spread. A Bayesian hierarchical model provides a statistically valid way of quantifying the uncertainties in all of the numbers and distributions, and thus obtaining reasonable uncertainty estimates.

We use 'informative prior distributions' to help constrain some of the values for which there is not enough data to estimate them precisely. For example, in the Residential Air Conditioning model (given below in full) we assume that the amount by which the mean per-meter event impact can vary from event to event, for an event at a given outdoor air temperature, is probably less than 0.4 kW. That is, if we knew precisely the actual load impact (as opposed to the estimated load impact) for many events that took place at 80 F, we think most of them would be within 0.4 kW of the overall mean of those events. We provide this 'prior information' by assuming the sigma parameter in the model is drawn from a half-normal distribution with standard deviation 0.2; see the model below, which, like all of the models, is implemented in the Stan language.

Many people new to Bayesian statistics balk at the idea of 'informative prior distributions', which basically serve to put a thumb on the scale. Ironically, some people who would have no problem with simply applying a program-wide estimate of, say, the mean load shed for electric vehicle battery chargers, balk at the idea of assuming the value varies from LCA to LCA but constraining the amount of variation. But assuming there is a single program-wide value is equivalent to allowing variation from LCA to LCA but setting the amount of variation to exactly zero!

Models were fit separately for: the two hours prior to the event; the event itself; the four hours after the event. This was done to capture pre-cooling and rebound if present. These phenomena were observed for the air conditioning loads (residential and nonresidential) but not for the other load types.

Code for the Residential Air Conditioning Model

```
data {
  int <lower = 0> N_data_rows;
  int <lower = 0> N_pred_rows;
  int <lower = 0> N_groups;
  int <lower = 0> N_months;
  int             month[N_data_rows];
  int             group[N_data_rows];
  int             month_pred[N_pred_rows];
  int             group_pred[N_pred_rows];
  vector [N_data_rows] impact_per_participant;
  vector [N_data_rows] impact_err;
  vector [N_data_rows] frac_impact;
  vector [N_data_rows] frac_impact_uncert;
  vector [N_data_rows] temp70;
  vector [N_pred_rows] temp70_pred;
  vector [N_data_rows] temp;
  vector [N_pred_rows] temp_pred;
}

parameters {
  real   <lower = -1, upper = 1>  theta_group;  // mean load shed
                                                // (below 70 F) of
                                                // dist'n from which
                                                //group means are drawn.
  real   <lower = 0, upper = 1>   tau;      // dispersion parameter
                                            // for dist'n of group
                                            //  means around overall
                                            // group mean.
  real   <lower = 0, upper = 1>   sigma;    // dispersion parameter
                                            // for true impact around
                                            // group mean.
  real   <lower = 0, upper = 1>   sigma_temp; // how much do different
                                              // groups vary in
```

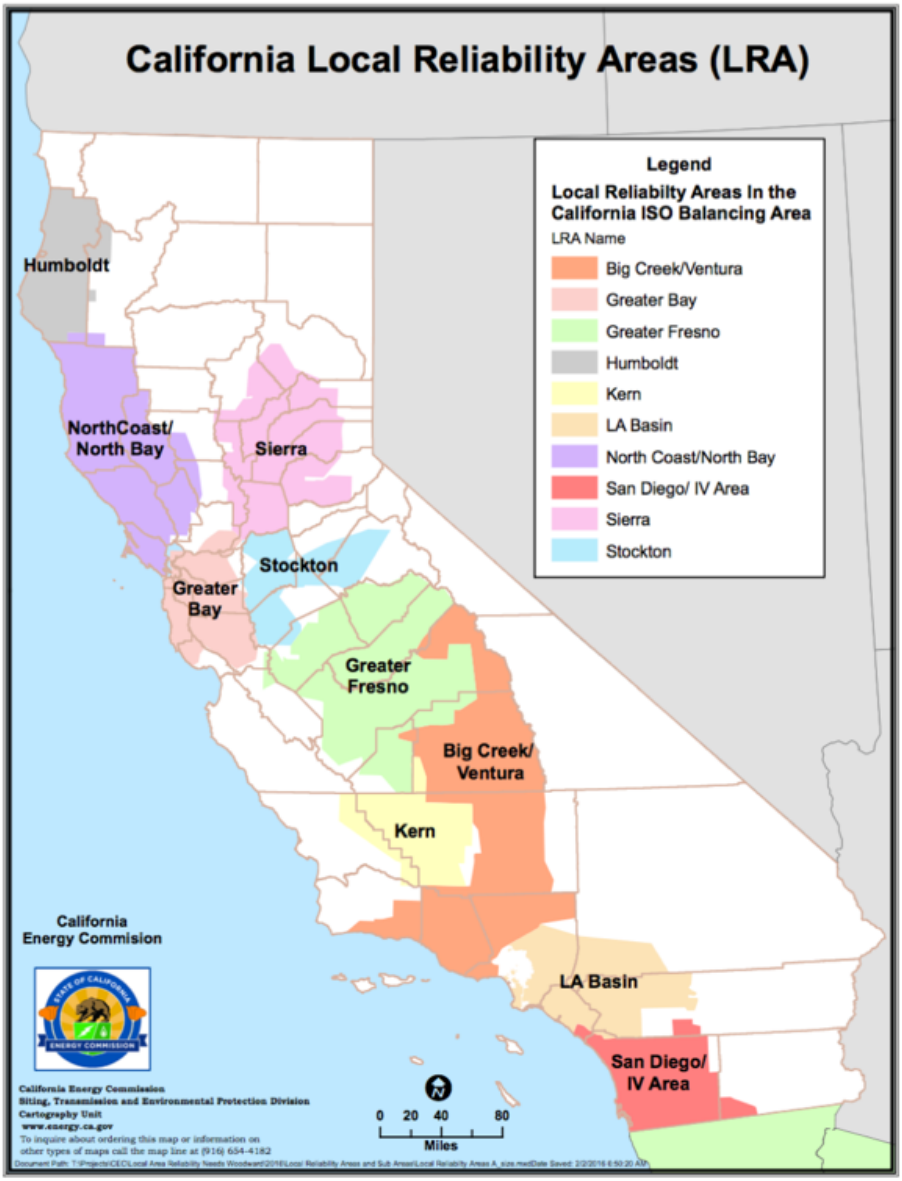
```
                                                    // temperature response?
  real   <lower = 0>                beta_temp_mean;
  vector <lower = 0> [N_data_rows] true_impact;
  vector <lower = 0, upper = 1> [ N_groups] mean_shed;
  vector <lower = 0> [N_groups]    beta_temp;
}

model {
  sigma ~ normal(0,0.2); // How much can true impact differ from
                         // predicted impact, in either direction,
                         // if we knew the prediction parameters
                         // perfectly? (This is variation due to
                         // parameters not in the model).
  tau   ~ normal(0,0.03);  // How variable are intercepts?
  theta_group ~ normal(0, 0.02); // We know that below 70 F there
                            // can't be much load impact, if any.
  beta_temp_mean ~ normal(0.0, 0.03);  // We expect something like 0.5
                      // kW at 80 or 85 F, from other programs,
                      // so have a prior that is consistent with
                      // that but has lots of probability below that.
                                            // Pool towards 0 to be
deliberately conservative.  0.03 would mean 0.3 kW per 10F.
  sigma_temp  ~ normal(0, 0.02);     // Some variation between LCAs.

  mean_shed ~ normal(theta_group, tau ); // distribution of individual
                                         // groups around overall mean
  beta_temp      ~ normal(beta_temp_mean, sigma_temp);

  for (row in 1:N_data_rows) {
    if (temp[row] > 50)
      true_impact[row] ~ normal( mean_shed[group[row]] +
                       beta_temp[group[row]] * temp70[row], sigma);
    else
      true_impact[row] ~normal( 0, 0.0001);
```

```
    impact_per_participant[row] ~ student_t(6, true_impact[row],
                                                    impact_err[row] );
  }
}

generated quantities {
  vector [N_pred_rows] pred_impact;

  for (row in 1:N_pred_rows) {
    if (temp_pred[row] > 50)
      pred_impact[row] = normal_rng( mean_shed[group_pred[row]] +
               beta_temp[group_pred[row]] * temp70_pred[row], sigma);
    else
      pred_impact[row] = 0.0;
  }
}
```

Appendix F: LCA geography

This image illustrates the geography of California's LCAs (aka Local Reliability Areas on the map).

Figure 43. Map of California's LCAs (aka LRAs)



Original at: https://ww2.energy.ca.gov/maps/reliability/Local_Reliability_Areas.pdf

Appendix G: Data Cleaning for Analysis

An important part of every analysis is preparing data for modeling by checking that the data makes sense, is consistent with what is needed and removing unnecessary or problematic data that can cause issues. For this evaluation, we went through a series of steps on the combined residential and nonresidential data that resulted in removing some data points from the analysis data set. We describe those steps here and include Table 18 which shows how many rows of meter data and how many meters remain after each step.

Table 18: Table of drops for analysis data

	Nonresidential		Residential	
Drop Reason	*Meter Count*	*Row Count*	*Meter Count*	*Row Count*
Original Data	771	6,510,744	6,881	53,889,068
Holidays and PSPS days	771	6,030,907	6,878	49,756,831
Weekends	771	4,282,540	6,876	35,315,080
Incomplete days with less than 24 hours of data	770	4,258,056	6,872	35,166,288
Customers with <20 days of data	766	4,257,024	6,796	35,151,000
Residential customers with no events			6,689	34,671,192

We dropped data for a variety of reasons, mostly because the meters data fell on days incompatible with demand response. These are full explanations for each drop reason:

1. Drop holidays and PSPS days – We removed NERC holidays from the data, and dropped PSPS days, since on both of these day types, customers use energy much differently than on regular weekdays
2. Drop weekends – We removed weekends from the data because customers use energy differently on the weekends than weekdays
3. We dropped individual days from meters where one or more hour of data is missing as this causes problems with many algorithms
4. We dropped customers with fewer than 20 days of data. We consider 20 days an absolute minimum to develop baselines and model customers. These 76 residential and 4 non-residential customers enrolled after all events
5. We dropped residential customers who were enrolled by Leap after all events.

The hours of meter data and customers that we removed from the analysis data were all because that information was not useful to the analysis, rather than dropping customers who participated in events. The final analysis data contains very complete information for participating customers on non-holiday, non-PSPS weekdays.

Appendix H: Data Cleaning and Analysis for Residential Batteries [removed from public version]

Appendix I: Response to PAO's draft report review

In this appendix we reproduce the CPUC Public Advocate Office's (PAO) comments on the DRAFT load impact evaluation, submitted on 5/15/2020, and provide detailed and point by point responses. The PAO was critical of several aspects of the evaluation, but we believe several of the comments are based on misinterpretations of the context of the cited numbers or figures. Since that confusion can likely be traced to the manner in which the results were presented in the DRAFT deliverables, we hope to clarify our understanding of Leap's resource in this appendix.

CDA commentary will be in Time New Roman italics.

First, some necessary background:

(1) This is Leap's first year calling events in support of LIP evaluation and their first evaluation under the LIPs. All of the events called were test events and they structured their tests to incrementally call all of their enrolled customers over the course of the summer of 2019. As a result, ***the aggregate performance of any given test event reflects only a small fraction of their enrolled customer base****.*

(2) Leap's DR resource is composed of a mix of ***very heterogeneous load types****, including agricultural pumping, commercial EV chargers, commercial battery systems, commercial AC, residential AC, residential battery systems, etc. Although there is nothing "wrong" with doing so, no IOU program in existence aggregates across load types in this manner. The heterogeneity of the resource required careful thought around the methods of impact evaluation, but also* ***presented particular challenges around questions of enrollment and participant counts across load types****.*

(3) The LIPs are clear and prescriptive on the meaning of "Typical Event" in the context of the report and table generators. Protocol 8 states: **"An average event day is calculated as a day-weighted average of all event days."** *And then clarifies:* **"it is the sum of the impacts in each hour for each event day divided by the number of event days. The reason to think of this as a day-weighted average is because the weights to use when calculating the standard errors are squared."** *This definition does not allow for the averages to be weighted by participant count, which is the correct way to compute the expected value of events with varying participation numbers and they certainly don't represent the performance of the full set of known participants, although that is often how "Typical Event" numbers are interpreted.*

(4) Taking all three of the above into account – the incremental nature of the test events, the heterogeneity of participants, and the requirement that we perform straight averages to obtain Typical Event numbers, ***the Typical Event numbers are largely meaningless to any question of Leap's aggregate potential****. However, most IOU program "Typical Event" calculations are reasonable estimates of full resource potential, so there is bound to be confusion around the meaning and reasoning behind Leap's Typical Event. For example, one large pump could be used to achieve similar impact to 1000 residential AC customers, so it does not mean much to say that the average enrollment between a hypothetical event dispatching one pump and another dispatching 1000 residential thermostats was 500.5. Similarly, the unweighted combination of the uncertainties across events is very large due to the large variation in customer size, load types, and which customers were called. The uncertainty for Typical Events primarily represents variation across events rather than uncertainty in the resource capacity.*

PAO's comments on the DRAFT evaluation report and underlying data:

1. Leap's 2019 ex post estimates show no event impact. CDA reports ex post nonresidential load impacts for 26 Leap Demand Response (DR) events in 4 months of 2019. The average estimated impacts in April, May and June include zero in the 80% confidence interval (DRAFT figure 14) indicating that Leap's DR events did not have a verifiable impact for those months. The remaining month, August, does not include zero in the confidence interval, but the lower bound appears to be almost zero.[2] Therefore, given that Leap uses the highly imprecise confidence measure level of 80%, it is likely the August interval would also include zero if CDA reported at more generally accepted level such as 90% or 95%.

We have updated to reporting 90% intervals, but we note that these comments conflate confidence interval extent and expected impacts – the mean is the expected impact. The confidence range is a measure of certainty and can be used to make statements like "if the same event were repeated over and over under the same conditions, the resulting confidence intervals would be expected to include the true impact 90% of the time", but the errors on events involving small groups of customers crossing zero cannot be used to accept the "null" hypothesis that "the resource delivers no value". The mean IS the expected impact, with much of the uncertainty due to noise in the baselines, especially across relatively few participants. Further, and most importantly, large uncertainties from the specific sub-groups of customers who participated in events each month do not necessarily result in large uncertainties in performance estimates of the full resource if it were to be called all at once.

If you want to draw conclusions about the full resource, you have to look at estimates of performance based on the full resource. As evaluators, we tend to think of ex post as our opportunity to study the various influences on average impacts per-participant. Those are the basis for the modeling (i.e. controlling for outside temperature, load type, time of day, etc.) that takes place in ex ante. However, we've realized that we failed to provide a concise representation of the "Full resource" ex post as opposed to "Typical event" ex post aggregate performance and uncertainties in the report and table generators. We have corrected this deficiency, so assessments of the resource as observed can be made more easily moving forward.

Furthermore, the uncertainty seen in the ex post impact estimates tables indicates that Leap may not have delivered any value to ratepayers in 2019. Table 1 shows average per meter results for a typical Leap event from 6pm - 7pm in PG&E territory.

Table 1: Average Per Meter Results for a Typical Leap Event in 2019, PG&E Territory[3]

Hour Ending	Estimated Reference Load (kWh/hour)	Observed Event Day Load (kWh/hour)	Estimated Load Impact (kWh/hour)	Uncertainty Adjusted Impact (kWh/hr)- Percentiles				
				10th %ile	30th %ile	50th %ile	70th %ile	90th %ile
18	28.1	26.4	1.743	-213.66	-86.40	1.74	89.89	217.15
19	52.7	38.7	13.993	-241.79	-90.67	13.99	118.66	269.78
By Period:	**Estimated Reference Energy Use (kWh)**	**Observed Event Day Energy Use (kWh)**	**Estimated Change in Energy Use (kWh)**	**Uncertainty Adjusted Impact (kWh/hour) - Percentiles**				
				10th %ile	**30th %ile**	**50th %ile**	**70th %ile**	**90th %ile**
Daily	1,077	1,085	-8.41	n/a	n/a	n/a	n/a	n/a
Event Hours	52.7	38.7	13.99	-314.56	-120.45	13.99	148.44	342.55

The large differences in the uncertainty adjusted impacts,[4] which range from load reductions many times greater than the average customer's peak demand to load increases of a similar magnitude, show how meaningless the estimated impacts are. For example, these figures show a 10% probability that the impact during these 2019 event hours was either less than -312.56 kWh/hr or that it exceeded 342.55kWh/hr. Given that the estimated reference load during event hours was 52.7 kWh, these wildly swinging ex post results do not demonstrate that Leap provided ratepayers any actual load reduction benefits.

Once again, we point out that our hands were tied by Protocol 8 when calculating "Typical Event" performance. Typical Event numbers are the simple average across called events, but because Leap called their resource incrementally across many test events, that average has little to no bearing on the question of their full resource potential. We have now added this missing information by calculating "Full Resource Estimates" of event performance based on per-participant event performance across all events multiplied by enrollment, being careful to compute all per-participant and aggregate values by load type before rolling them up to full totals since participant counts lose their meaning when averaged across load types.

2. Leap's ex ante estimates should be discounted considering the uncertainty present in their ex post estimates and the lack of transparency into Leap's ex ante estimate methodology. The results of CDA's ex post analysis are used to inform their ex ante impact estimates based on the expected per meter load shed of a particular customer type. Given the concerns raised in the previous section, the estimated impact per meter is entirely unreliable. The data as presented does not verify that Leap provided any significant impacts or even had the capability to provide energy when a DR event is called. Therefore, there is no empirical evidence on which to base Leap's forecasted load impact.

Resource wide per-meter estimated impacts are not "entirely unreliable" for reasons already stated above – most of the figures on ex post are based on incremental events that called just a small

fraction of the full resource. New report tables provide standard errors and confidence intervals on the ex post numbers aggregated using 2019 enrollment and the confidence intervals are clearly tighter than those from individual events and are not close to including zero. This updated look at the full resource should constitute more clear evidence for Leap's resource value.

Moreover, Leap did not provide CDA customer enrollment forecasts.[6] Instead, Leap provided megawatt capacity forecasts, essentially skirting the LIP process by performing its ex ante estimates outside of the regulated process.

Due to the heterogeneity of their resource, Leap plans and forecasts in terms of capacity. After consideration, CDA saw no value in making the forecast look like enrollment by having Leap do the conversion themselves (see the paragraph below). Instead, CDA elected to accept the forecast in terms of capacity.

We see no place in the LIPs where the forecast is required to be made in terms of enrollment. They do make clear that the forecast is not to come from the evaluators and that the evaluator should adopt a consultative stance on topics related to what the forecast contains: **"For example, forecasting the size or makeup of the participant population at some future point in time is not part of impact estimation. Rather, impact estimation concerns estimating demand response given assumptions about the size and makeup of the participant population that are provided to the evaluator by someone else (e.g., regulators, planners or some other stakeholder). Having said that, the evaluator has an important role in guiding the development of data needed to make such estimates, in that he or she must tell the interested user what information is needed."** *In the case of Leap's plans for the future, the most salient issue is that due to the heterogeneity of Leap's resource (i.e. coming from a wide variety of load types with widely varying participant "sizes" and expected impacts) and the fact that their "product" is load impact capacity, they conduct all of their internal planning and modeling in terms of capacity as well. To convert their capacity recruitment goals into enrollment, they would need to know their per-customer impact, so they face a chicken and egg problem. It seems to us that forecasting in capacity relative to 2019 is a reasonable solution to that problem and any conversion they would make to enrollment would be a superficial change.*

Our updated Enrollment Forecast section confirms that CDA normalized the capacity forecast at 2019 = 1 and tracked subsequent years as multiples of the initial year. This was done separately for each load type/sub-LAP. Then we applied those scaling factors to 2019 actual enrollment to obtain calibrated enrollment estimates. Finally, those calibrated enrollment numbers were used to multiply the ex ante per-customer predictions. In this manner, we applied the relative growth implied by Leap's forecast without deterministically producing their capacity goals. However, because ex ante capacity was largely dictated by the forecast, we advise the relevant authorities to review the enrollment numbers behind the ex ante estimates. If the enrollment seems plausible, then the capacity should be judged plausible and vice versa.

Leap is therefore proposing to sell California ratepayers a product into which customers have no insight. The ex-ante load impact estimates rely on confidential internal proprietary growth models. The Commission should give no weight to Leap's ex ante forecasts unless more transparent and evidence-based estimates are provided.

As their evaluator, we have seen and grown familiar with several aspects of the set of information Leap considers confidential and believe their stated purpose wanting to avoid public disclosure of sensitive information to their competitors is legitimate. For a private company in a competitive field, it is common for growth plans in particular to be sensitive. Further, there is nothing unusual about the exclusions - it is routine for sensitive information to be redacted from public evaluation reports. The reason that practice has been deemed acceptable is that private versions of all reports and data are furnished to the CPUC (as Leap has done in this instance), which has a duty to verify that the public interest is being served even when considering the redacted portions of evaluations. As a part of the PUC, PAO is authorized to access such confidential information for legitimate review purposes, so it does not appear to us that the PAO has a structural disadvantage in learning what Leap's forecasts are based on. Given that the CPUC has access to the confidential forecasts that they can judge for themselves, there is no basis in the protocols or precedent from prior evaluation work that they should give them no weight because they are classified as confidential.