EVALUATION PLAN

PY2024 Emergency Load Reduction Program Evaluation

January 17, 2025

Prepared for:
San Diego Gas & Electric

By:
Savannah Horner
Demand Side Analytics

TABLE OF CONTENTS

1 INTRODUCTION

This evaluation plan lays out the analysis approach and requirements for evaluating impacts for San Diego Gas & Electric (SDG&E)'s Emergency Load Reduction Program (ELRP), which includes the goals of accurately estimating hourly ex-post load impacts for ELRP rates in 2024 and producing ex-ante load impact forecasts through 2035.

Given these goals, there are two main objectives for this evaluation plan. The primary objective is to engage in science and avoid after-the-fact analysis and decisions where there is a temptation to modify models to find the desired results. This requires documenting the hypothesis, specifying the intervention, establishing the sample size and the ability to detect a meaningful effect, identifying the data that will be collected and analyzed, identifying the outcomes that will be analyzed and segments of interest, and documenting in advance the statistical techniques and models that will be used to estimate energy savings and demand reductions. The goal is to leave little to no ambiguity regarding what data will be collected or how the data will be analyzed. The secondary objective is to comply with the California Load Impact Evaluation Planning Protocols[1].

This evaluation plan is laid out such that reporting requirements, methods, and considerations that apply across the ELRP subgroups are summarized first. There are also details specific to evaluating each subgroup of the program, which stem from a variety of factors including characteristics of each program subgroup and the availability and quality of treatment and control populations for analysis. The considerations for each program are addressed in their own sections, along with detailed plans for evaluation.

1.1 CALIFORNIA LOAD IMPACT PROTOCOLS

The California Load Impact Protocols require that for every demand response evaluation, an evaluation plan be produced that establishes a budget and schedule for the process and develops a preliminary approach to meeting the minimum evaluation and reporting requirements. The evaluation plan should also develop a plan to determine what additional requirements, if any, will be met in order to address the incremental needs that may arise for long term resource planning or in using load impacts for other applications, such as customer settlement or CAISO operations. At a high level, the requirements for a load impact evaluation are to provide:

- Impact estimates for each of the 24 hours on various event day types for event-based resource options and other day types for non-event based resources;
- Estimates of the change in overall energy use in a season and/or year;

[1] The full set of load impact protocols can be found here: http://www.calmac.org/events/FinalDecision_AttachementA.pdf

- Uncertainty adjusted impacts, reported for the 5th, 50th, and 95th percentiles, reflecting the uncertainty associated with the precision of the model parameters and potentially reflecting uncertainty in key drivers of demand response, such as weather;
- Outputs that utilize a common format for ex post evaluation. A slightly different reporting format is required for ex ante estimation;
- Ex ante estimates for each day type;
- Various statistical measures so that reviewers can assess the accuracy, precision and other relevant characteristics of the impact estimates;
- Ex ante estimates that utilize all relevant information from ex post evaluations whenever possible, even if it means relying on studies from other utilities or jurisdictions;
- Detailed reports that document the evaluation objectives, impact estimates, methodology, and recommendations for future evaluations.

1.2 SUMMARY OF STUDY DESIGN AND EVALUATION CRITERIA

Table 1 lists the study design question in the California Load Impact Protocols and details how the evaluation plan addresses each study design issue for each program. More detail can be found in each of the subsequent program-specific sections.

Table 1: Study Design Questionnaire

#	Study design issue	A.1, A.2, A.5, B.2	A.4, A.6
1	Will the evaluation rely on a control group? If so, how will it be developed and what comparisons between the treatment and control group will be made?	Yes: either through a site-specific regression model in which matched control hourly usage is used as a right-hand-side variable or a difference-in-differences model in which a matched control is selected with replacement on a stratified random sample of nonparticipants.	Yes: a matched control group with replacement on a stratified random sample of nonparticipants. Matching will be conducted using a methods tournament that will include multiple propensity score matching Euclidian distance matching specifications. The best performing method will be identified using out of sample bias and fit statistics.
2	Will the evaluation rely on pre-intervention data to establish a baseline?	Intervention only occurs on subset of days. Comparison to non-event days will be investigated, but not used as baseline.	
3	Will the study rely on a sample or include the full population receiving the intervention? If a sample is used, does it meet 90/10 precision requirements?	Full population	
4	Is the study designed to detect a specific effect size? And, if so, how was statistical power assessed?	N/A – full population analyzed for all relevant subgroups	
5	What is the study's threshold for statistical significance?	95% confidence using a two-tailed test	
6	What is the size of the control and treatment groups, if applicable?	A.1: ~700 All other groups: generally < 100 Synthetic controls will be selected from NMEC granular profiles and matched control will be selected from non-participants.	A.4: ~600 Matched controls will be identified for A.4 (no A.6 events were called in PY 2024).

#	Study design issue	A.1, A.2, A.5, B.2	A.4, A.6
7	How will the evaluation address outliers?	Individual customer regressions and customer specific models will be used to ensure best fit for large customers.	Customers for whom a matched control group cannot be identified (due to score distance) will not be included. We expect it to be less than 1% of participants.
8	How will the evaluation address attrition?	Ex post impacts are estimated for all dispatched customers. Ex ante will incorporate any information about changes/improvements to dispatch.	
9	How will standard errors be calculated?	Robust standard errors for individual customer regressions and standard errors produced by difference-in-differences.	Standard errors produced by difference-in-differences.
10	Will estimates be developed for subcategories? If so, please define them.	Yes, segmentation is reported in the sections below.	
11	Will energy savings be estimated?	No	
12	Will overlap with energy efficiency programs be estimated?	No	

2 GENERAL APPROACH AND METHODS

In general, the methods to be used in the ex post and ex ante impact estimations for ELRP will be familiar to evaluators, program managers and regulators. In this section, we review the key considerations for unbiased impact estimation and the evaluation methods used to produce results. Details for each specific program will be summarized in the subsequent sections.

The primary goal of any load impact evaluation is to answer two key questions: what were the historic ex post load impacts in the prior evaluation period, and what are the estimates of program load impacts going forward? This second question is of particular importance, as it can be leveraged for long term resource planning, DR impacts for resource adequacy, and other progress reporting. In this document, we focus instead on developing a plan that ensures that unbiased ex post estimates are produced and fed into a robust ex ante estimation process in a way that is transparent and logical. To that end, the evaluation plan lays out key issues to be addressed in the process of developing ex post and ex ante impacts, summarized in Table 2.

Table 2: General Considerations for ELRP Load Impact Evaluations

Evaluation Consideration	Framework
What is the target level of confidence and precision in the impact estimates?	We intend to target 95% confidence using a two-tailed test.
Will both ex post and ex ante impacts be produced?	Yes, with the exception of ex post impacts for subgroups for which no events were called during the evaluation period (e.g. A.6).
What, if any, changes are expected over the forecast horizon to either the program or participant characteristics that should be incorporated into ex ante estimates?	SDG&E program staff will provide a summary of expected program changes, which will be incorporated into the analysis. DSA is responsible for developing ex ante enrollment forecasts, based on the assumptions discussed by DSA and SDG&E program staff.
Will impact persistence be explicitly incorporated into the analysis?	Program impacts can be compared to impacts from previous years, but ELRP is too new of a program for a formal analysis of impact persistence.
Is M&V activity needed to address the issue of persistence or of program changes?	As impact evaluations are conducted annually, no additional M&V activities are expected to be leveraged to monitor persistence.
Will impacts be developed for geographic sub-regions? If so, what are these sub-regions?	Yes, impacts will be reported by LCA, SubLAP, and climate zone.

Evaluation Consideration	Framework
Will impacts be developed for sub-hourly intervals?	Impacts will be developed at the granularity of interval data available. We expect this to be hourly for most subgroups.
Will impacts be developed for participant sub-segments? If so, what are these sub-segments?	Yes, results will be segmented by subgroup-appropriate customer types, as well as size and industry for commercial customers. Subgroups with aggregators or service providers will also be segmented by the various providers.
Will impact estimates be developed for additional day types beyond what the protocol specifies?	Impacts will be estimated for the day type the protocol specifies (each event day and Average Event Day) as impacts are not assumed to persist across non-event days.
Whether any additional investigations be conducted to determine why the impacts are what they are, rather than simply reporting the estimates?	Ongoing involvement with SDG&E program staff should provide expert context to program performance, but no additional metering or analysis will be performed.
Are there expected to be free riders or structural winners among program participants? If so, will there be efforts to identify their number or frequency within all participants?	The incidence of free ridership is expected to vary based on program design and participant makeup. In general, programs that rely on control groups will address issues of free ridership.
Whether a control group will be used for impact estimation and how will it be constructed to avoid introducing bias?	Matched control groups and synthetic controls will be used for programs where there is a comparable group of non-participants.
Whether common methodology or data will be used across multiple utilities that have implemented the same DR resource.	A common methodology will be used for PG&E, SCE, and SDG&E program impacts. However, no data will be pooled for modeling and no participant information will be shared across IOU's.

2.1 KEY RESEARCH QUESTIONS

Different evaluation methods will be applied to each subgroup, given the distinct program subgroups and populations enrolled. However, the overall goals for each subgroup's evaluation are the same–to answer these key research questions:

- What were the demand reductions due to program operations and interventions in 2024[2] – for each event day and hour and for the average event? How do these results compare to the ex post results from the prior year and why?
- How do load impacts differ for customers who have enabling technology and/or are dually enrolled in other programs?
- How do weather and event conditions influence the magnitude of demand response?
- How do load impacts vary for different customer sizes, locations, and customer segments?
- What is the ex-ante load reduction capability for 1-in-2 and 1-in-10 weather conditions? And how well do these reductions align with ex-post results and prior ex-ante forecasts?
- What concrete steps can be undertaken to improve program performance?

2.2 DEMAND RESPONSE EVALUATION METHODS

The primary challenge of impact evaluation is the need to accurately detect changes in energy consumption while systematically eliminating plausible alternative explanations for those changes, including random chance. Did the dispatch of demand response resources cause a decrease in hourly demand? Or can the differences be explained by other factors? To estimate demand reductions, it is necessary to estimate what demand patterns would have been in the absence of dispatch – this is called the counterfactual or reference load. At a fundamental level, the ability to measure demand reductions accurately depends on four key components:

- The effect or signal size – The effect size is most easily understood as the percent change. It is easier to detect large changes than it is to detect small ones. For most DR programs, the percentage change in demand is relatively large.
- Inherent data volatility or background noise – The more volatile the load, the more difficult it is to detect small changes. Energy use patterns of homes with air conditioners tend to be more predictable than industrial load patterns.
- The ability to filter out noise or control for volatility – At a fundamental level, statistical models, baseline techniques, and control groups – no matter how simple or complex – are tools to filter out noise (or explain variation) and allow the effect or impact to be more easily detected.
- Sample/population size – For most of the subgroups in question, sample sizes are irrelevant because we plan to analyze data for the full population of participants either using AMI data or end use battery data. Sample size considerations aside, it is easier to precisely estimate average impacts for a large population than for a small population because individual customer behavior patterns smooth out and offset across large populations.

---

[2] The evaluation will cover events occurring from May 2024 to October 2024.

A key factor for many, but not all, demand response resources is the ability to dispatch the resource. The primary intervention – demand response dispatch – is introduced on some days and not on others, making it possible to observe energy use patterns with and without demand reductions. This, in turn, enables us to assess whether the outcome – electricity use – rises or falls with the presence or absence of demand response dispatch instructions.

In general, there are seven main methods for estimating demand reductions, as summarized in Table 3. The first four only make use of use patterns during days when DR is not dispatched to calculate the baseline. The latter three methods incorporate non-event data but also use an external control group to establish the baseline. The control group consists of customers who are similar to participants and experienced the same event day conditions but are not dispatched during events. Control and participant groups should have similar energy usage patterns when the intervention is not in place and diverge when the intervention is in effect. The only systematic difference between the two groups should be that one is dispatched for events while the other group is not.

Table 3: Methods for Demand Response Evaluation

General Approach		Method	Method Description
Use non-event days only to establish the baseline	1	Day matching baseline	This approach relies on electricity use in the days leading up to the event to establish the baseline. A subset of non-event days in close proximity to the event day are identified (e.g., Top 3 of 10 prior days). The electricity use in each hour of the identified days is averaged to produce a baseline. Day matching baselines are often supplemented with corrections to calibrate the baseline to usage patterns in the hours preceding an event – usually referred to as in-day or same-day adjustments.
	2	Weather matching baseline	The process for weather matching baselines is similar to day-matching except that the baseline load profile is selected from non-event days with similar temperature conditions and then calibrated with an in-day adjustment.
	3	Regression models (interrupted time series)	Regression models quantify how different observable factors such as weather, hour of day, day of week, and location influence energy use patterns. Regression models can be informed by electricity use patterns in the day prior (day lags) and in the hours before or after an event (lags or leads) and can replicate many of the elements of day and weather matching baselines.
	4	Machine learning (w/o external controls)	Most machine learning approaches (e.g., random forest, neural networks, etc.) rely exclusively on non-event day data to establish the baselines. The algorithms test different model specifications and rely on a training and testing datasets (out-of-sample testing) to identify the best model and avoid overfitting.

General Approach		Method	Method Description
Use non-event days plus a control group to establish the baseline	5	Matched control groups	Matching is a method used to create a control group out of a pool of nonparticipant customers. This approach relies on choosing customers who have very similar energy use patterns on non-event days and a similar demographic and geographic footprint. The non-event day data is incorporated by either analyzing the data using a regression model, a difference-in-differences model, or both.
	6	Synthetic control groups	This approach is similar to matching except that multiple controls are used and weighted according to their predictive power during a training period. A key advantage of this approach is that it can be used to produce results for individual customers.
	7	Randomized control trials	Participants are randomly assigned to different groups, and one group (the "control" group) is withheld from dispatch to establish the baseline. The control group provides information about what electricity use would have been in the absence of DR dispatch – the baseline. The estimate is refined by netting out any differences between the two groups on hot non-event days (difference-in-differences).

Approaches that use an external control group typically provide more accurate and precise results on an aggregate level when there are many customers (i.e., several hundred). They also make use of non-event days to establish the baseline but have the advantage of also being informed by the behavior of the external control group during both event and non-event days. Except for synthetic controls, the two fundamental limitations to control groups have been: the limited ability to disaggregate results, and the inability to use control groups for large, unique customers. The precision of results for control group methods rapidly decrease when results are disaggregated, and a control group cannot be used to estimate outcomes for individual customers (except for synthetic controls).

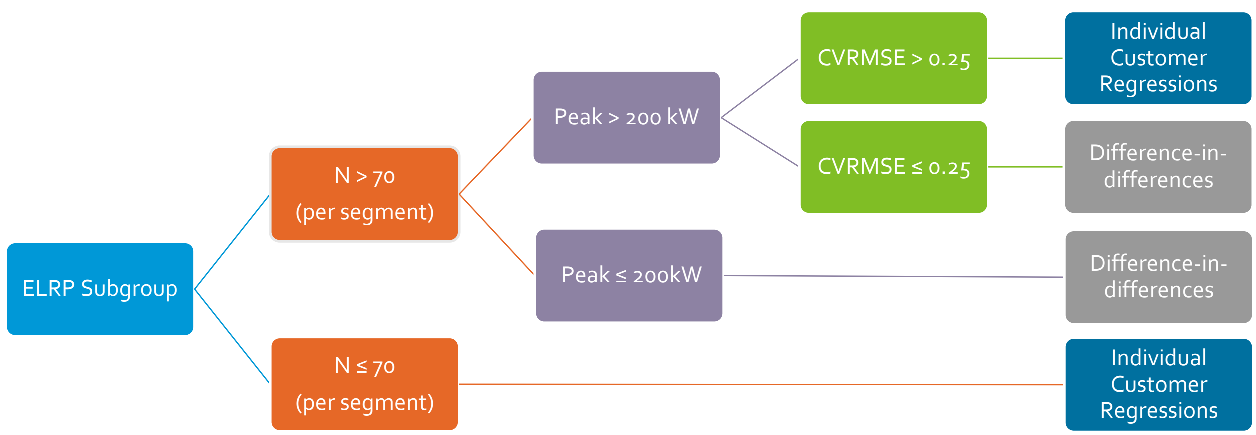
Methods that rely only on non-event days to establish the baseline – such as individual customer regressions – are typically more useful for more granular segmentation. Individual customer regressions have the benefit of easily producing impact estimates for any number of customer segments. Because they are aggregated from the bottom up, the results from segments add up to the totals. However, the success of individual customer regression hinges on having non-event days comparable to event days. When most of the hottest days are event days, as has been the case historically, estimating the counterfactual requires extrapolating trends to temperature ranges that were not experienced during non-event days. This produces less accurate and less reliable demand reduction estimates for the hottest days when resources are needed most.

2.3 MODEL SELECTION

The primary challenge of impact evaluation is the need to accurately detect changes in energy consumption while systematically eliminating plausible alternative explanations for those changes, including random chance. Was the introduction of the ELRP program the primary cause of a customer's change in energy usage or were there other factors involved? To estimate a change in energy consumption, it is necessary to estimate what that energy consumption would have been in the absence of the intervention—the counterfactual or reference load.

The change in energy use patterns was estimated using a combination of difference-in-differences with matched controls and individual customer regressions. Figure 1 summarizes the selection framework that will be used to determine the appropriate method for each site. Most sites will utilize a difference-in-difference model, except for in cases where there are not enough sites in a given segment (customer size and subLAP) or for sites with an annual peak above 200 kW and daily usage patterns which exhibit substantial statistical noise (CVRMSE above 0.25).

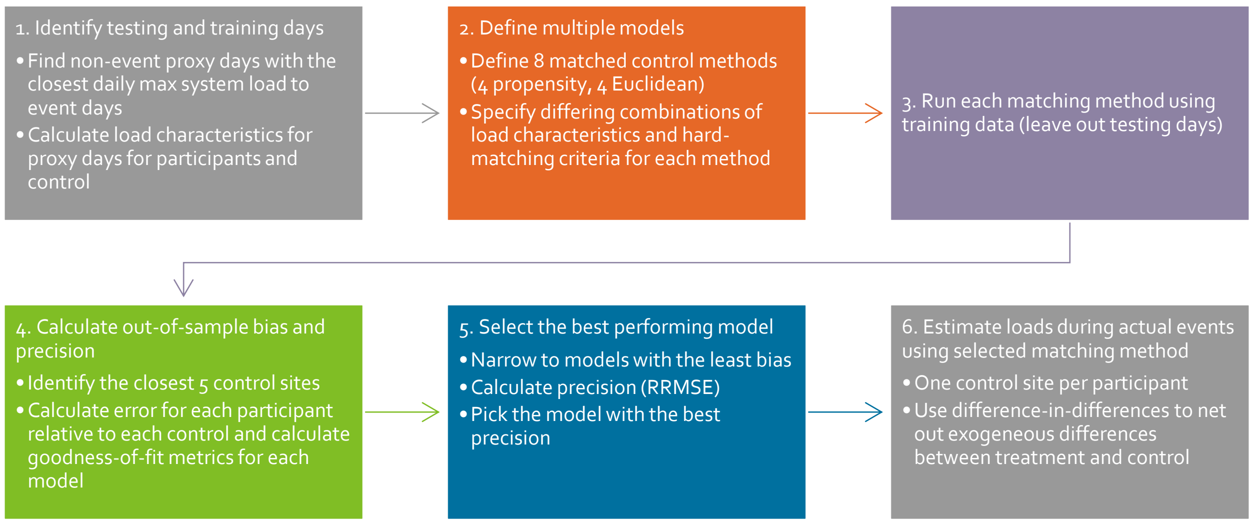
Figure 1: Ex Post Methodology Selection Framework



Site-specific models for individual customer regressions will be selected among dozens of potential specifications, which will include synthetic controls using one or more matched control sites to help control for factors outside of the ELRP events. Similarly, the difference-in-differences approach will use a matched control group to net out changes in energy usage patterns not due to the ELRP events. As such, regardless of evaluation methodology, each participant site will be matched to one or more non-participant using a matching tournament where match quality is compared across eight different matching models to identify the best performing model.

Figure 2 summarizes the process that will be used to select matched controls for the difference-in-difference analyses and synthetic controls for the individual customer regressions. To identify the control pool sites that best match each participant site's energy use patterns on event-like, proxy days (similar in weather and system conditions to event days), eight matching methods will be tested. These methods include different matching algorithms (e.g. Euclidean and propensity matching) and different site characteristics. Matching methods include different combinations of proxy day load characteristics such as load factor, load shape, and weather sensitivity. Control candidates will also be "hard-matched" on subLAP, net metering status, and size bin[3].

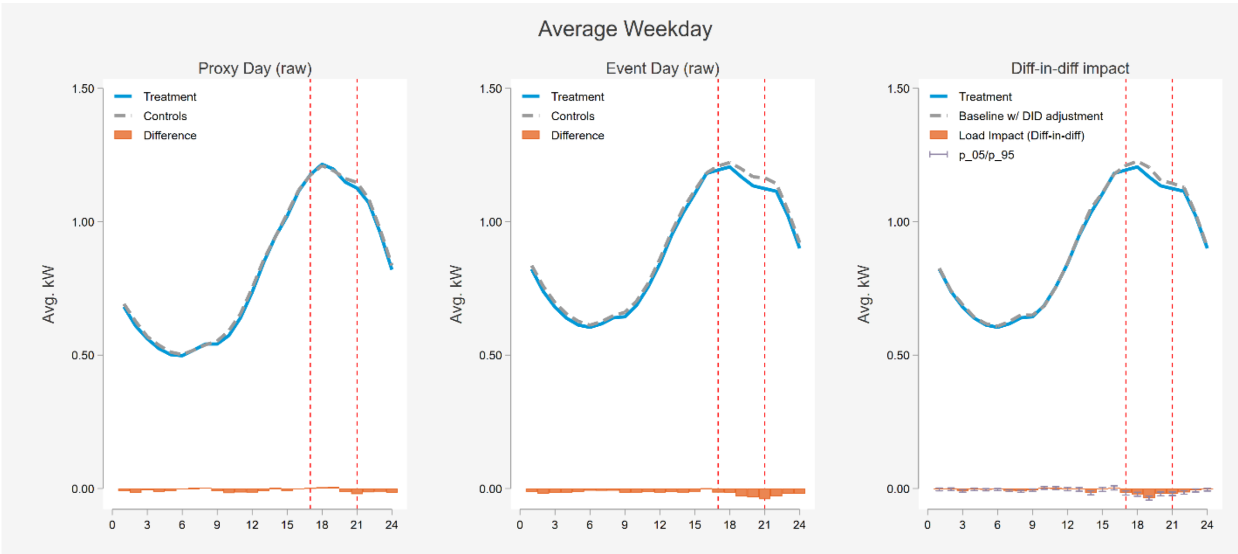
Figure 2: Out of Sample Process for Control Group Selection



As described above, difference-in-differences with matched controls will be the primary evaluation methodology used, except in cases where there were few sites or large sites with noisy load patterns. Figure 3 below demonstrates the mechanics of a difference-in-difference calculation. In the first panel, average observed loads on proxy days are shown for participants and for their matched controls. The difference between these two is the first "difference" and quantifies underlying differences between participants and their controls not attributable to event participation. Note that this first difference is very small, indicative of a high-quality match and sufficient sample size to neutralize the noise inherent in individual customer loads. The second panel shows the average observed participant and matched control loads on event days. The gap between these two is the second "difference" which includes both the difference due to event participation and the underlying first difference observable on non-event days. The third panel shows the average event day loads after netting out the proxy day difference from the event day control load. The result is the difference-in-differences impact.

[3] Bins will be constructed using average usage on event-like, proxy days. For solar customers, bins will be constructed based on system size.

Figure 3: Difference-in-Differences Calculation Example



In cases where a difference-in-differences approach is not deemed appropriate due to insufficient sample size or for large sites with noisy loads, site-specific individual customer regression models will be selected using another out of sample tournament to select the most accurate regression model specification for each participant site. To implement out of sample testing, the top 50 system load days, excluding event days, will be randomly divided into testing and training datasets. Bias and fit metrics will be calculated using the testing dataset and the model with the best fit (lowest Root Mean Squared Error) will be selected among models with the least bias (Mean Absolute Error[4]). Site specific load impacts will be estimated using the winning model for each site.

2.4 EX POST IMPACTS

Once the counterfactual event day load has been developed, the difference between that reference load and the observed load is the program impact. Impacts will be reported:

- For each hour on each event day
- For the average event hour on the average event day

As alluded to earlier, ex post impacts will also be reported out for particular sub-segments of enrolled participants. While the exact segments will vary depending on the subgroup, the typical set of segments include the following:

[4] MAE will be used rather that Mean Average Percent Error (MAPE) to ensure robustness for sites with loads very close to zero, common for sites with solar or other generation.

- **Region:** Local Capacity Area, SubLAP, and climate zone
- **Industry:** for non-residential customers, identified by customer's NAICS code
- **Size:** peak demand less than 20kW, between 20-200kW, and greater than 200kW
- **Dual Enrollment:** either dually enrolled with another program or not
- **NEM/Solar Status:** for residential programs with high penetration of rooftop solar, identified by non-NEM, solar only, storage only, and solar + storage

Program specific and portfolio adjusted impacts will be developed for each subgroup. The fundamental difference that necessitates having these two sets of results is grounded in the ability of customers to participate in more than one energy saving program.

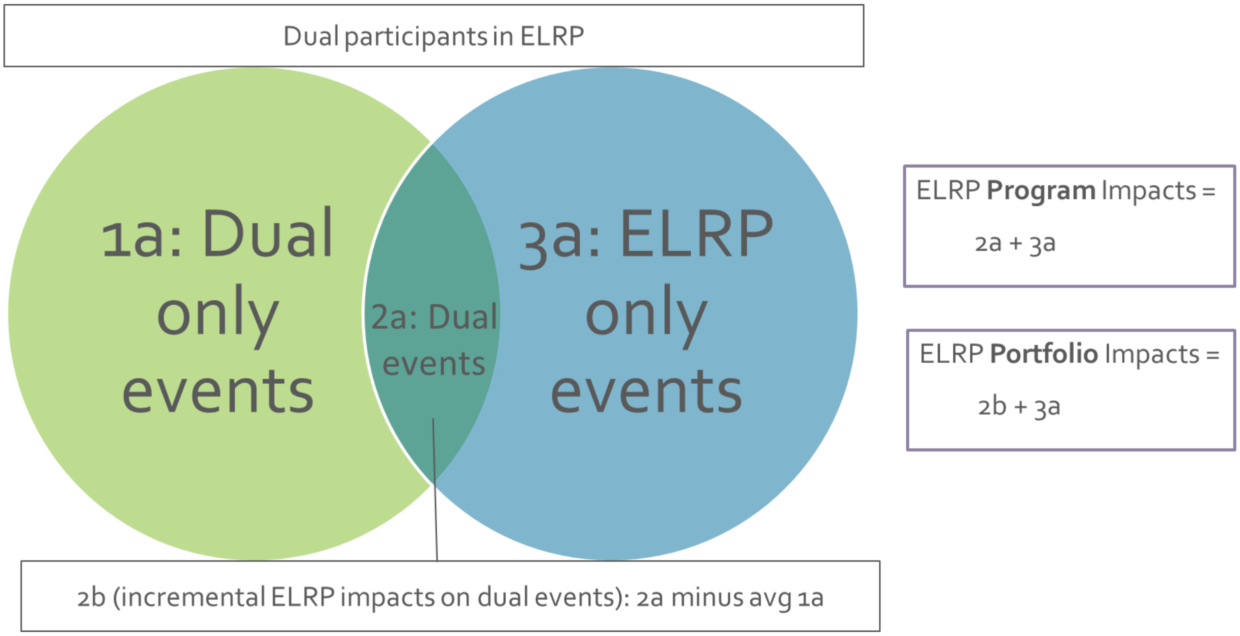
Since customers are allowed to participate in more than one energy saving program, proper attribution of savings estimates is essential, to avoid double-counting. Ex post results are properly attributed by calculating the incremental impacts, or the load reduction beyond what was predicted or committed on dually called event hours (Table 4).

Table 4: Ex Post Load Impact Attribution Strategy for Customers Dually Enrolled and Dispatched

Dual Group	Study	Ex-post
ELRP A1 + CPP	ELRP	Full Impacts reported
	CPP	Impacts removed from program average; duals' impacts on dual events not in report
ELRP B2 + CBP	ELRP	Any impacts beyond nomination
	CBP	Impacts are capped at nomination

As described by the portfolio-adjusted definition, the ELRP evaluation will need to allocate impacts for dual enrolled customers in a way that avoids double counting. For CBP, the FSL value or the CBP nomination level can be used as a threshold. For CPP, a split in impacts will need to be determined between the two programs for dual customers on overlapping events (if any). The ELRP evaluator will determine this split using the methodology in Figure 4.

Figure 4: Dual Enrollment Impact Allocation



NET VS DELIVERED LOAD

Compensation rules for vary by ELRP subgroup, as summarized in Table 5. Evaluation analyses will be conducted in alignment with these rules and will therefore be conducted on net loads for all subgroups.

Table 5: ELRP Subgroup Export Compensation Rules

Sub-Groups	Delivered or Net Load used for LIP Evaluation?
A1	Net
A2	Net
A4 VPP Aggregator	Net and Delivered
A5 EV/VGI Aggregator	Net
B2 CBP	Net

* Exporting accounts can elect to include exports in their compensation calculations. For these accounts, loads used for evaluation will align with those used for compensation.

2.5 EX-ANTE IMPACTS

A key objective of the DR evaluations is to quantify the relationship between demand reductions, temperature, hour-of-the-day, and dispatch strategy. The purpose of doing so is to establish the demand reduction capability under 1-in-2 and 1-in-10 weather conditions for planning purposes and, increasingly, for operations. When possible, we rely on the historical event performance to forecast ex-ante impacts for future years for different operating conditions.

At a fundamental level, the process of estimating ex-ante impacts is simple:

1. Use at least two years of historical performance data
2. Decide on an adequate segmentation to reflect how the customer mix evolves over time
3. Estimate the relationship between reference loads and weather
4. Use the models to predict reference loads for different weather conditions (e.g., 1-in-2 and 1-in-10 weather year conditions)
5. Estimate the relationship between weather and percent impacts
6. Predict percent reductions for different weather conditions (and/or dispatch hours)
7. Combine the reference loads (#4) and percent reductions (#6) to produce per-customer impacts
8. Multiply per-customer impacts by the enrollment forecast

The process can be used to develop ex-ante estimates of demand reduction as a function of temperature, event start time, and event duration. It can be used to develop estimates for 1-in-2 and 1-in-10 weather year planning conditions, and it can be used to develop time-temperature matrices useful for estimating reduction capability for operations or a wider range of planning conditions.

The conversion of ex post impacts to an ex ante forecast should be transparent and understandable to outside stakeholders. In general, the differences between the two are due to several key distinctions:

1. **Customer Mix:** Difference in participant population mix or forecasted enrollment
2. **Weather:** Ex post observed weather may be hotter or colder than ex ante planning conditions
3. **Event Time:** Ex post events may not occur during the RA window for which ex ante impacts are developed
4. **Historical Data:** Ex ante data should explicitly incorporate multiple years of impacts, so average impacts may change when additional years of ex post data are included
5. **Program Design:** If dispatch strategy, eligible months, or program participation options change, ex post impacts may not represent the future capability of the program

As part of the reporting process, we will capture the impact each of these changes has on the difference between ex post and ex ante impact estimates.

Finally, as the results of demand response impact evaluations are increasingly used to support operational concerns, the evaluation team will also provide time-temperature matrices for all subgroups. These matrices will rely on the ex ante impact estimates to predict, for different event start times, durations, and weather conditions, what the average customer hourly impact could be. This will be provided to SDG&E program staff separately from the ex ante load impact tables.

At SDG&E's discretion, the evaluation team may produce, in parallel to the current ex ante reporting requirements, ex ante estimates to PY2024 programs in the upcoming reporting format as described in the ongoing Resource Adequacy proceedings. As proposals in this proceeding are not yet finalized or approved, the evaluation team makes no attempt to summarize the specific reporting requirements here. The new proposal is expected to be finalized in the next six months and will apply to the 2024 compliance year.

For each subgroup, a slice-of-day table will be provided in addition to the standard weather year ex-ante impact tables. A slice-of-day table shows the hourly impacts for the worst day of each month based on the year selected.

Program Specific versus Portfolio Adjusted Impacts

Program specific and portfolio adjusted impacts will be developed for each subgroup. The fundamental difference that necessitates having these two sets of results is grounded in the ability of customers to participate in more than one energy saving program. Dual enrollments make proper attribution of savings estimates essential, to avoid double-counting. Ex post results are properly attributed by calculating the incremental impacts, or the load reduction beyond what was predicted or committed on dually called event hours.

Program specific ex ante estimates, which are the unadjusted impacts of the program, are calculated by using ELRP-only and dually enrolled customers on all ELRP event days. Summing up program specific aggregate ex-ante estimates across all evaluation reports could generate double counting of impacts. Portfolio adjusted ex ante estimates are the population's incremental savings generated by ELRP dispatch. These impacts avoid double counting across evaluation reports, which allows for summing up aggregate ex-ante estimates across all evaluation reports to get an estimate of SDG&E's portfolio of DR programs.

Table 6 defines the dual enrolled programs for consideration in each subgroup. If there are no dual enrollments allowed or there were no dual events in a given season, the program impacts will equal the portfolio impacts.

Table 6: Ex Ante Load Impact Attribution Strategy for Customers Dually Enrolled and Dispatched

Dual Group	Study	Ex-Ante Program Specific	Ex-Ante Portfolio Adjusted
ELRP A1, A6 + CPP	ELRP	ELRP and overlapping events, single and dual customers	CPP event average removed from impacts
	CPP	CPP and overlapping events, single and dual customers	Ex ante impacts estimated based on ex post data from non-ELRP event days
ELRP B2 + CBP	ELRP	ELRP and overlapping events, single and dual customers	Any impacts beyond nomination
	CBP	CBP and overlapping events, single and dual customers	Impacts are capped at nomination

2.6 EXECUTIVE SUMMARY AND CPUC ENERGY DIVISION REQUESTS

A requirement over the last several years has been to provide supplemental reporting to the Energy Division for long term planning. For all programs in SDG&E's PY2024 portfolio, including the statewide programs, several additional reporting features are due to the CPUC on or before November of 2025. These requirements are as follows, with both a public and confidential version enclosed:

1. Ex Ante Load Impacts in plain Excel format, due on or before April 1st of each year:
    a. Portfolio aggregate ex-ante load impacts for 1-in-2 weather year monthly system peaks for each of the 10 ex-ante forecast years, for both the IOU's service area and each LCA within the service area
    b. Portfolio aggregate ex-ante load impacts for 1-in-10 weather year monthly system peaks for each of the 10 ex-ante forecast years, for both the IOU's service area and each LCA within the service area
2. Portfolio aggregate ex-ante load impacts by program for 1-in-2 year August system peak for each of the full ex-ante forecast period years, disaggregated by WECC busbar. Due by November 1
3. Portfolio aggregate ex-ante load impact by program for the 1-in-2 weather year monthly system peak in the final year of the forecast, for all program operating hours (not just RA window). Document the methods used to estimate non-RA hour impacts. Due by November 1.

Demand Side Analytics will construct these tables.

3 COMMERCIAL SUBGROUPS (A.1, A.2, A.5, B.2)

ELRP subgroups A.1, A.2, A.5, and B.2 are comprised of primarily commercial customers who have different program eligibility requirements and load patterns than the residential subgroups A.4 and A.6. This has several implications for our evaluation approach; it determines our strategy for matching, modeling loads and event day impacts, and forecasting ex ante impacts. This section details our evaluation strategy for the commercial subgroups.

3.1 PROGRAM CHARACTERISTICS THAT INFLUENCE EVALUATION

Subgroups A.1, A.2, A.5, and B.2 are targeted at various types of non-residential resources, including large customers, aggregated resources, dual participants in other programs (such as BIP and CBP), and exporting DERs. While a few subgroups have a few thousand participating sites, like A.1, most have very few. Further, aside from small commercial A.1 participants, the commercial subgroups tend to be large with unique loads. Given these challenges, it is likely not feasible to find a matched control, for each customer in the commercial subgroups, that sufficiently explains energy usage patterns not due to ELRP events. To evaluate these unique subgroups, we will be using a combination of difference-in-differences with matched controls and individual customer regressions.

Ten ELRP events have been called to date in summer 2024 for the commercial subgroups. Not every commercial subgroup was dispatched for every event, and for some events start and end times varied by subgroup. Table 7 summarizes characteristics of the commercial subgroups that are relevant to the evaluation approach.

Table 7: ELRP Commercial Subgroup Characteristics Relevant to Evaluation Approach

Metric	A.1, A.2, A.3, A.5, B.2
Historical events	▪ 2021 – 3 ▪ 2022 – 10 ▪ 2023 – 9
Dual participation	▪ With CPP, CBP
PY2023 Ex-ante Estimates	▪ 2024 Typical Event Day 1-in-2 Ex Ante: ➢ 26.65 MW (A.1) ➢ -0.03 MW (A.2) ➢ 0.02 MW (A.5) ➢ 1.61 MW (B.2)

IMPLICATIONS FOR ANALYSIS

- A.1 & B.2
  - ✓ Large industrial customers may have few feasible control candidates with similar load patterns, so we will use individual customer regressions.
  - ✓ Where there is a sufficient subset of small commercial customers with similar load patterns, we will use difference-in-differences.
- All other groups: Small number of participants with unique loads may not be sufficient for precise impact estimates, may be difficult to draw inferences for ex ante.

The above factors were taken into consideration in selecting our planned evaluation approach, presented below.

3.2 EVALUATION APPROACH

Table 8 summarizes our proposed evaluation approach. For clarity, we present key components of the ex-post (Table 8) and ex-ante (Table 9) load impact estimation separately. The ex-post evaluation is direct and relies on a simple, transparent method.

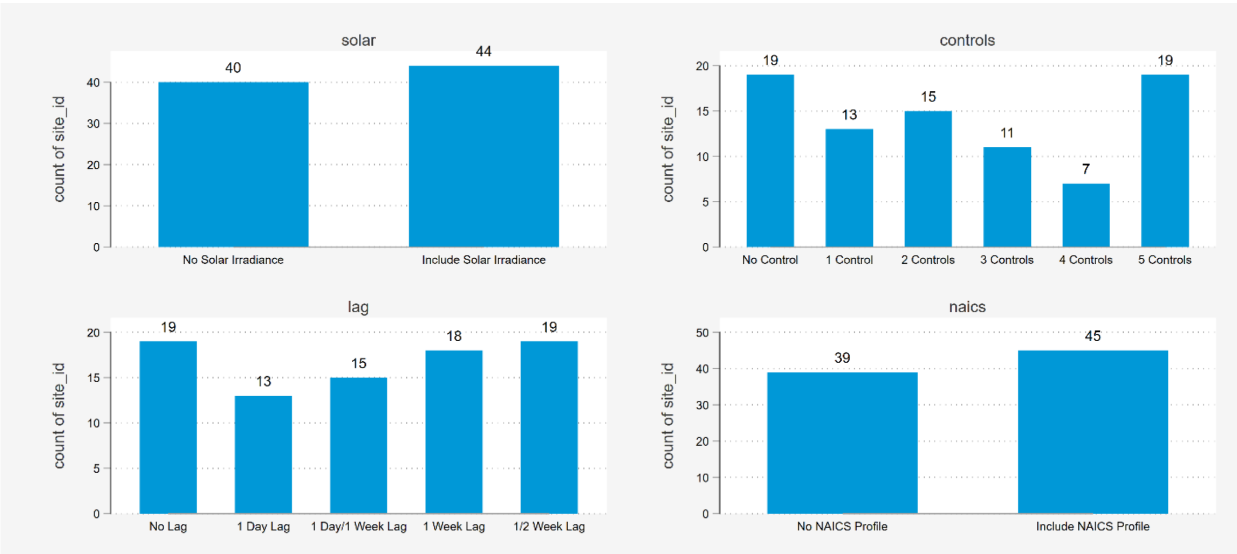
Table 8: A.1, A.2, A.5, B.2 Ex-Post Evaluation Approach

Methodology Component	Demand Side Analytics Approach
1. Population or sample analyzed	For the commercial subgroups, our plan is to analyze the full population of participants and utilize synthetic controls where a matched control group cannot be constructed.
2. Data included in the analysis	The analysis will include all PY2024 data. Additional data may be included if event conditions are substantially hotter than non-event days.
3. Use of control groups	For large commercial customers with unique load patterns, synthetic controls may be included in site specific models as right-hand variables. Where there is a sufficient number of small commercial customers in A.1 and B.2, we will be constructing a matched control group. The control group will be selected using non-event day load patterns, geographic location, and other customer characteristics (e.g., industry) to develop propensity scores within each stratum. A matching model tournament will be used to identify the best performing Euclidean matching or propensity matching specification based on the quality (out-of-sample bias and fit) of each matching method.
4. Model selection	An out-of-sample model selection tournament is used to pick the best performing model for each site across multiple parameters and dozens of model specifications. In the out-of-sample process, data is systematically left out of the model then predicted on to assess counterfactual performance—a well performing model should predict loads reasonably well on days which were not used in the model. The final model is identified based on the least bias (% Bias) and the best fit (Relative RMSE) metrics. An out-of-sample process can also be used to select site specific synthetic controls. The model parameters that will be included in the site-specific model tournament include

Methodology Component	Demand Side Analytics Approach
	industry profile, number of synthetic controls, solar irradiance and lags meant to capture various scheduling patterns, e.g., daily, weekly, bi-weekly, etc. The differences-in-differences models will compare hourly electricity use during events and outside of event hours for both the program participants and the matched controls. The control group experiences the same weather and other conditions as the participants over time, but they are not dispatched for an event. Thus, the control group's usage during an event serves to remove any differences between the treatment and control group that remain after the matching process.
5. Segmentation of impact results	The results will be segmented by: ▪ Region: LCA, subLAP, and climate zone ▪ Industry: for non-residential customers, identified by customer's NAICS code ▪ Size: peak demand less than 20kW, between 20-200kW, and greater than 200kW ▪ Dual Enrollment: either dually enrolled with another program or not ▪ Customer service level, e.g. whether or not a customer service representative has been assigned (A.1 only) The main segment categories are building blocks. They are designed to ensure segment-level results add up to the total, to enable production of ex-ante impacts, and to allow for busbar level analysis.

Figure 5 shows the different model parameters that were included in the site-specific model tournament for SDG&E's PY2024 ELRP evaluation, including industry profile, number of synthetic controls, solar irradiance and lags meant to capture various scheduling patterns, e.g., daily, weekly, bi-weekly, etc. The figure also tabulates the number of sites whose winning model included each parameter. The wide spread across parameters indicates that it was important to allow for individually tailored models to be selected for each participating site.

Figure 5: Parameters Tested, Inclusion in Best Performing Site-Specific Models, PY 2023



Using the standard ex ante estimation process relies on deriving a weather-based impact model to predict impacts given weather conditions, either directly or as a percent of reference loads. However, in our experience evaluating events for non-residential ELRP subgroups, there has been no clear relationship between weather and impacts. This is logical given that the bulk of enrolled resources are either are dispatchable storage (A.4, A.5) or curtailed load from large C&I customers whose load is largely not weather sensitive. Furthermore, the majority of loads and reductions can be driven by a handful of very large customers with unique load and response profiles, which also presents challenges for estimating reference loads. DSA will carefully assess the weather sensitivity of PY 2024 load impacts for SDG&E ELRP subgroups. Ex ante impacts will only be differentiated by weather condition if ex post impacts exhibit clear weather sensitivity.

Ex ante reference loads will be developed for each ELRP subgroup based on the load patterns observed in PY 2024 for the PY 2024 participant population. Reference load and impact forecasts for future years will be scaled to enrollment forecasts provided by SDG&E and will reflect the level of granularity of these forecasts. For example, if the share of small versus large commercial participants in A.1 is expected to remain relative steady, a total A.1 forecast can be applied to scale the A.1 impacts and reference loads. If this share is expected to change meaningfully, however, it is recommended that enrollment forecasts be segmented.

Table 9: A.1, A.2, A.5, B.2 Ex-Ante Evaluation Approach

Methodology Component	Demand Side Analytics Approach
1. Years of historical performance used	Where possible, we plan to use three years of historical data to estimate how demand reductions vary based on dispatch hours and weather conditions and to estimate the reductions available under planning conditions.

Methodology Component	Demand Side Analytics Approach
2. **Process for producing ex-ante impacts**	The key steps will be: ▪ Use three years of historical performance data for relevant customers. ▪ Decide on an adequate segmentation to reflect changes in the customer mix. ▪ Estimate the relationship between reference loads and weather. ▪ Use the models and ex ante weather conditions to predict reference loads for 1-in-2 and 1-in-10 weather year conditions. ▪ Estimate the relationship between impacts or percent impacts and reference loads. ▪ Use the models to predict impacts for 1-in-2 and 1-in-10 weather year conditions. ▪ Incorporate the enrollment forecast.
3. **Accounting for changes in the participant mix**	Because the customer mix may change, changes in the participant mix need to be accounted for when developing forecasts of reduction capability under planning conditions. From the outset, we produce a detailed segmentation – building blocks – so we are able to account for changes in the customer mix over the historical and forecast periods.

4 RESIDENTIAL SUBGROUPS (A.4, A.6)

ELRP subgroup A.4 and A.6 are primarily comprised of residential customers and aggregations of residential customers. Due to the large number of A.4 enrollments and suitable controls, it is feasible to find a well-matched control group that makes measuring impacts simple and straightforward. This has several implications for our evaluation approach; it determines our strategy for matching, modeling loads and event day impacts, and forecasting ex ante impacts. This section details our evaluation strategy for the residential subgroups.

4.1 PROGRAM CHARACTERISTICS THAT INFLUENCE EVALUATION

Subgroup A.6 was not dispatched for events during PY 2024, so ex post impacts will not be evaluated and many evaluation considerations for this subgroup will be irrelevant in this program year. However, ex ante impacts will still be produced for A.6.

Subgroup A.4 was dispatched for five events for SDG&E. Ex post and ex ante impacts will be evaluated for A.4. Table 10 summarizes key information about the program that is relevant to the evaluation.

Table 10: ELRP A.4 Program Characteristics Relevant to Evaluation Approach

Metric	Value
Historical events	▪ 2022 – 10 ▪ 2023 – 2
Dual participation	▪ A.6 CPP
PY2023 Ex-ante Estimates	▪ 2024 Typical Event Day 1-in-2 Ex Ante: ➢ 1.14 MW (A.4) ➢ 17.2 MW (A.6)

IMPLICATIONS FOR ANALYSIS

- A.4: Residential loads. Primary analysis done using whole building meter data.
  - ✓ Can supplement/validate with end use data as available (e.g., battery data, thermostat runtime data). End use data eliminates noise of non-controlled loads but cannot support matched control selection.
- A.6: Residential loads. Large default population will necessitate performing evaluation on sample, census may be preferred for small opt-in population.

4.2 EVALUATION APPROACH

Table 11 and Table 12 summarize our planned approaches for the ex-post and ex-ante evaluations, respectively.

Table 11: A.4, A.6 Ex-Post Evaluation Approach

Methodology Component	Demand Side Analytics Approach
1. Population or sample analyzed	Our plan is to analyze the full population of A.4 participants with an appropriate matched control group.
2. Data included in the analysis	The analysis will include all PY2024 data. Additional data may be included if event conditions are substantially hotter than non-event days.
3. Use of control groups	A matched control group will be employed for residential customers. Control customers will be pulled from a stratified random sample, which ensures that large and/or unique participants are still likely to find an appropriate match. The control group is selected using non-event day load patterns, geographic location, and other customer characteristics (e.g., industry) to develop propensity scores within each stratum. For each participant, the nearest neighbor based on propensity scores or Euclidean distance is identified. Our typical process is to specify 10 to 20 combinations of stratification, matching methods, and variables used for scoring. Thus, we pick 10 to 20 potential matched control groups and assess their accuracy. This is accomplished by merging the hourly interval using days that were not included in the matching process. We select "event-like" proxy days for our out-of-sample testing in order to accurately capture the performance of the model under event-like conditions (typically very hot weekdays). This allows us to assess out-of-sample how well each candidate control group predicts the participant group's load patterns on their own and to calculate metrics for bias (MPE) and for estimation noise (MAPE and CVRMSE). Thus, the DSA approach ensures a control group that has nearly identical load patterns as participants in the absence of an event.
4. Model selection	The differences-in-differences models will compare hourly electricity use during events and outside of event hours for both the program participants and the matched controls. For customers participating in ELRP events, we should observe: 1. Nearly identical usage patterns between the treatment and control groups on non-event days 2. A change in load during events 3. No similar change for the control group 4. The timing of the change in energy use should coincide with the event start The control group experiences the same weather and other conditions as the participants over time, but they are not dispatched for an event. Thus, the control group's usage during an event serves to remove any differences between the treatment and control group that remain after the matching process. DSA will supplement/validate with end use data as available (e.g., battery data). End use data eliminates noise of non-controlled loads but cannot support matched control selection.

Methodology Component	Demand Side Analytics Approach
5. Segmentation of impact results	The results will be segmented by: ▪ Region: LCA, subLAP, and climate zone ▪ NEM/Solar Status: for residential programs with high penetration of rooftop solar ▪ Dual Enrollment: either dually enrolled with another program or not The main segment categories are building blocks. They are designed to ensure segment-level results add up to the total and to enable production of ex-ante impacts, including busbar level results.

For the A.6 subgroup, the key difference between ex post and ex ante is the production of weather-normalized reference loads, whereby participant loads are expressed as functions of key weather variables. Then, using the ex ante weather forecasts provided by SDG&E for both 1-in-2 and 1-in-10 weather years, reference loads are predicted using the same regression function. In general, for weather sensitive programs, hotter weather produces higher load impacts as there is more cooling load to shed. The relationship between temperature and percent impacts are modeled using a simple regression, and then applied to the reference loads to generate ex ante impacts.

For the A.4 subgroup, which is comprised of battery storage responding to dispatch signals, impacts can be assumed to be a function of the battery capacity made available by participants. Ex ante reference loads will be developed based on the load patterns observed in PY 2024 for the PY 2024 participant population. Reference load and impact forecasts for future years will be scaled to enrollment forecasts provided by SDG&E and will reflect the level of granularity of these forecasts.

Table 12: A.4 Ex-Ante Evaluation Approach

Methodology Component	Demand Side Analytics Approach
1. Years of historical interconnected dispatchable generation capacity	Where possible, we plan to use three years of historical data on growth of interconnected dispatchable generation capacity in SDG&E territory.
2. Process for producing ex-ante impacts	The key steps will be: ▪ Forecast total nameplate capacity interconnected in SDG&E territory. ▪ Estimate technical potential (amount of capacity available for dispatch). ▪ Estimate feasible potential (refine enrollment assumptions and incorporate program marketing assumptions). ▪ Use the forecasts to predict hourly impacts for 1-in-2 and 1-in-10 weather year conditions (only a function of installed capacity, not weather-dependent).
3. Accounting for changes in the participant mix	Any anticipated changes in the participant and technology mix will be incorporated into the ex-ante forecasting.

5 QUALITY CONTROL PROCEDURES

The Demand Side Analytics team takes analysis accuracy seriously. We have several processes in place to ensure all data management, analysis, and reporting are delivered with the highest quality. A summary of our philosophy, however, is enumerated below:

1. There is clear oversight in each project by an expert in Demand Response evaluation. Our senior staff are familiar with the types of programs being evaluated, the preferred methods and their respective strengths and weaknesses, and the California demand response landscape. We understand these programs and their evaluation challenges.
2. Whenever possible, we rely on automated reporting and tabulation. This allows us to go from data validation to reports quickly and efficiently, without errors caused by version control, manual data entry, or copy and paste errors.
3. We understand the reporting requirements to conform to the California Load Impact protocols. Because of our background, we don't anticipate surprises in the format, content, or timeline of the key project deliverables, which means that SDG&E get the right information at the right time in a clear, accessible format.

5.1 DATA CHECKS

The first step for quality control is to make sure that all data that had been requested is both accounted for and does not contain spurious values. To that end, we have implemented a detailed checklist for our demand response evaluations that investigates common data pitfalls for each type of data typically used in a demand response evaluation. A summary of these questions typically includes:

1. **Interval Data:** Is the data in the right units? Adjusted for Daylight Savings and any grid export/net demand? Is there a full panel of data for all customers? Are there outliers in terms of customer size? Did we receive all the interval data for the customers we requested?
2. **Customer Characteristics:** Do we have all the relevant participant and control groups? Do we have DR enrollment data for all customers and were they affected by other interventions during the analysis period? Do we have all the characteristics that are needed for reporting?
3. **Treatment and Event Data:** Do we have the correct event days identified? Are the event days and hours properly coded? Can we visually see when customers are reducing loads during events?
4. **Weather Data:** Is the DST adjustment in the weather data consistent with that of the interval and event data? Is it in the right time zone and units?

Because incorrect data will lead to incorrect results, any issues that are identified to be significant to the evaluation will be addressed with the SDG&E team to ensure quick resolution.

5.2 ANALYSIS CHECKS

Analysis checks are critical to a successful evaluation, and where our expertise in DR evaluations will provide value. Because of our familiarity with these demand response programs and the California load impact protocols, we are able to quickly identify results that do not make sense and either correct the issue or identify the reason why results differ from our initial assumption. While analysis checks tend to be program specific, the general considerations are:

1. **Analysis Dataset Construction:** Is the control group constructed appropriately? Is it statistically indistinguishable from the treatment group on days when no customer was dispatched? What are the result of out of sample testing? Given model precision and bias, will we be able to detect the expected effect?
2. **Ex post results:** Are the results generally in line with prior years, given no substantial program changes? Are all customers dispatched as expected? Do weather sensitive programs see greater impacts on hotter days? Do reference load patterns follow the same trend as the raw data with regards to temperature? What are the distributions of impacts - are there large customers that are driving the majority of impacts? Are there particular customer segments that respond differently?
3. **Ex ante results:** Given the differences between ex post and ex ante weather and participation, do reference loads look appropriate for each day type and weather year? What about percent impacts? Have we captured the effects of dual enrollment for program and portfolio impacts appropriately? Have changes to program design or enrollment been captured in the ex ante forecasts?

The focus of these questions is to ensure that there are no surprises in the evaluation report and that all results are situated in their full context. In collaboration with the SDG&E team, the evaluation team will work to frequently share draft findings and raise any issues as they arise.

5.3 REPORTING CHECKS

Many iterations are expected in the process of producing draft and final evaluation reports, load impact tables, and other results memos. In those cases, opportunities arise for omissions, copy/paste errors, and gaps in reporting updates. To the extent possible, the evaluation team relies on automated reporting and table generation, where the latest version of the analysis is automatically written into a report. This ensures that reports and load impact tables are consistent in their results, and that all values are updated whenever an updated version of the analysis is implemented.

5.4 PROJECT MANAGEMENT CHECKS

As discussed in the kickoff meeting, Alana Lemarchand will be the key contact for all project management topics. They will both be responsible for ensuring that the project remains on time and on budget and will identify bottlenecks or issues likely to affect the project timeline as soon as possible to the SDG&E team. As part of this process, monthly reporting on budget, key tasks completed, upcoming deliverables, and any changes to the schedule will be provided to the SDG&E team.

6 TIMELINE

Figure 6, below, shows the next steps for the evaluation of SDG&E's ELRP program.

Figure 6: Timeline of Key Deliverables

	Task	Deliverables	Timing
1	Project Management	Regular Meetings	September 2024 – March 2025
		Kick-Off Meeting	8/29/2024 Memo due 3 business days after kick-off meeting
2	Evaluation Plan	Draft Evaluation Plan Final Evaluation Plan	Draft Plan due 10 business days after kick-off meeting Final Plan due 5 business days after comments received on draft
3	Data Collection and Validation	Data Request	Due 10 business days after kick-off meeting
4	Ex-Post Results	Draft and Final Result Spreadsheets	Draft Results by 12/02/2024 Final Results by 12/16/2024
6	Ex-Ante Results	Draft and Final Result Spreadsheets	TTM Results by 1/17/2025 Draft Results by 1/24/2025 Final Results by 2/7/2025
7	Documentation & Reporting	Draft and Final Evaluation Report	Draft Report by 2/7/2025 Final Report by 2/28/2025