Beyond Automation:
What Real Homes Reveal About Smart Thermostat Flexibility

Joseph Bourg, Felipe Godinez and Steven Rowsey, Olivine Inc.
Jingjing Liu, Lawrence Berkeley National Laboratory

ABSTRACT

Aggregations of residential smart thermostats are widely accepted in the U.S. as load-flexible devices able to support grid-scale peak load management and load shaping strategies. As part of the California Load Flexibility Research and Deployment Hub (CalFlexHub) administered by Berkeley Lab, Olivine implemented two technology demonstration projects to test and evaluate the reliability and impacts of residential smart thermostats responding to dynamic retail price signals. The per site load reduction observed was generally in agreement with results published in utility thermostat demand response (DR) program evaluation reports. However, analyses of control signal rejection, opt-out rates, and participant interviews indicate that reliable load-flexibility from smart thermostat control depends significantly on user behaviors and experiences influencing device interactions. In this report, we assess prevailing assumptions regarding precooling, particularly that precooling enhances occupant comfort, decreases opt-outs and reduces the degradation of load reduction during the later hours of a control period. Our findings indicate that precooling may have a negative impact on comfort and opt-outs, and that precooling's positive impact on load reduction may be overstated. We further advance temperature as the primary determinant of demand-flexibility potential rather than the position of an hour within a control period as prior studies have concluded. All of these findings are limited by the small sample size, and act only as guidance on what future studies with larger sample sizes and the ability to implement more accurate baseline methodologies should investigate.

Introduction

Smart thermostats represent one of the most scalable residential demand-flexibility resources in the U.S., collectively representing an estimated 19% of electricity consumption via HVAC (EIA 2024) across 34 million devices (Walton 2022). Yet real-world reliability remains inconsistent due to the complex interaction of user behavior, device design, and economic incentives. As utilities increasingly rely on these devices for load flexibility, achieving consistent and reliable load-shifting and peak management benefits require balancing customer comfort, incentive design, and the degree of control accessible to program operators.

Through field demonstration projects for the California Load Flexibility Research and Deployment Hub (CalFlexHub), Olivine implemented seasonal testing of dynamically price-responsive smart thermostat load flexibility at 63 residential sites using two thermostat brands. Olivine's existing integrations with OEM cloud control pathways, utilized in Virtual Power Plant (VPP) operations and wholesale market aggregations, were leveraged to assess three primary objectives: (1) validate and demonstrate the model of third-party Automated Service Providers (ASPs) to optimize residential thermostat response to dynamic retail price responses; (2) understand the user experience and behavioral adaptations when residential thermostats are used for dynamic load-flexibility; and (3) quantify load and cost impacts through coordinated group-based testing.

This paper presents initial recommendations for improving the performance and reliability of smart thermostat-based load flexibility, highlighting areas for future investigation, and identifies pathways for effectively integrating dynamic retail price response into residential offerings without compromising customer experience or cost-effectiveness.

Background on Residential Thermostat Studies

Prior studies and reports from utility residential thermostat control programs were evaluated, including Load Impact Evaluations of Southern California Edison's (SCE) Smart Energy Program, San Diego Gas & Electric's (SDG&E) AC Saver Program, and Pacific Gas & Electric's (PG&E) Smart Thermostat Control Pilot. The studies assessed for this report included only those conducted within California, partially to support a more direct comparison with our own data, and due to a lack of publicly available data sets from other states. These studies generally did not collect data on participant opt-outs or device-level responses to controls, though one study found that opt-out rates were 14.8% in California across the 2019 program season, with opt-outs reducing total load impact of the aggregations by the same percentage (Sarran et al. 2021, 5). Utilities infer from participants surveyed that opt-out rates are relatively low (Jiang, Smith, and Ramee 2021). However, routine disenrollment of participants with frequent opt-outs and low performance likely influences this trend. Rates of devices rejecting control signals are not discussed. Setback setpoint changes ranged up to 4°F, but no further data on setpoint changes was included in the reports.

Control groups are the most common baseline methodology used in prior studies, with 0.54 kW being the average per participant load impact reported across all evaluated reports, a value consistent across device brands and between low income and non-low-income participant groups (Smith and Totten 2020, 2021; Smith and Farr 2022; Bode et al. 2024; Farr, Deych, and Smith 2025). Event duration was identified as the key indicator of load impact, with the largest reductions occurring in the first event hour, and a steep decline in subsequent hours - an observation supported by the data presented in these reports. Figure 1 shows the average results from the reports reviewed by event hour and outdoor temperature bands.

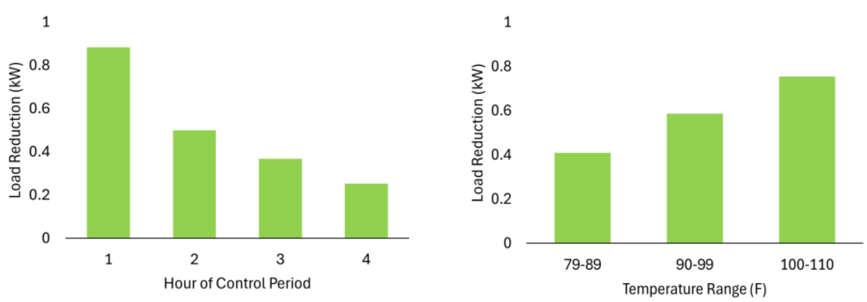

Figure 1. Per site load impact by event hour and temperature range from past reports. *Source*: Smith and Totten 2020, 2021; Smith and Farr 2022; Bode et al. 2024; Farr, Deych, and Smith 2025

A recurring claim in these reports is that precooling improves customer comfort, reduces opt-outs, and mitigates the decline of load impact in subsequent event hours (Smith and Totten 2020, 2021; Smith and Farr 2022; Bode et al. 2024; Farr, Deych, and Smith 2025). However, with a lack of device response data, or any direct comparison between control periods with and without precooling, no persuasive evidence is presented to support this claim.

Methodology

Field Test Sites

Field testing encompassed 75 thermostats installed across 63 participating households - a mix of single-family detached homes, townhomes and multifamily complexes served by PG&E and SCE. Each site was equipped with central HVAC cooling equipment and located within climate zones 9 through 14. Forty-one of the households were identified as Low Income and/or burdened by air pollution ("LI") based on CalEnviroScreen scores. The results and conclusions presented in this report are limited by this small sample size and should be further evaluated by future studies. Figure 2 shows the site distribution by thermostat brand and LI status.

Figure 2. CEC climate zones and LI mapping. Participants are in zones 9 through 14.

Participant Recruitment

The majority of participants recruited for Brand 1 testing had previously participated in a VPP pilot studying the feasibility of behavioral DR among customers located in areas classified as LI. These participants did not have smart thermostats in their homes prior to the CalFlexHub project. As part of enrollment, they were provided with new devices along with professional installation.

Participants recruited for Brand 2 testing, in contrast, already had smart thermostats installed in their homes. These participants learned about the CalFlexHub project through incentive offerings presented in their mobile apps or through email campaigns distributed by the OEM. Olivine worked with Brand 2 to target customers in a select set of zip codes within targeted climate zones. A one-time enrollment incentive was offered to encourage enrollment. It is not known how many alternative programs each customer could choose from, but over time the list of target zip codes was expanded, and the enrollment incentive was increased to recruit a statistically significant number of participants in a timely manner.

Thermostat Control

Olivine's software platform retrieved day-ahead 24-hour seasonal price profiles from the CalFlexHub Price Server. These profiles were designed by Berkeley Lab to support testing of dynamic retail price responsiveness. Thermostat control schedules were optimized to minimize participant energy costs during setback periods (load-decrease intervals) and mitigate the impacts of precool and subsequent snapback. All precooling hours (load-increase intervals) were scheduled immediately prior to the onset of each setback period. Setback schedules were sent to the respective OEM cloud systems via device-control API services. The OEM cloud platforms converted the device control request payloads into device-level dispatches relayed to thermostats in the field. Figure 3 is a selection of seasonal price profiles designed by Berkeley Lab.

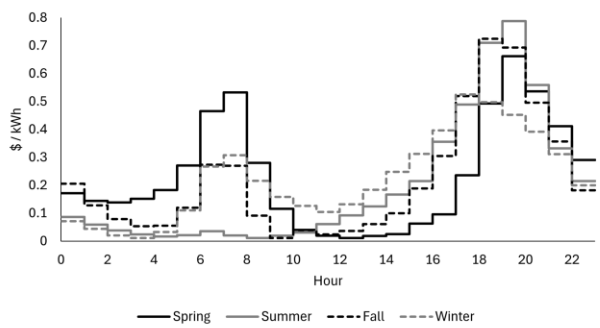

Figure 3. Example 24-hour price signal profiles

Device control API services provided by OEMs are proprietary communications pathways available to third party ASPs for a fee. Fees include annual subscriptions and per device costs. Table 1 shows the control features available and utilized during testing of both brands. Note that Brand 2 provides control parameters that prevent users from manual override or opt out. This feature was not utilized in these projects. All participants were able to override control signals at any time.

Table 1. Features available via control API services utilized in testing Brand 1 and Brand 2.

			Precool control			Setback control	
	Device Level Targeting	Block Override	Setpoint Offset	Duration	Gap Duration	Setpoint Offset	Duration
Brand 1	✓	—	—	—	—	—	✓
Brand 2	✓		✓	✓	✓	✓	✓

Checks indicate features utilized in the creation of control events. "—" indicates features are not available through device control API services. A blank cell indicates that the feature is available but was not utilized during testing. These controls are determined by user preference settings and set by OEM systems.

Testing Notifications

Email notifications were sent to participants one week prior to the start of each seasonal test period. These notices alerted participants to upcoming control events, specified dates for when testing would begin and end, reminded participants of the purpose for testing automated load flexibility, and described how thermostats were expected to respond during control events. In addition to the email outreach, thermostat users were alerted to the device control events through web and mobile application interfaces implemented by the OEMs. Brand specific messaging within these interfaces communicated the upcoming and active setpoint control periods, specifying start and end times. Both OEMs employed messaging that framed control events as opportunities for cost savings, in one case the interface displayed a piggy bank icon to communicate an active control event.

Seasonal Test Periods

Table 2 details the dates and dynamic retail price response strategies used for control signal optimization in each of the seasonal test periods for both thermostat brands.

Table 2. Control parameters applied during each seasonal test period.

Device brand	Rate profile	Start date	End date	Precool period		Setback period	
				Offset (˚F)	Duration	Offset (˚F)	Duration
Brand 1	Winter, energy, single-day	2/13	2/26	1˚ to 4˚ F*	0 to 1*	1˚ to 4˚ F*	2 consecutive highest price hours during week 1 4 non-consecutive highest price hours during week 2
Brand 1	Spring, energy + GHG***, single-day	5/15	5/26	1˚ to 4˚ F*	0 to 1*	1˚ to 4˚ F*	2 consecutive highest price hours during week 1 4 non-consecutive highest price hours during week 2
Brand 1	Summer, energy, single-day	9/ 11	10/6	1˚ to 4˚ F*	0 to 1*	1˚ to 4˚ F*	2 consecutive highest price hours during weeks 1 & 2 3 consecutive highest price hours during weeks 3 & 4
Brand 2	Summer, energy, single-day	7/15	7/28	0˚ or 3˚ F 0˚F week 2	0 or 1 0˚F week 2	3˚ F	2 consecutive highest price hours
Brand 2	Fall, energy, multi-day	10/21	11/ 3	1˚ or 2˚ F**	0.5 or 1**	3˚ F	3 consecutive highest price hours

Brand 2	Spring, energy, multi-day	5/12	5/25	1˚ and 2˚ F 1˚F first 30-min 2˚F second 30-min	1	1˚ to 3˚ F Followed price magnitude per hour	3 consecutive highest price hours.
Brand 2	Summer, energy, multi-day	7/ 7	7/20	0˚ or 2˚ F Staggered daily	0 or 1 Staggered daily	3˚ F	3 consecutive highest price hours.

* Defined by user preference settings; ** 1-hour with -2˚F offset when precool price was less than snapback price and 30-minute with a -1˚F offset when precool price was higher than the snapback price; *** The SpringHDP and GHG signals were combined using an additive method with a GHG scaling coefficient of 1

Between February 2023 and July 2025, Olivine conducted 7 seasonal test periods. The results of all the seasonal test periods are included in this report. Brand 1 tests occurred in winter, spring and summer of 2023. Brand 2 tests occurred in summer and fall of 2024, and spring and summer of 2025. Excluding a 4-week test in summer 2023, all test periods were two consecutive weeks. The number of price responsive control days was limited to three days per test week. Table 2 details the control parameters for each test period.

Summer 2024 and summer 2025 Brand 2 test periods included control events both with and without precooling. In addition, a novel control scheme was implemented for spring 2025 Brand 2 testing. This scheme attempted to differentiate precooling and setback setpoint offsets relative to the hourly price signal. The intention was to set greater setpoint offsets during higher price hours and lower setpoint offsets during the lower high price hours. To achieve this price responsive setpoint modulation, a new control signal was sent at the start of each hour during setback periods. As a result, participants were notified of control events up to 7 times throughout each control period.

Data Collection

Olivine received 15-minute interval whole premise meter data from PG&E and SCE through the Green Button Connect (GBC) system. Meter data was typically available with a 3-day delay. For Brand 1 thermostats, Olivine retrieved device level state data via API after each control event. This data included 5-minute interval data for various state fields including mode (ON, OFF, COOL, HEAT), cooling and heating setpoint, and compressor runtime (seconds/interval). For Brand 2 thermostats, device-level response information was retrieved via API and included each device's final reported status to control events. Olivine used device state and device status data, respectively, to categorically organize device level responses for each control event. Table 3 provides definitions for each device-response status category.

Table 3. Device response status categories.

Status	Description
completed	Customer participated for the full duration of the event.
rejected-load-control	Event was rejected. Thermostat control mode did not match the event type. For example, for a summer (cooling) load shed event, if the thermostat is in heating, auto-heating, or off modes, the event is rejected.

partial-opt-out	Customer opted out of the event at some point prior to the event end time. For example, the customer (manual) or a smart home system modifies setpoint, or changes thermostat mode during control period.
received	Customer has successfully received event requests, and event occurs at some time in the future. Note this status should be superseded once a control period is completed, or there is a partial-opt-out.
No response	No device response status data was returned for the thermostat. Indicates the control signal was not received. This could be due to several factors, but generally the thermostat was disconnected from Wi-Fi.

Participant Education

During Brand 1 testing, a subset of thermostats immediately rejected control signals. To address this issue, informational materials were developed to provide participants with guidance on configuring their thermostats to enable automated controls, including a video with step-by-step instructions. An email campaign was implemented with links to the video published online. No other efforts were made to intervene in how participants used the thermostats.

Measurement and Verification

Whole-premise meter data were used to assess load impacts for both thermostat brands. In addition, for Brand 1, HVAC fan runtime was converted to approximate load using a nameplate capacity of 0.3 kW. The final analysis and comparisons used a 10-in-10 baseline methodology.

The standard 5-in-10 and 10-in-10 baseline methodologies were compared. For the Brand 2 summer 2024 test period, both methodologies produced reliable counterfactual load estimates. For the Brand 2 fall test period, a 10-in-10 baseline produced the more reliable counterfactual. In addition, control group baselines were also constructed using sites that immediately rejected control signals; however, these sites represented a small sample size and consistently exhibited lower loads than responsive sites. As a result, the control group baselines did not provide reliable counterfactuals and were used only for comparative assessment of baseline performance. Figure 4 shows comparisons of the baseline methodologies for two test seasons.

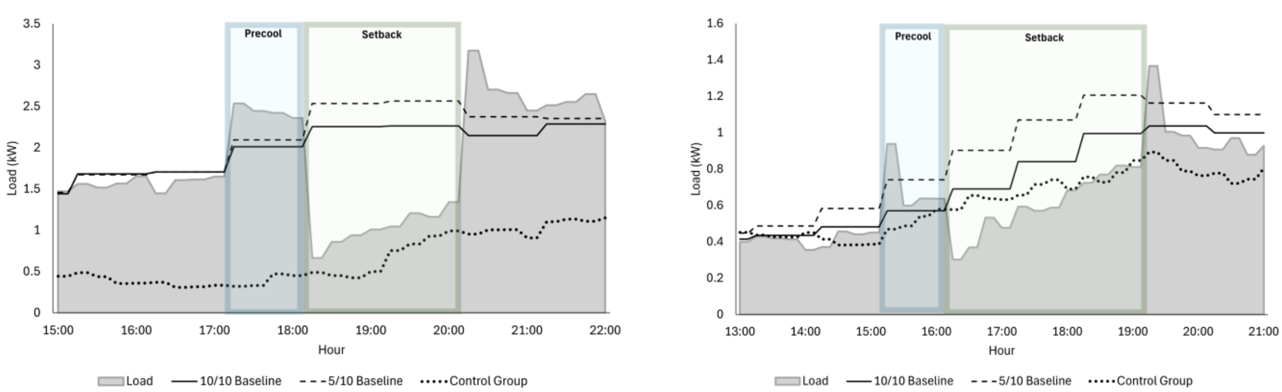

Figure 4. Brand 2 baseline comparisons for summer 2024 (left) and fall (right).

Outlier sites were identified and excluded from the dataset used in the load-impact calculations. All test days for a given site were excluded if the site load was regularly observed to have significantly higher than expected residential home load. Specific test days were also excluded for individual sites if there appeared to be EV charging activity of significant magnitude to impact either the test day load or baseline estimates, or if the calculated baseline was not reliable in the hours immediately preceding the control period.

Field Test Results

Device Response

Across all test periods, the thermostat aggregations were dispatched in 54 individual control events with a total of 126 hours of thermostat setback. Figure 5 shows the distribution of control-signal outcomes by status category for each test season and thermostat brand.

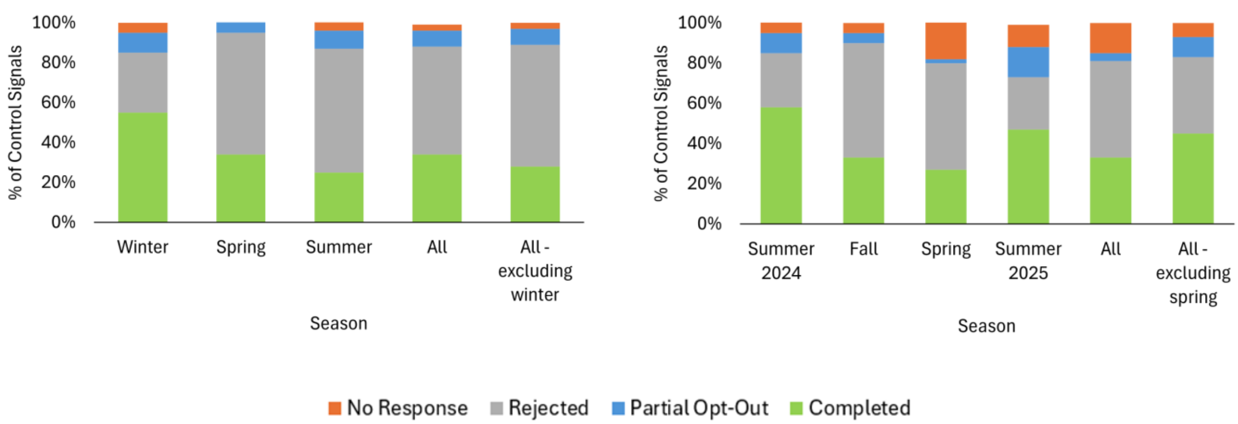

Figure 5. Device response status by test season for Brand 1 (left) and Brand 2 (right)

The highest rate of control signal rejections occurred during the Brand 2 spring 2025 testing period. This outcome was attributed to the novel control scheme implemented during that season, which required frequent hourly setpoint updates and likely contributed to participant notification fatigue and general dissatisfaction. Because no device reported a “completed” status for all control signals on any test day, the dataset for this period was deemed insufficient to produce reliable load-flexibility estimates and was therefore excluded from subsequent impact analyses. Over the remaining test periods, opt-out rates were generally consistent between the brands. Brand 1 participants having participated in a DR program prior to testing did not appear to influence response rates. No Response rates were generally lower for Brand 1, possibly due to the professional installation of the thermostats that was provided. Brand 1 exhibited higher rates of control-signal rejections, especially when excluding the spring testing period for Brand 2, and the winter test period for Brand 1, during which heating was controlled rather than cooling. One plausible explanation is that Brand 1 thermostats include a permanent setpoint-hold feature that allows customers to disable automated adjustments, thereby preventing the device from responding to control signals. This feature is not available on Brand 2 thermostats, which rejected control signals only when the thermostat was not in cooling mode. Lower average outdoor temperatures during the Brand 1 test periods may have contributed to this trend.

The thermostats consistently rejecting the control signals, in almost all cases, appeared to be manually controlled by users. This sub-set of participants were using the smart thermostats as traditional digital thermostats, resulting in behavioral usage patterns that were incompatible with dynamic retail price response and automated load-flexibility objectives. This pattern may have been influenced by the Brand 1 participants being provided with the smart thermostats just prior to testing, thus not allowing them time to become familiar with the available functionality.

A comparison of response patterns was made between households located in and outside of regions identified as Low Income (LI) designated areas revealed. No significant differences in LI and non-LI response rates were observed for Brand 2. For Brand 1, the number of non-LI participants was too small to support meaningful comparisons of the response rates.

Temperature and device response. Temperature was determined to be the primary factor influencing device response rates. As temperatures increased, control-signal rejection rates decreased, and completion rates increased while opt-out rates remained steady. Figure 6 shows the distribution of device-response response statuses for each device dispatch across temperature bands.

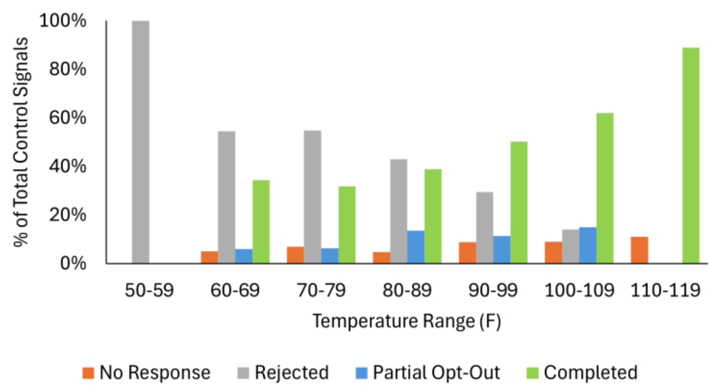

Figure 6. Response statuses by temperature bands from the Brand 2 test periods.

Precooling and device response. The Brand 2 summer test periods were designed to evaluate the influence of precooling on both load-flexibility capacity and participant opt-out rates with alternating test days with or without precooling. This design allows for comparison of with and without precooling across adjacent days which helps with keeping similar weather conditions. The number of thermostats reporting completed device-control status decreased on test days that included precooling, along with a small increase in opt-outs. The status distribution holds across temperature bins as shown in Figure 8. The vast majority of opt-outs occurred during the setback period, rather than the precool period. Additionally, participants tended to overcompensate slightly when opting out, adjusting their setpoints to about a degree beyond their pre-event setpoint. Figure 7 compares the response statuses for Brand 2 test days with and without precooling.

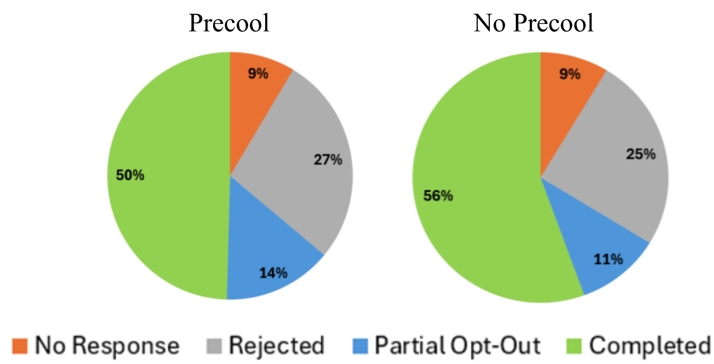

Figure 7. Response statuses for Brand 2 on days with (left) and without (right) precooling.

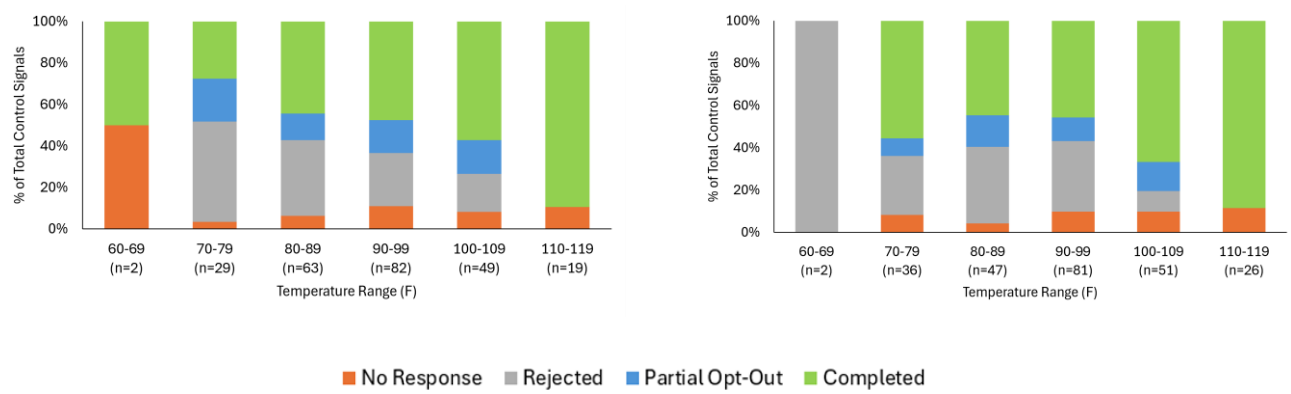

Figure 8. Response statuses by temperature bins for Brand 2 on days with (left) and without (right) precooling.

Energy and Cost Impacts

Similar to device status outcomes, there were no significant differences observed in the energy impact from households located in or outside of low-income (LI) designated areas.

Table 4 summarizes the energy and cost impacts for Brand 1 by seasonal test period. These metrics only include sites with a completed device control status, excluding outliers. Note that the winter test period utilized a heating setback with a setpoint decrease, while all other test periods utilized a cooling setback with a setpoint increase. The cost impacts are based on the seasonal price profiles developed by Berkeley Labs.

Table 4. Energy and cost impacts metrics for Brand 1 seasonal test periods.

Metric	Unit	Winter 2023	Spring 2023	Summer 2023
24-hr energy cost	$/day	$1.58	$2.04	$3.83
	% of baseline value *	51%	78%	59%
Weighted average 24-hr energy price	$/kWh	$0.21	$0.19	$0.29
	% of baseline value *	91%	87%	86%

High price period load reduction: 2-hr	kW	1.11	0.61	1.62
	% of baseline load *	95%	100%	96%
Load shed time window		6-8am Spring signal	5-7pm Spring signal	6-8pm Summer signal
Baseline load reflected (*)		* = % of AC or HP load		

Table 5 summarizes the energy and cost impacts for Brand 2 by seasonal test period. These metrics only include sites with a completed device control status, excluding outliers. Note that spring testing is excluded due to high control signal rejection rates.

Table 5. Energy and cost impacts metrics for Brand 2 seasonal test periods. *n refers to the number of test days within the test season.

Metric	Unit	Summer 2024 (10-in-10; n* = 12)			Fall (10-in-10; n* = 10)			Summer 2025 (10-in-10; n* = 12)		
		Week 1	Week 2	Total	Week 1	Week 2	Total	Pre-cool	No Pre-cool	Total
24-hr Energy Cost	$/day	7.43	7.50	7.47	4.92	4.16	4.48	12.83	10.49	11.57
	% of baseline energy cost per day	88%	81%	85%	87%	99%	93%	97.6%	97.8%	97.7%
Weighted Average 24-hr Energy Price	$/kWh	0.25	0.23	0.24	0.28	0.28	0.29	0.46	0.41	0.44
	% of baseline weighted average price	92%	87%	89%	97%	94%	96%	96.4%	98.0%	97.2%
Load Reduction during 3-hr High Price Period (4-7 PM)	kW	NA	NA	NA	0.43	0.12	0.24	0.77	0.60	0.68
	% of baseline load	NA	NA	NA	40%	17%	28%	42.4%	37.0%	39.7%
Load Reduction during 2-hr High Price Period (6-8 PM)	kW	1.20	1.26	1.23	NA	NA	NA	N/A	N/A	N/A
	% of baseline load	56.4%	52.7%	54.5%	NA	NA	NA	N/A	N/A	N/A

Temperature and event duration impact on energy. The load shed capacity estimated for residential smart thermostats was determined to be consistent across device brands when controlled for temperature. There was a strong positive correlation between load shed capacity and outdoor temperature, though the correlation was observed to reverse at temperatures above 110°F. The discontinuity in the trend of load shed capacity and temperature may likely result from increased HVAC cycling required earlier and more frequently in the setback period to maintain temperature, even with the increased setpoint. A weather matching baseline was applied to the sites on test days with these high temperatures in an attempt to account for this difference. This alternative baseline methodology did not result in a reliable counterfactual.

Figure 8 shows the load shed for Brand 1 and Brand 2 thermostats on all test days, as a function of the maximum daily outdoor temperature during all test periods, as well as a sample site's load on a test day with a maximum daily temperature of 113°F showing clear cycling mid-setback period.

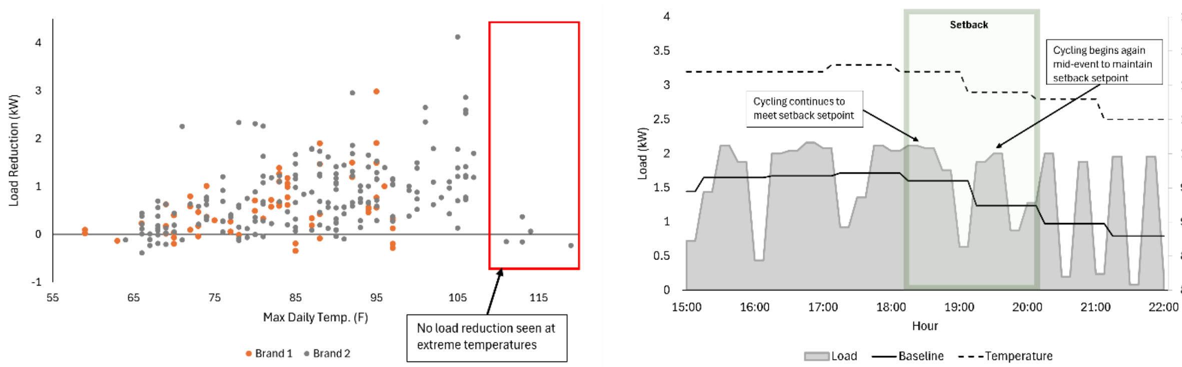

Figure 8. Average daily setback load reduction per site as a function of the local high temperature (left), and a sample site's load and 10-in-10 baseline from a high temperature test day.

The strong correlation between load reduction capacity and temperature holds when accounting for the test hour within the control period. The degradation of load impact as the control period persists increases with temperature. This is particularly apparent in the third hour of the control period. These results support the claim that temperature is the primary determinant of load reduction capacity. Figure 9 shows load reduction by temperature band and control period hour for all Brand 2 sites and for those sites with a completed device response status.

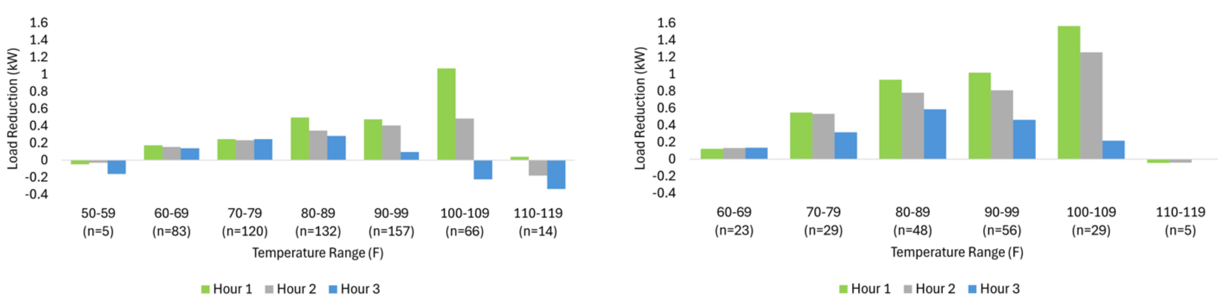

Figure 9. Average load reduction by control period hour and temperature band for all Brand 2 sites (left), and only non-outlier Brand 2 sites with a completed device response status (right).

Precooling and energy impact. When comparing the Brand 2 test days with and without precooling, minimal impact on average load reduction was observed. Though, test days with precool appeared to have lower degradation of load reduction in the later test hours, particularly in the third hour. This trend is most apparent when load impact for all sites is compared, rather than just sites with a completed device response status. Figure 10 shows the load reduction by control hour on days with and without precooling.

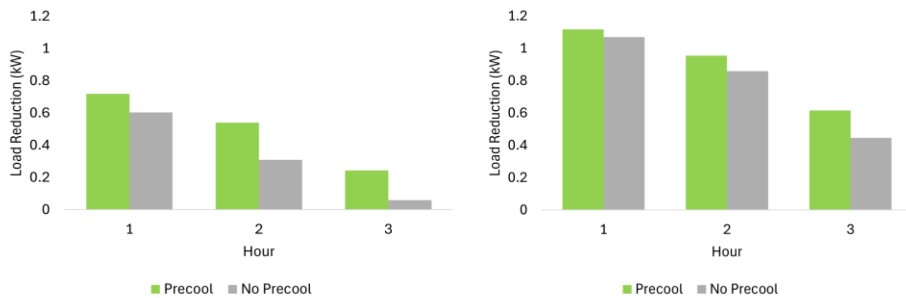

Figure 10. Load reduction on precool and non-precool test days for both Brand 2 summer test periods. Shown by control period hour with all sites included (left), and sites with completed device control (right).

Subject Interviews

Olivine collaborated with the CalFlexHub Task 2 research team from UC Davis to develop subject interview protocols. Olivine provided the UC Davis team with a cross-section of participants as interview candidates and notified participants that an outside team would contact them to request an interview. Contact information was provided for seven Brand 1 participants and ten Brand 2 participants. Candidates included households with unresponsive devices, households designated and not designated as DAC or LI, sites identified as outliers, and sites that demonstrated expected device responsiveness. The UC Davis team conducted outreach, scheduled interviews, and carried out discussions via video conference or telephone. Olivine was not present for the interviews.

The UC Davis research team conducted interviews with a total of nine participants. None of the Brand 1 participants interviewed had used a smart thermostat before and none were using setpoint scheduling features. One Brand 1 participant reported noticing lower energy bills after installation of the smart thermostat, while another reported routinely switching the thermostat to OFF mode when cooling was not desired, a method of operation that impedes dynamic retail price response. None of the Brand 1 interviewees reported seeing the in-app or on-device control period notifications. In contrast, Brand 2 participants reported seeing only in-app notifications (Sanguinetti et al. 2024).

Participants from both brands reported greater sensitivity to indoor temperatures when setback periods followed a precooling period (Sanguinetti et al. 2024). One participant noted a "larger than typical temperature swing" (Sanguinetti et al. 2024), even though setback setpoints remained constant on test days with and without precooling. This comment is consistent with observations from summer 2024 testing when more opt-outs occurred during week 1, which included precooling, than in week 2, which did not include precooling.

Conclusions and Discussion

Past studies have frequently claimed that precooling improves customer comfort, decreases opt-out rates and reduces the degradation of load reduction during the later hours of a control period. Olivine's findings do not support the first two claims and indicate that the third effect may be less pronounced than previously assumed. As discussed in the Subject Interviews section, participants reported that the larger temperature swing resulting from the transition from precool to setback periods had a negative effect on comfort, which is supported by the observation that opt-outs largely occurred during the setback period rather than the precool period. Slightly higher opt-out rates were also seen during test days with precool versus those without (see Figure 7). While precool test days did show improvements in load reduction in the third hour of the control period, the impact was not substantial when looking only at sites with a completed device control status (see Figure 10). This improvement was much more pronounced when considering all sites, suggesting that past studies may be overestimating the significance of precooling's impact by not considering device response data. These findings underscore the need for future research that explicitly controls for precooling to isolate its effects on both load flexibility capacity and on user comfort across a variety of conditions, ideally with detailed device-level response data to ensure accurate interpretation. Utilities and ASPs should exercise caution when implementing precool periods until sufficient evidence supports their effectiveness.

Past studies have frequently asserted the position of an hour within a control period is the dominant predictor of load impacts, outweighing other factors such as time of day or ambient temperature (see Figure 1). While our findings confirm that event-hour position influences load-shed performance, temperature was observed to be the primary determinant of load reduction performance as discussed in the Temperature and Event Duration Impact on Energy section. Past studies' control periods occurred in a limited temperature range (79-108°F) and therefore did not capture the pronounced effects observed for both lower and extreme high temperatures. In our results, load reduction erodes sharply at temperatures below 80°F or above 110°F compared to linear projections against temperature, and the differences in load impacts across control-period hours were substantially diminished under these conditions (see Figure 9). Additionally, device control completion rates decreased at lower temperatures, further reducing available load shed (see Figure 6). It would be inaccurate to assume that available load reduction capacity remains reasonably consistent across temperature ranges simply by controlling for event hour position. Future research should prioritize data collection across a broader temperature range, with particular emphasis on extreme heat conditions given that grid emergency events are typically triggered at high temperature thresholds. This would help to inform utilities of available thermostat load reduction during grid emergencies.

The load impact results in this study are comparable with those from similar reports, though lower when including sites that rejected control signals in the average results. This pattern persists when only comparing load reduction within the same temperature range and event hour position, as illustrated in Figure 1 and Figure 9. Setpoint change magnitude, baseline methodology, rejection rates, and opt-out rates likely contribute to the differences. The magnitude of setpoint changes during setback periods may have been higher during the control periods evaluated in past reports than those used in our testing, potentially contributing to the difference in load reduction across studies. However, this hypothesis cannot be verified without access to setpoint change data. Standard baseline methods (e.g., 5-in-10 or 10-in-10) do not consistently provide reliable counterfactual load profiles due to the inherent stochastic nature of residential home load profiles. Control group baselines, used in the majority of smart thermostat reports surveyed, are likely the more robust methodology, as they reduce reliance on historical load shapes that may deviate for reasons unrelated to thermostat control. However, ASPs cannot rely on control groups without a sufficiently large number of non-participating but comparable sites available to support accurate matching between participants and non-participants (see Figure 4). As a result, the accuracy of load reduction reporting currently achievable by ASPs is inherently limited.

Accurate impact measurement is critical to demonstrating value to participants and incentivizing development of price responsive device features. Program designers should consider robust baseline methodologies that do not rely on wide access to customer data, such as load forecasting or runtime data comparison, to help reliably expand smart thermostats as a load-flexible resource. Current utility programs and market incentives rarely cover both customer acquisition costs and ongoing OEM device fees. Broader adoption and promotion of open communication protocols such as OpenADR or Matter may offer a more cost-effective pathway for ASPS by reducing reliance on proprietary control channels.

Utility administered residential thermostat programs actively remove participants who consistently underperform or opt-out of control periods. This practice may partially account for the higher load reduction values reported. In addition, variations in incentive structures and program messaging - especially during grid emergencies - could have discouraged customers from opting out during utility-initiated events, contributing further to the difference in load reduction across studies. Without access to device response data, however, these hypotheses cannot be fully validated. We propose that systematic collection and reporting of device response data in future utility pilot and program evaluation would improve the interoperability and comparability of smart thermostat load flexibility research, particularly considering the significant impact opt-outs have on aggregate load reduction. For example, in our study we observed that Brand 1 participants had more control signal disruptions (see Figure 5). We found these participants tended to use their thermostats manually, interfering with device control signals. Thermostats configured with setpoint schedules were far less prone to control signal disruption. It is our determination that many of these issues can be attributed to a lack of familiarity with smart devices used as an energy resource and load flexibility in general. The substantial reduction in control signal completion rates during the Brand 2 spring testing season also highlights user behavior as a limiting factor to smart thermostat load response that merits further investigation if aggregators or utilities intend to implement more complex control schemes. Although more frequent setpoint changes that better align with hourly changes in price could hypothetically have increased cost savings if all control signals were completed, repeated control notifications in a short timeframe appear to have diminished users' tolerance for continued thermostat control. Future research could investigate differences in customer characteristics, familiarity with smart home devices, recruitment strategies and customer motivation, control-signal completion rates, and ultimately the predictability of residential thermostat load response.

Acknowledgements

This work was supported by the California Energy Commission's Electric Program Investment Charge. We are grateful to the CEC's Matt Fung for his support of the California Load Flexibility and Deployment Hub (CalFlexHub) program.

References

Sanguinetti, A., Alston-Stepnitz. E., Outcault, S., Taylor, M. 2024. *Load Flexibility: Keeping Users in the Loop with "Invisible" Technologies*. Berkeley: Lawrence Berkeley National Laboratory. calflexhub.lbl.gov/wp-content/uploads/sites/41/2024/08/Load-Flexibility-Keeping-Users-in-the-Loop-with-Invisible-Technologies.pdf

Bode, J., Gauther, P., Karlovic, P., and Gordon, H. 2024. *2022-2023 Load Impact Evaluation of Pacific Gas and Electric's Smart Thermostat Control Pilot*. San Francisco: California Measurement Advisory Council. www.calmac.org/%5C/publications/PY2022-2023_Smart_Thermostat_Control_Pilot_Evaluation_Report_Final_%28003%29.pdf

EIA (Energy Information Administration). 2024. *2020 RECS Survey Data.* Washington, DC: EIA. U.S. Energy Information Administration - EIA - Independent Statistics and Analysis

Farr, D., Deych, D., and Smith. 2025. *PY2024 SCE Smart Energy Program Load Impact Evaluation.* San Francisco: California Measurement Advisory Council. www.calmac.org/publications/PY2024_SCE_DR_Program_Report_SEP_FINAL_PUBLIC.pdf

Jiang, G., Smith, J., and Ramee, C. 2021. *2021 Process Evaluation of San Diego Gas and Electric's Demand Response Programs.* San Francisco: California Measurement Advisory Council. www.calmac.org/%5C/publications/SDG&E_Process_Evaluation_-_Full_Report_-_Final.pdf

Sarran, L., Gunay, H., O'Brien, W., Hviid, C., Rode, C. 2021. *"A Data-Driven Study of Thermostat Overrides During Demand Response Events."* Energy Policy 153 (1): 2-11.

Smith, J. and Totten. 2020. *Smart Energy Program Load Impact Evaluation.* San Francisco: California Measurement Advisory Council. www.calmac.org/publications/SCE0442.01_PY2019_SCE_DR_Program_Report_SEP_Public.pdf

—. 2021. *2020 Smart Energy Program Load Impact Evaluation.* San Francisco: California Measurement Advisory Council. www.calmac.org/publications/SCE0451.01_PY2020_SCE_DR_Program_Report_SEP_Public.pdf

Smith, J., and Farr, D. 2022. *2021 Smart Energy Program Load Impact Evaluation.* San Francisco: California Measurement Advisory Council. www.calmac.org/%5C/publications/SCE0463.01_PY2021_SCE_DR_Program_Report_SEP_Public.pdf

Walton, R. 2022. *Slow Adoption of Smart Thermostats in the US Misses Big Potential Energy Savings: S&P.* Washington, DC: Utility Dive.