Behavior-Aware Pre-Event Customer Selection for Residential Thermostat Demand Response

Lingdong Zhou, Jose E. Aguilar Escamilla, Huazheng Wang, and Xiangqi Zhu
School of Electrical Engineering and Computer Science
Oregon State University, Corvallis, OR, USA

***Abstract*—Thermostat demand response programs aim to reduce air-conditioning use, but some customers opt out before an event ends. This paper examines how behavior history can help aggregators select customers before each event. Predicted runtime reduction is weighted by completion probability, estimated from participation history alone or combined with event and household information. From ecobee's summer 2024 Donate Your Data records, we reconstruct 58,194 sessions and replay selections for 10,912 of them in later events. At 10% selection, history weighting nearly halved early exits, from about 20 to 11 per 100 selected sessions, relative to ranking by predicted runtime reduction alone, with no significant difference in average runtime reduction. The gain held in cross-validation and at every tested fraction. The combined model predicted early exits better but selected no better lists. We thus show that two counts per thermostat, completions and early exits, make call lists markedly more reliable.**

***Index Terms*—Demand response, direct load control, smart thermostats, customer selection, early-exit prediction, residential HVAC.**

I. Introduction

Aggregators use smart thermostats, water heaters, electric vehicles, and other residential loads to shift electricity consumption for demand response (DR) and virtual power plants [1], [2]. Cooling is especially useful during summer peaks because building thermal inertia slows changes in indoor temperature, allowing compressors to run less [3]–[5].

Before each event, an aggregator selects part of its thermostat fleet to reduce compressor runtime. Aggregate models estimate how much load a group of devices can shift and when that capacity is available [3], [6]. Individual customers, however, can override control before the scheduled end, even when large runtime reductions are forecast. Such overrides reduce savings, recur within the same customers, and vary with household, weather, and event conditions [7]–[10].

Khorasani Zadeh *et al.* predicted individual overrides from prior behavior, comfort preferences, and event conditions [11]. Chen, Nie, and Li selected customers by combining potential response with completion probability and tested this objective in simulations [12]. These simulation results raise two questions for recorded thermostat outcomes: does the rule reduce early exits while keeping response high, and how do different models for completion probability change the selected lists?

This paper examines how behavior history can help aggregators select customers before each event. The paper's contributions are threefold: (i) we test completion-weighted selection [12] (Fig. 1) on recorded rather than simulated outcomes; (ii) we show that history-based weighting nearly halved early exits in runtime-ranked lists (20.0% to 10.6% at 10% selection) without significantly changing runtime reduction; the gain held in cross-validation and across list sizes; and (iii) we show that counts of past completions and early exits suffice, as a richer model with event and household information predicted exits better but never selected significantly more reliable lists.

Section II describes the data and measurements. Section III presents prediction, selection, and validation methods. Section IV reports selection and prediction results, Section V discusses their implications, limitations, and future work, and Section VI concludes the paper.

II. Data and Measurements

A. Thermostat Sessions and Early Exit

We use de-identified five-minute thermostat telemetry and building/HVAC metadata from ecobee's Donate Your Data program, June 3–September 27, 2024 [13], [14]. Reported cities link thermostats to hourly outdoor temperature from the Open-Meteo archive [15].

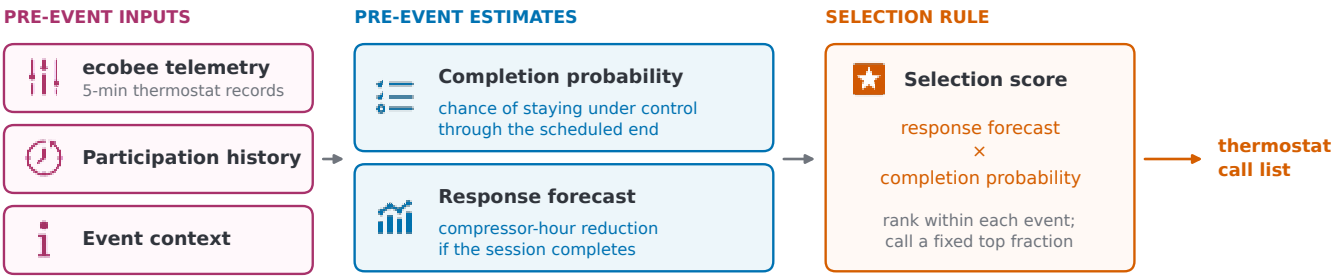


Fig. 1. Workflow for pre-event customer selection. Predicted response under completed control is weighted by completion probability to rank customers within each event and select a fixed top fraction.

Free-text DR labels give no event IDs or schedules. A session represents one thermostat's participation in an event, reconstructed from at least three consecutive DR rows. Sessions are grouped by state/province (jurisdiction), labeled DR program type, and start time rounded to 15 min; events require at least 20 sessions. The most common session end, rounded to five minutes, estimates the event end. Following Sarran *et al.* [7], a session ending at least 15 min before this estimate is an *early exit*; others are completions. Exit rates at 10, 20, and 30 min thresholds are 18.12%, 17.24%, and 16.27%, respectively.

Reconstruction yields 58,194 sessions from 9,896 thermostats in 236 events, including thermostats without history; 17.69% end early.

B. Compressor-Runtime Response

Response compares compressor runtime with an estimated no-DR baseline. Compressor duty is the fraction of a window with the compressor running. Candidate days fall in the preceding 90 days, match the event day's weekday/weekend type, and exclude DR days and US or Canadian federal holidays. Eligibility requires at least 80% telemetry coverage and no DR rows in three windows: the event window, the two pre-event hours, and an adjustment period four to two hours before the event. Outdoor temperature in the two pre-event hours also needs 80% coverage. With at least five eligible days, we choose the five closest in mean temperature in those two hours and average their duty over the same local clock window. To adjust for event-day conditions, we cap the adjustment-period duty difference (event day minus matched days) at ± 0.20 and add it to the baseline average, then bound adjusted duty to $[0, 1]$.

For session i, let $\bar{c}_i$ and c_i denote baseline and observed mean compressor duty over the estimated event window, and T_i its duration in hours. Signed response is

$$r_i = (\bar{c}_i - c_i)T_i. \tag{1}$$

One compressor-hour means one hour less runtime than baseline; negative values mean more. Evaluation retains both signs over the full estimated window, including after early exit, without assessing power, comfort, or rebound.

To check the baseline, we estimate duty on each session's most recent eligible non-DR day using only earlier days. Duty-fraction mean absolute error is 0.160 with bias -0.016, versus 0.179 and -0.099 using the two pre-window hours as the baseline. Non-DR days limit earlier-event contamination, but baseline error remains a source of uncertainty [16], [17].

Exit prediction uses all reconstructed sessions. Response measurement and cross-validation replay require cooling at onset, pre-event telemetry, at least 80% event-window coverage, and a matched-day baseline. The later selection replay uses response-eligible sessions from the later exit-prediction test (Table I). The response model is trained and checked on completed sessions; replay compares hypothetical lists with recorded outcomes, including early exiters.

III. Methods

Chen *et al.*'s objective weights customer-reported load reduction by completion probability under a payment budget [12].

TABLE I
Analysis cohorts. Indented rows are completers in the preceding cohort.

Analysis	Sessions	Thermostats	Events
Early-exit prediction	58,194	9,896	236
Response measurement/replay	35,887	7,029	223
completers	28,991	6,581	223
Later exit-prediction test	17,621	7,190	71
Later selection replay	10,912	4,667	70
completers	9,038	4,194	70

With equal costs and a fixed list size, we replace reported reduction with forecast runtime reduction under completed control, weighted by completion probability:

$$s_i = \widehat{r}_i^{\,\text{complete}}(1 - \widehat{q}_i), \tag{2}$$

where $\widehat{r}_i^{\,\text{complete}}$ forecasts response under completed control and $\widehat{q}_i$ estimates early-exit probability. Candidates are ranked by s_i within each event. With C indicating completion, R signed response, and z pre-event information, the identity $\mathbb{E}[CR \mid z] = P(C = 1 \mid z)\mathbb{E}[R \mid C = 1, z]$ motivates the score. The target CR credits response only upon completion, giving early exits zero. Their pre-event scores can be positive because outcomes are unknown. Replay credits full-window signed response, including early-exiter reductions.

A. Predicting Early Exit

Four models predict exits; the two using history provide completion probabilities for selection. The population model assigns the training-set exit rate. A context-only Light Gradient Boosting Machine (LightGBM) model uses calendar and time variables, location, building/HVAC attributes, occupancy metadata, and pre-event indoor-temperature and compressor-runtime summaries. Exit models exclude outdoor weather and event duration. Reconstructed duration uses the session end times defining early-exit labels, so including it would leak outcome information. Response forecasts use it as a schedule substitute shared across rankings. Weather enters baseline matching; weather forecasts are not evaluated as exit predictors.

The beta-binomial empirical-Bayes (EB) model [18] estimates exit probability from x_i early exits in n_i prior thermostat sessions:

$$\widehat{q}_i^{\,\text{EB}} = \frac{x_i + \alpha}{n_i + \alpha + \beta}. \tag{3}$$

Each training fold fits prior counts α (early exits) and β (completions) by maximum likelihood. Without history, the estimate is the fleet mean $\alpha/(\alpha + \beta)$; individual outcomes gain weight as history grows. Counting past sessions equally assumes they remain informative about the next event.

A context+history LightGBM classifier adds prior-session count, EB and raw exit rates, the most recent outcome, and the length of the trailing run of consecutive early exits in the available history to the contextual variables [19]. All processing and fitting use only training-fold data. Because Eq. (2) uses probability values as well as rankings, nested Platt scaling calibrates the LightGBM probabilities [20], [21].

TABLE II
REPLAY OUTCOMES AT A 10% SELECTION FRACTION. COMP.-H/SESSION IS COMPRESSOR-HOURS PER SELECTED SESSION OVER THE ESTIMATED EVENT WINDOW.

Validation	Selection rule	Early exit	Comp.-h/session
Event-held-out CV	Uniform random	19.18%	0.499
	Response only	23.93%	1.416
	Response × EB history	14.62%	1.432
	Response × context+history	14.92%	1.432
Later-event test	Uniform random	17.18%	0.515
	Response only	20.00%	1.418
	Response × EB history	10.58%	1.407
	Response × context+history	11.73%	1.408

Average precision (AP) measures how well early exits rank near the top, with the exit rate as the random-ranking reference [22]. Area under the receiver operating characteristic curve (AUC) is the probability that an early exit ranks above a completion, with half credit for ties (0.5 for random ranking). The Brier score is mean squared probability error against the 0/1 outcome; lower is better. AP comparisons use 1,000 paired bootstrap replicates jointly resampling thermostats and events.

B. Forecasting Response Conditional on Completion

The LightGBM regressor learns compressor-duty reduction from completed sessions using the matched-day baseline, recent compressor runtime and coverage, indoor temperature, event timing, location, building/HVAC attributes, and occupancy metadata. Replay uses reconstructed duration in place of the missing announced schedule to define baseline windows and convert predicted duty reduction to compressor-hours, $\widehat{r}_i^{\text{complete}}$; performance with announced schedules remains untested.

Forecast checks on held-out completed sessions use mean absolute error (MAE), Spearman correlation, and measured response among each event's top 10% by forecast. The benchmark is training-fold mean duty reduction times estimated duration.

C. Event-Level Selection

At fraction f, each rule selects $k_e = \lceil f n_e \rceil$ of event e's n_e response-eligible sessions (Table I). Four rules comprise uniform random selection and ranking by $\widehat{r}_i^{\text{complete}}$ alone or weighted by EB or context+history completion probability. The random benchmark weights each outcome by selection probability k_e/n_e, averaging over all possible lists.

The top k_e maximize total score at fixed list size, without guaranteeing the largest realized response. We compare selected early-exit rate and mean signed response per session at $f = 10\%$ (primary) and $f \in \{2, 5, 20, 30, 50\}\%$ (sensitivity).

D. Validation Designs

The chronological *later-event test* is primary. Models and history use only data before August 15, 2024, 19:00 UTC and remain fixed. Exit training uses 40,464 sessions in 163 events; response training uses 19,902 completions in 151 response-eligible events. Two cutoff-crossing events are excluded. Replay eligibility is retrospective, requiring full-window response data.

In five-fold *event-held-out cross-validation*, each event enters one test fold: 236 events for exit prediction and 223 with measurable response for selection replay. Each training fold refits models, calibration, and the EB prior; test history uses only earlier training-fold outcomes from the same thermostat.

Selection comparisons use 2,000 paired bootstrap replicates that jointly resample thermostats and events with the same weights across rules [23]. Sensitivity intervals are descriptive, without multiple-comparison adjustment.

E. Robustness Checks

Robustness checks address temporal stability, inferred boundaries, and geographic concentration. Selection outcomes are compared in two chronological halves of 35 later events, with EB calibration comparing mean probabilities and observed exit rates across all exit-prediction sessions in each half.

Boundary sensitivity uses events where at least 70% of sessions share a rounded end time, reflecting agreement within the records. Geographic concentration is $n_{\text{eff}} = (\sum_j n_j)^2 / \sum_j n_j^2$, where n_j counts selected sessions in jurisdiction j. The metric gives the number of equally sized jurisdictions with the same concentration. Counts match across rules because list sizes match within events. Omitting each jurisdiction checks its influence.

Mean scores are compared with mean CR on each weighted rule's 10% list to check score accuracy, crediting early exits with zero response.

IV. RESULTS

A. Selection Outcomes with Completion Weighting

At 10% in the 70-event later selection replay (Table II), random selection had expected outcomes of 17.2% early exits and 0.515 compressor-hours per selected session. Response-only selection averaged 1.418 h per selected session, with 225 of 1,125 sessions ending early (20.0%). Its early-exit rate exceeded random selection by 2.8 percentage points (95% confidence interval (CI) -0.6 to 6.7), a nonsignificant difference.

EB weighting of the same forecast reduced early exits to 119 of 1,125 sessions (10.58%), 106 fewer than response-only selection. The paired change was -9.42 percentage points (95% CI -13.63 to -5.73). Mean response was 1.407 h versus

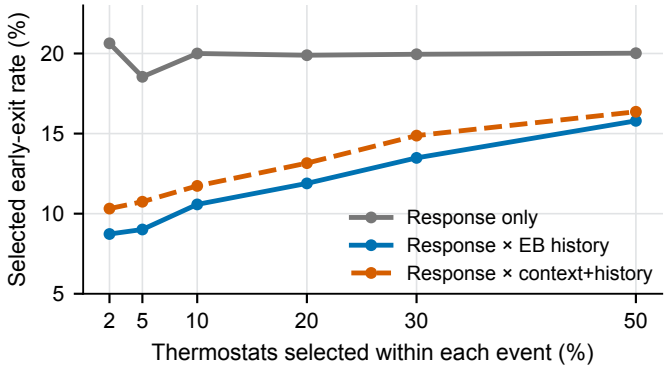


Fig. 2. Replay early-exit rates in 70 later events.

1.418 h, a change of −0.011 (95% CI −0.074 to 0.056). The difference was nonsignificant, but its lower confidence limit corresponds to a 0.074 h loss, or 5.2% of the response-only mean.

In cross-validation, EB weighting changed early-exit rate relative to response-only selection by −9.31 points (95% CI −11.82 to −6.94) and response by +0.016 h (95% CI −0.021 to 0.054). Response-only selection significantly exceeded the random benchmark's early-exit rate by 4.7 points (95% CI 2.3 to 7.2).

EB weighting reduced early-exit rate relative to response-only selection at every later replay fraction (Fig. 2), by 11.9 points at 2% and 4.2 at 50%. Response differences were −0.014 to +0.049 h per session; all 95% intervals included zero.

At 10%, context+history weighting exceeded EB's early-exit rate by 1.16 points in the later selection replay (95% CI −1.05 to 3.48) and +0.30 in cross-validation (95% CI −0.68 to 1.22); both differences were nonsignificant. At 20% and 30% in the later replay, context+history exceeded EB by 1.27 and 1.39 points, with unadjusted 95% intervals of 0.01 to 2.59 and 0.36 to 2.45, respectively.

EB weighting reduced early exits versus response-only selection in both 35-event halves (−9.9 and −9.1 points) and high end-time agreement events (−8.5 points in 59 later events; −8.9 in 198 cross-validation events). Early-exit intervals excluded zero; response intervals included zero. High agreement favors fewer exits; relative reductions in this subset were 47% in later replay and 39% in cross-validation, close to full-sample values.

Effective jurisdiction counts were 6.5 in later replay and 4.9 in cross-validation. Omitting any jurisdiction changed the early-exit difference between EB and response-only selection by less than one percentage point.

B. Past Outcomes as a Pre-Event Warning Signal

Next-session exit rates were 12.02% after completion and 44.21% after early exit (risk ratio 3.68, 95% CI 3.35–4.01). After two completions, the rate was 9.86%, versus 57.54% after two exits (risk ratio 5.84, 95% CI 5.15–6.54; Fig. 3). For mixed outcomes, rates were 31.65% when the more recent outcome was an exit and 24.79% when it was a completion.

In cross-validation (Table III), context and history each improve on the fleet-rate reference; their combination has the highest AP and AUC and lowest Brier score. Fleet-rate AUC is below 0.5 because predictions vary across folds. Context+history's AP gain over EB was 0.037 (using unrounded values; 95% CI 0.009–0.059).

TABLE III
EARLY-EXIT PREDICTION IN EVENT-HELD-OUT CROSS-VALIDATION (58,194 SESSIONS, 236 EVENTS).

Pre-event information	AP	AUC	Brier
Population (fleet rate)	0.172	0.487	0.146
Context only	0.303	0.673	0.138
Participation history (EB)	0.398	0.716	0.128
Context + history	0.434	0.744	0.125

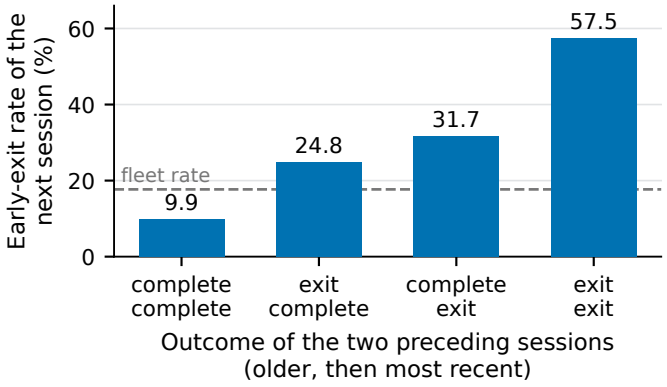


Fig. 3. Next-session early-exit rate by the two preceding outcomes, older first, over the 41,091 sessions with at least two prior sessions; the dashed line is the 17.69% rate over all 58,194 sessions.

Context+history's cross-validation calibration slope was 1.02 (ideal: one). Expected calibration error, measuring the gap between predicted and observed exit rates, was 0.005 (ideal: zero). Later exit-prediction AP and AUC were 0.391 and 0.744. Mean EB probabilities exceeded observed exit rates by 3.7 and 3.9 percentage points in the two halves of the 70 replay events.

C. Response Forecast and Score Accuracy

On 9,038 later completed sessions, forecast MAE was 0.387 h versus the benchmark's 0.634 h, with Spearman correlation 0.760. Among these completers, each event's top 10% by forecast averaged 1.550 h of measured response, versus 0.615 h for all completers. Cross-validation MAE was 0.446 h with Spearman correlation 0.733.

On each rule's 10% list, mean score error (predicted score minus observed CR) was +0.003 h for EB and −0.015 h for context+history in cross-validation, and −0.033 h and −0.008 h, respectively, in the later selection replay.

V. DISCUSSION

A. Operational Meaning

Avoided early exits preserve control the aggregator would otherwise lose: in later replay, early exiters left a median 95 min, or 49% of the event window, uncontrolled.

EB weighting needs only a thermostat identifier, two counts, and a response forecast. Our later-event test kept counts fixed at the cutoff; deployment could update them after each event.

B. Prediction Accuracy and Decision Quality

Building on thermostat override prediction [11], we examine how risk estimates affect selection. Simple EB weighting was never significantly outperformed by context+history at any tested fraction from 2% to 50% in either design, despite its lower AP (Section IV-A). AP evaluates risk ranking across all sessions, while selection ranks each event's candidates by response times completion probability. Better risk prediction therefore need not yield more reliable lists.

Chen *et al.*'s online-learning simulations favored context over past participation counts [12]. Our models differ in history

features and type, so comparing them using recorded outcomes cannot isolate context's value. Weather forecasts and announced duration remain untested as exit predictors. EB ignores outcome order. The stronger association between recent outcomes and the next exit (Fig. 3) suggests testing greater weight on recent outcomes.

EB overestimated early-exit risk in both later-event halves, while completion weighting still reduced selected early exits. Historical probabilities may therefore need recalibration; the cause of the bias remains unresolved.

C. Scope, Limitations, and Next Steps

Equipment-level power measurements are needed to evaluate kWh or MW for market settlement; indoor temperature, comfort, and rebound require separate evaluation. Event ends are inferred rather than taken from program schedules, and the forecast and selection rule remain untested in a live program with announced schedules. City-based weather matching can link thermostats to the wrong location; its effect on baseline and selection remains unquantified.

The data cover one summer of ecobee Donate Your Data cooling events. Without an invitation roster, replay covers participants with measurable response in the observed jurisdictions and DR types, leaving full-fleet selection untested. Other seasons, platforms, program designs, and populations need evaluation. Consecutive-day calls could raise early exits, but the later replay has no sessions on a third consecutive day to assess repeated-event fatigue.

Predicting full-window response for completed and early-exit sessions could credit reductions from both, allowing selection to balance response targets, shortfall costs, and continued control. DR-Gym can support tests across repeated events [24].

VI. Conclusion

In this single-season replay, weighting response forecasts by completion probability nearly halved early exits among selected sessions. At 10% selection, EB weighting lowered the rate from 20.0% under response-only selection to 10.6% without a significant change in response. The reduction held at every tested fraction and in cross-validation. Context+history predicted early exits more accurately, yet never significantly outperformed EB in selection. Adding only completion and early-exit counts to response forecasts can therefore help aggregators build markedly more reliable call lists.

Acknowledgment

We thank ecobee and its customers for supporting this research.

References

[1] U.S. Department of Energy, "Pathways to Commercial Liftoff: Virtual Power Plants," U.S. Department of Energy, Washington, DC, USA, Tech. Rep., 2023.

[2] H. Li, Z. Wang, T. Hong, and M. A. Piette, "Energy Flexibility of Residential Buildings: A Systematic Review of Characterization and Quantification Methods and Applications," *Advances in Applied Energy*, vol. 3, p. 100054, 2021.

[3] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Aggregate Flexibility of Thermostatically Controlled Loads," *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 189–198, 2015.

[4] W. Zhang, J. Lian, C.-Y. Chang, and K. Kalsi, "Aggregated Modeling and Control of Air Conditioning Loads for Demand Response," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4655–4664, 2013.

[5] N. Lu, "An Evaluation of the HVAC Load Potential for Providing Load Balancing Service," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1263–1270, 2012.

[6] H.-G. Kwag and J.-O. Kim, "Reliability Modeling of Demand Response Considering Uncertainty of Customer Behavior," *Applied Energy*, vol. 122, pp. 24–33, 2014.

[7] L. Sarran, H. B. Gunay, W. O'Brien *et al.*, "A data-driven study of thermostat overrides during demand response events," *Energy Policy*, vol. 153, p. 112290, 2021.

[8] V. Tomat, M. Vellei, A. P. Ramallo-González *et al.*, "Understanding patterns of thermostat overrides after demand response events," *Energy and Buildings*, vol. 271, p. 112312, 2022.

[9] O. Olawale, J. Reyna, and B. Gilbert, "Factors Affecting Override Behavior during Demand Flexibility from High-Resolution Smart Thermostat Data," in *2022 ACEEE Summer Study on Energy Efficiency in Buildings*, Monterey, CA, USA, 2022, nREL/CP-5500-82442.

[10] P. J. Wildstein, M. T. Craig, and P. Vaishnav, "Participant Overrides Can Halve the Reliability Value of Direct Load Control Programs," *Energy and Buildings*, vol. 299, p. 113606, 2023.

[11] Z. Khorasani Zadeh, M. Ouf, H. B. Gunay *et al.*, "Development of Prediction Models for Thermostat Override Behavior in Direct Load Control Events," *Energy and Buildings*, vol. 301, p. 113707, 2023.

[12] X. Chen, Y. Nie, and N. Li, "Online Residential Demand Response via Contextual Multi-Armed Bandits," *IEEE Control Systems Letters*, vol. 5, no. 2, pp. 433–438, 2021.

[13] ecobee Technologies ULC, "Donate Your Data," [Online]. Available: https://www.ecobee.com/en-us/donate-your-data/, n.d., accessed: Jan. 21, 2026.

[14] H. Li, W. O'Brien, V. Loftness *et al.*, "A critical review of use cases and insights from a large dataset of smart thermostats," *Advances in Applied Energy*, vol. 19, p. 100236, 2025.

[15] Open-Meteo, "Historical Weather API," [Online]. Available: https://open-meteo.com/en/docs/historical-weather-api, n.d., accessed: Sep. 30, 2026.

[16] A. Todd, P. Cappers, C. A. Spurlock, and L. Jin, "Spillover as a cause of bias in baseline evaluation methods for demand response programs," *Applied Energy*, vol. 250, pp. 130–142, 2019.

[17] M. L. Goldberg and G. K. Agnew, "Measurement and Verification for Demand Response," National Forum on the National Action Plan on Demand Response, Tech. Rep., 2013.

[18] C. N. Morris, "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, vol. 78, no. 381, pp. 47–55, 1983.

[19] G. Ke, Q. Meng, T. Finley *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Adv. Neural Inf. Process. Syst. 30*, 2017, pp. 3146–3154.

[20] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers*, A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. MIT Press, 1999, pp. 61–74.

[21] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities with Supervised Learning," in *Proceedings of the 22nd International Conference on Machine Learning*. ACM, 2005, pp. 625–632.

[22] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLOS ONE*, vol. 10, no. 3, p. e0118432, 2015.

[23] A. B. Owen, "The Pigeonhole Bootstrap," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 386–411, 2007.

[24] J. E. A. Escamilla, L. Zhou, X. Zhu, and H. Wang, "Towards affordable energy: A gymnasium environment for electric utility demand-response program," *Reinforcement Learning Journal*, vol. 7, 2026.