sustainability

MDPI

Article

Reliable Interpretation of Extreme-Heat Energy Response for Sustainable Campus Management: Baseline, Calendar, and Chilled-Water Metering Controls

Cenk Cihangir [1,*], Lale Özyılmaz [2] and Yusuf Coşkun [3]

1 Department of Industrial Design Engineering, Istanbul Health and Technology University, 34445 Istanbul, Türkiye
2 Department of Avionics, Faculty of Applied Sciences, Istanbul Gelişim University, 34310 Istanbul, Türkiye; lozyilmaz@gelisim.edu.tr
3 Department of Mechatronics Engineering, Istanbul Health and Technology University, 34445 Istanbul, Türkiye; yusuf.coskun@istun.edu.tr
* Correspondence: cenk.cihangir@istun.edu.tr

Abstract

Open hourly meter data are often used to infer how buildings respond to extreme heat, but the interpretation can depend on the analytical baseline, institutional calendar, and meter configuration. We examined 60 education and office buildings in the Building Data Genome Project 2 using repeated whole-week validation, temperature-matched placebo tests, 13 academic-calendar definitions, and same-building, common-hour carrier comparisons. A leakage-resistant change-point model that incorporated a previous-day, load-derived operating-regime proxy reduced blocked-validation error relative to a conventional model in 58 of 60 buildings; temperature binning provided no additional predictive benefit after regime control. Recovery-related signals were uncommon: rebound was significant in 2 of 49 eligible buildings, load-shape disruption in 1 of 49, and the 72 h duration diagnostic in 3 of 60. At Fox, the BASE calendar supported term and break comparisons through 39–42 °C, but neither building-use group differed significantly after false-discovery-rate adjustment; estimates above 42 °C were inconclusive. In the matched sample of 13 education and 10 office buildings, no electricity-only, chilled-water-only, or weighted specification provided robust evidence of a between-group difference. Across 99 combinations of carrier weight, normalization, and temperature bin, every 95% interval included zero. Reliable campus benchmarking therefore requires out-of-sample baseline assessment, explicit calendar sensitivity, common-hour meter matching, and carrier-specific uncertainty reporting before building-use rankings or resilience-related interpretations are made.

Keywords: sustainable campus energy management; building energy performance; extreme heat; district cooling; chilled-water metering; operating-regime proxy; academic calendar; measurement and verification

Check for updates

Academic Editor: Antonio Caggiano

Received: 18 July 2026
Revised: 30 July 2026
Accepted: 4 August 2026
Published: 12 August 2026

1. Introduction

Buildings and construction are central to climate mitigation and to the management of heat-related energy demand. The latest Global Status Report for Buildings and Construction estimates that the sector accounts for 28% of global energy consumption and 37% of global CO_2 emissions; operational emissions reached 9.9 Gt CO_2 in 2024 [1]. During the same year, electricity use in buildings rose by more than 600 TWh, or 5%, and accounted for almost 60% of the increase in global electricity demand. Increased air-conditioning use during severe heatwaves was one of the main drivers [2]. These trends make the careful interpretation of measured building loads under hot conditions important for benchmarking, campus operations, and heat-event assessment.

Hotter summers increase cooling electricity demand, and operational meter data allow this effect to be examined across large building samples. Urban overheating has been associated with higher cooling demand and a smaller reduction in heating demand [3]. Reviews document the scale of this burden across the building stock [4], while recent outage records show that heatwaves can also reduce electricity-system reliability [5]. At building level, concurrent heat and power interruptions create practical trade-offs between energy efficiency and resilience that affect operational planning [6].

Sustainable building performance in hot climates also depends on the design and control of façades and daylighting systems. Recent multi-objective studies of corrugated, perforated, and separated shading systems have examined trade-offs among useful daylight, glare, illuminance uniformity, and external view in deep-plan and very-hot-climate offices [7,8]. These studies address the physical design of high-performance buildings. The present study examines the following complementary measurement question: whether open meter data support reliable conclusions about heat response when the baseline, academic calendar, and energy-carrier coverage are considered together.

University campuses are a useful setting for this question. They contain several building types within a single estate, are often metered at building level, and commonly use district cooling. In such systems, a central plant produces chilled water and distributes it through a pipe network instead of relying on a separate electric chiller in each building [9–11]. The Building Data Genome Project 2 (BDG2) provides an open portfolio for this purpose, with hourly electricity, chilled-water, and steam records, together with weather data and building metadata [12].

District cooling affects both how a campus is cooled and how its energy use is recorded. The thermal energy drawn for cooling is measured at the chilled-water interface and is separate from the building's electricity consumption. This distinction is important when heat response is interpreted. A district-cooled building may appear only weakly temperature-sensitive on its electricity meter because much of its cooling demand is recorded on another meter.

The availability of these data has broadened the use of metered consumption from estimating total energy use to examining building response and recovery during heat events. Metered consumption has been used to estimate institutional heatwave loads and peak demand [13], infer cooling and HVAC operation from whole-building electricity and outdoor temperature [14,15], and derive indicators such as peak demand, load factor, and ramp rate from smart-meter records [16]. Some studies describe these quantities in terms of resilience or recovery.

Metered load is nevertheless an indirect indicator of building behavior. Electricity demand varies with weather, but also with the baseline used to define normal operation, the building's operating schedule, and the way energy is supplied and metered. Although each factor is recognized in the literature, they are seldom examined together within one operational dataset. An apparent difference between building types, or an apparent absence of difference, may therefore reflect the analytical baseline, calendar, or meter architecture rather than the buildings themselves.

This study examines this interpretive risk through four linked checks. Repeated whole-week validation and a four-model ablation identify the predictive baseline. Temperature-matched placebo tests assess whether apparent event effects are specific to the heat episode. Academic-calendar stratification is evaluated under 13 prespecified definitions, with explicit checks of coefficient estimability and data support. The final check uses the same buildings and common valid hours to separate electricity and chilled-water responses and to test alternative carrier weights and normalization methods. Stability is quantified using 10,000 common building-level resamples, effect sizes, and multiplicity control. Figure 1 summarizes the workflow.

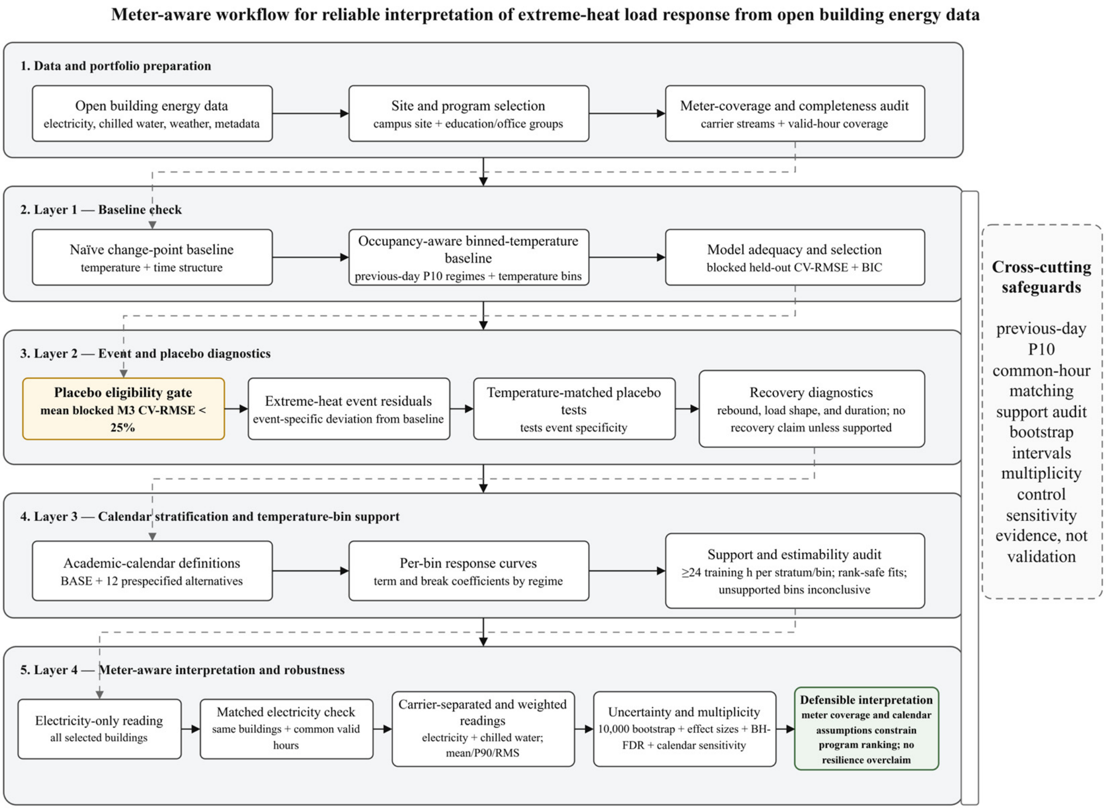

Figure 1. Workflow for the reliable interpretation of extreme-heat load response from open building energy data. The analysis begins with portfolio selection and meter-coverage screening, followed by baseline comparison, event and placebo diagnostics, academic-calendar and temperature-support checks, and matched carrier analysis with sensitivity testing. Placebo diagnostics were restricted to buildings with a mean blocked M3 CV-RMSE below 25%. Temperature-bin coefficients were interpreted only when the relevant stratum contained at least 24 training hours; unsupported or non-estimable bins were retained as inconclusive. Previous-day P10 operating regimes, common-hour matching, bootstrap confidence intervals, multiplicity control, and sensitivity analyses were applied throughout. These analyses assess robustness but do not provide physical validation.

The study addresses four research questions. RQ1 asks whether a leakage-resistant, load-derived operating-regime baseline improves out-of-sample prediction relative to a conventional change-point baseline, and whether any improvement arises from operating-regime control or temperature binning. RQ2 asks whether event- and recovery-related metrics can be distinguished from temperature-matched non-event periods. RQ3 examines how the academic-calendar definition and temperature-bin support affect the estimated load response and its interpretable temperature range. RQ4 asks whether an electricity-only difference between office and education buildings remains stable after same-building and common-hour restriction, carrier separation, alternative carrier weights, normalization choices, bootstrap uncertainty, and multiplicity control.

The study contributes a reproducible procedure that combines leakage-resistant operating-regime control, placebo calibration, prespecified calendar sensitivity, explicit support and estimability checks, and meter-coverage auditing in one analysis based on public data. The four-model ablation identifies whether predictive improvement arises from operating-regime control or temperature representation. The matched carrier analysis then assesses whether an apparent difference between office and education buildings remains consistent under changes in sample composition, carrier visibility, index construction, normalization, and uncertainty treatment. The results do not support a chilled-water-driven reversal of the building-use ranking. Rather, the ranking is too unstable to justify a directional claim.

The scope is intentionally limited. The dataset contains separate electricity and chilled-water loads together with building-use and operating-state information, but no data on envelope construction, form, orientation, fabric properties, indoor temperature, or occupant exposure. Accordingly, the study addresses building use, operation, and cooling infrastructure rather than design morphology. It does not measure thermal comfort, occupant resilience, or passive survivability, nor does it assume that either building-use group is inherently more resilient. Estimates without adequate data support, particularly above 42 °C, are reported as inconclusive.

The remainder of the paper is organized as follows. Section 2 reviews empirical studies of heat-response interpretation, predictive baselines, operational calendars, and metering structure. Section 3 describes the dataset and the four analytical checks. Section 4 reports the baseline, placebo, calendar, and meter-aware results. Section 5 discusses the implications for the use of open meter data. Sections 6 and 7 present the limitations and conclusions, respectively.

Figure 1 shows the full analytical sequence.

2. Related Work

2.1. Extreme Heat and Building Energy Demand

The influence of high outdoor temperature on building energy use is well established. A review covering many cities estimated that urban heat islands increase cooling energy use by about one-fifth on average, while partly reducing heating demand [17]. Field measurements in Athens found that urban cooling loads could approximately double and that peak cooling electricity could triple, relative to less dense surroundings [18]. City-scale evidence from Tokyo shows the same direction of effect in both residential and commercial buildings [19]. Much of the earlier evidence was simulation-based. Subsequent analyses of operational meter records have provided direct empirical evidence of urban-overheating effects [3], making measured-data studies increasingly attractive. The BDG2 portfolio used here also supported the ASHRAE Great Energy Predictor III competition [20], continuing a long tradition of predicting building loads from measured data [21].

2.2. Inferring Heat Response and Resilience from Measured Energy Data

Research based on smart-meter data has expanded rapidly. Such records have been used to model residential demand under changing climate conditions [22], quantify the effect of extreme temperatures on household electricity use [23], and review the methods and datasets used across the wider residential smart-meter literature [24].

Energy indicators are sometimes interpreted using the language of resilience, although formal definitions of thermal resilience usually rely on indoor thermal conditions and simulation. Examples include thermal-resilience benchmarking frameworks [25], definitions of resilient cooling [26], and applications to educational buildings [27] and care facilities [28]. These studies assess outcomes that energy meters cannot observe directly. The question addressed here is narrower as follows: whether meter records alone justify resilience-related language. Table 1 summarizes empirical studies that infer heat response, cooling behavior, or recovery-related quantities from operational energy data, together with studies of predictive-baseline assessment, occupancy or calendar control, and meter or carrier structure. Thermostat-based work on heat and outage planning [29] is discussed in the text but is not included in the comparative table because it does not report building-level energy-meter coverage.

Table 1. Selected empirical studies based on measured data that inform heat-response interpretation and the following three analytical controls: predictive-baseline assessment, occupancy or calendar control, and energy-carrier or meter control.

Reference	Year	Data and Setting	Quantity Inferred	Baseline Out-of-Sample Validation	Occupancy/ Calendar Control	Energy-Carrier/Meter Control
[14] Khalil-nejad et al.	2021	432 commercial buildings; 15 min whole-building electricity and outdoor temperature	Cooling setpoint-setback savings and HVAC cooling share	No event-specific out-of-sample counterfactual baseline reported	Yes, occupied and unoccupied cooling periods	Whole-building electricity only; no building-level chilled-water coverage check
[15] Khalilnejad et al.	2020	432 commercial buildings; 15 min whole-building electricity	HVAC operating patterns, baseload, and scheduling savings	No event-specific out-of-sample counterfactual baseline reported	Yes, load-derived occu-pied/unoccupied operating patterns	Whole-building electricity only
[16] Guo et al.	2025	1.93 million residential consumers across 134 ZCTAs, Harris County, Texas; cooling seasons 2022–2023	Nine resilience-related electricity KPIs and future cooling-season demand	No event-specific counterfactual baseline reported; forecasting performance evaluated separately	Not reported	Residential electricity only
[3] Su et al.	2021	Seoul; seven years of measured building energy linked to 54 weather stations	Urban-overheating effect on cooling and heating energy	Not reported	Not reported	Cooling and heating energy distinguished; no chilled-water meter-coverage assessment
[30] Pickering et al.	2018	Six commercial buildings in California and Texas; 15 min electricity and weather	Operational signatures, HVAC scheduling, and temperature sensitivity	Not reported	Operational schedules inferred from load; no direct occu-pancy/calendar variable	Whole-building electricity only
[31] Kim et al.	2020	One campus office building; 3778 hourly observations over 157 days; electricity, occupant rate, and weather	Electricity prediction and sensitivity to occupancy and weather inputs	Train/test prediction performance reported; not a heat-event counterfactual baseline	Yes, measured occupant rate	Whole-building electricity only
[32] Gui et al.	2021	122 university buildings in subtropical Australia; measured energy and space-use data	Academic-calendar effects on university energy use	Not reported	Yes, semester/trimester calendars and teach-ing/research space use	No explicit energy-carrier or meter-coverage assessment reported

Table 1. *Cont.*

Reference	Year	Data and Setting	Quantity Inferred	Baseline Out-of-Sample Validation	Occupancy/ Calendar Control	Energy-Carrier/Meter Control
[33] Gaspar et al.	2022	83 academic buildings; electricity, natural gas, weather, and COVID-19 lockdown periods	Weather-adjusted avoided energy during lockdown	Weather-adjusted period comparison; no out-of-sample counterfactual baseline validation	Yes, pre-lockdown, lockdown, and occupancy phases	Electricity and natural gas; no end-use or chilled-water submetering
[34] Zhao et al.	2023	Operational district-cooling system serving multiple buildings; measured DCS energy and weather forecasts	District-cooling-system energy prediction	Predictive-model testing reported; not a building-level heat-response counterfactual baseline	No, occupancy influence stated as neglected	District-cooling thermal-energy/system measurements; no building-level meter-coverage comparison
[35] Raymand et al.	2023	BDG2; 1636 buildings; hourly electricity and chilled-water consumption	Building characteristics for smart-meter-based remote auditing	Cross-validation and held-out testing for ML pipelines; not a heat-response counterfactual baseline	Not reported	Electricity and chilled-water features; chilled water improved accuracy by 7.2–13.5%
[36] Fu et al.	2021	One classroom building in Texas; chilled water, domestic cold water, and electricity	Occupancy-proxy effects on chilled-water cooling-baseline accuracy	Model performance compared across periods; independent or blocked validation not established	Yes, domestic-water and electricity proxies tested	Chilled-water outcome with separate domestic-water/electricity proxy meters
This study	2026	BDG2; six campuses; hourly electricity, chilled water, weather, and metadata	Extreme-heat load response and interpretability of resilience-style claims	Yes, five-offset repeated whole-week blocked validation and four-model ablation (CV-RMSE, BIC, and residual diagnostics)	Yes, previous-day load-derived operating regimes; BASE plus 12 prespecified academic-calendar alternatives; per-bin rank and support audit	Yes, same-building/common-hour carrier separation; electricity/chilled-water endpoints; weight and mean/P90/RMS normalization sensitivity

A targeted, reproducible literature search was used to develop Table 1; it was not designed as a systematic or PRISMA review. The search covered English-language, peer-reviewed journal articles published between 2000 and 28 July 2026 in Scopus, Web of Science Core Collection, and Google Scholar. The following three groups of terms were combined: measured building energy with heat response or recovery; campus operation with occupancy or academic-calendar terms; and chilled water or district cooling with meter coverage and interpretation. Candidate records were checked against publisher pages and Crossref. Studies were included when they used measured operational data and addressed at least one of the following topics: temperature–energy relationships, event or recovery indicators, occupancy or calendar effects, carrier or meter structure, or measured-data baseline validation. Simulation-only studies, non-peer-reviewed material, duplicate versions, and forecasting studies without clear relevance to temperature, occupancy, calendar, or carrier structure were excluded.

The search was designed to identify a focused set of closely relevant studies rather than to support an exhaustive claim about the entire literature. Exact database hit counts are not reported because institutional access did not permit reproducible verification of record counts in Scopus and Web of Science. As Table 1 shows, earlier empirical studies usually address only part of the interpretability problem. Some use measured occupancy, load-derived operating patterns, academic calendars, or lockdown periods. Others distinguish electricity, gas, chilled water, or district-cooling measurements. Broader portfolio studies have also used measured electricity data to compare temporal patterns and energy use across commercial and office buildings [30,37,38]. Predictive testing appears in several forecasting and building-characterization studies, but it is rarely designed as leakage-resistant blocked validation of a heat-event counterfactual. Within the targeted literature set, we did not identify a study that combined baseline assessment, occupancy or calendar control, and an explicit test of whether separately metered chilled water changed an apparent building-use contrast. This conclusion is limited to the stated search procedure.

2.3. Why the Three Controls Matter

Each control addresses a different source of error in the interpretation of metered load.

The baseline model determines what is treated as normal consumption and, in turn, what is labeled as anomalous. Baseline choice affects predictive accuracy and the deviations inferred from it [39,40]. Measurement-and-verification guidance places out-of-sample assessment at the center of baseline credibility [41–43]. A model may fit its training data closely and still perform poorly during the unusual conditions of a heat event.

Occupancy and the operating calendar determine when and how a building is used. Incorporating measured occupancy can improve baseline accuracy and load prediction [44,45]. On a campus, term and break periods shape annual energy use and often overlap with the hottest part of the year [32]. Even the treatment of calendar boundaries can alter change-point model parameters [46]. Lockdown studies further show that nominally vacant academic buildings may continue to consume a substantial share of their usual energy [33]. Without an operating-period control, a weak response during hot weather may be attributed to demand-side performance when it simply reflects reduced use.

The energy carrier and meter arrangement determine whether cooling demand is visible in the selected data stream. When cooling is supplied as chilled water, thermal demand is recorded separately from building electricity, so an electricity-only analysis may omit much of the cooling response. In BDG2-based models, chilled-water features improved the estimation of building characteristics by 7.2–13.5% [35]. The present analysis tests whether this meter structure also affects the interpretation of heat-response differences between building-use groups.

3. Materials and Methods

The workflow is shown in Figure 1. Sections 3.1–3.7 describe its individual components.

3.1. Dataset, Site and Building-Use Selection, and Meter Coverage

BDG2 is an open dataset of hourly energy use for a large non-residential portfolio, accompanied by weather series and building metadata [12]. Where available, electricity, chilled water, and steam are recorded on separate building-level meters.

The dataset was selected because it provides electricity and chilled-water records for the same buildings, together with weather and metadata, across a large shared portfolio [12]. Its peer-reviewed descriptor and use in the ASHRAE Great Energy Predictor

III competition [12,20] support reproducibility. Although the records cover 2016–2017, the analytical issues examined here, including baseline sensitivity, calendar overlap, and carrier separation, remain relevant to current meter-based studies. The age of the records is considered in Section 6.

The analysis was restricted to campus sites and to education and office buildings, which were sufficiently represented for within-site comparison. Laboratory and healthcare buildings were too few within individual sites, whereas assembly buildings showed event-driven load patterns that were inconsistent with the baseline design used here.

Sites were selected according to realized outdoor temperature rather than nominal location. Fox reached the highest temperature range in the portfolio, with a 95th-percentile outdoor temperature of 39.4 °C and a maximum of 48.3 °C and was therefore used as the primary extreme-heat case. The full baseline and placebo workflow was repeated at Panther as a prespecified moderate-heat comparison using the same education and office design. Panther reached a 95th percentile of 31.7 °C and a maximum of 36.1 °C and was not treated as a second extreme-heat case. Bull, Rat, Eagle, and Hog were retained to provide context for calendar definition, temperature support, and meter coverage.

Electricity coverage was defined as the proportion of expected hourly timestamps with a valid electricity reading. Missing readings were neither treated as zero nor interpolated in the primary analyses; models used only hours for which all required variables were available. Buildings with electricity coverage below 75% were excluded. This prespecified threshold retained sufficient buildings for matched site and building-use comparisons while supporting whole-week blocked validation. No separate maximum-gap exclusion rule was applied. Within each site and building-use cell, eligible buildings were ranked first by coverage and then by valid-hour count, and the 15 best-covered buildings were retained. The longest uninterrupted missing period was recorded descriptively to distinguish overall sparsity from temporal discontinuity.

Chilled-water availability and coverage were assessed separately because they determine eligibility for the meter check in Section 3.5. A building without a chilled-water meter was treated as missing for that carrier, not as having zero cooling demand. This selection procedure improves comparability and numerical stability but may favor buildings with more reliable metering.

For each site, the heat event was defined as the contiguous 72 h period with the highest mean outdoor air temperature in the available record. The event itself, the preceding 72 h, and the following seven days were excluded from baseline fitting and validation. This prevented the event and its immediate surroundings from influencing model selection. At Fox, the selected event ran from 19 June 2017 at 10:00 to 22 June 2017 at 09:00, with a mean temperature of 40.02 °C and a maximum of 48.30 °C. At Panther, the event ran from 5 July 2016 at 17:00 to 8 July 2016 at 16:00, with a mean of 30.33 °C and a maximum of 36.10 °C. Table 2 summarizes the resulting site, building-use, temperature, and meter-coverage profile.

Outdoor air temperature was divided into fixed 3 °C bins. This representation makes the estimated response and its data support visible within each temperature band and avoids extending a continuous slope into sparsely observed extremes. The trade-off is lower within-bin resolution and potentially less statistical power than a continuous model. Coefficients were interpreted only when the corresponding bin had adequate training support.

Table 2. Dataset composition, temperature summary, and meter coverage by site and building-use group. Chilled-water (CW) coverage is the median share of study hours with a valid reading among CW-equipped buildings in each group. The penultimate column gives the highest temperature bin supported in both term and break under the BASE calendar. For the strict Fox carrier comparison, the same-building and common-hour audit retained 13 education and 10 office buildings. The final column reports the range of electricity coverage and valid-hour counts across the 15 retained buildings in each site and building-use cell.

Site	Building-Use Group	Buildings n	With CW (n)	CW Coverage, Median (%)	T P95 (°C)	T Max (°C)	Highest Supported Bin, Center (°C)	Electricity Coverage and Valid Hours, Range
Fox	Education	15	14	100	39.4	48.3	40.5	99.98–99.99% 17,540–17,543 h
Fox	Office	15	10	100	39.4	48.3	40.5	99.91–99.99% 17,528–17,542 h
Bull	Education	15	10	99.8	33.9	41.1	34.5	98.16–98.30% 17,221–17,246 h
Bull	Office	15	9	99.8	33.9	41.1	34.5	91.87–98.40% 16,117–17,263 h
Rat	Education	15	0	Not applicable	30.6	37.8	34.5	99.78–99.98% 17,506–17,541 h
Rat	Office	15	0	Not applicable	30.6	37.8	34.5	99.57–99.98% 17,468–17,540 h
Eagle	Education	15	14	100	28.9	35.6	31.5	98.79–99.93% 17,332–17,531 h
Eagle	Office	15	13	99.9	28.9	35.6	31.5	98.97–99.52% 17,363–17,460 h
Panther	Education	15	6	91.3	31.7	36.1	34.5	80.69–81.44% 14,157–14,287 h
Panther	Office	15	4	88.3	31.7	36.1	34.5	80.68–81.87% 14,155–14,363 h
Hog	Education	15	12	98.0	27.0	35.6	31.5	99.84–100.00% 17,516–17,544 h
Hog	Office	15	10	98.0	27.0	35.6	31.5	100.00–100.00% 17,544–17,544 h

3.2. Baseline Models

3.2.1. Candidate Baseline Models

Interpreting heat response requires a baseline that estimates expected consumption under normal operation. Baseline specification can affect predictive accuracy and the deviations later classified as anomalous [39,40]. Four candidate models were compared in an ablation design rather than selecting one model in advance.

M1 used a conventional heating-and-cooling change-point function with hour-of-day and weekend controls, following established measurement-and-verification and inverse-modeling practice [41–43]. M2 replaced the change-point terms with fixed 3 °C temperature bins and included no operating-regime control. M3 added a load-derived operating-regime factor to M1. M4 combined that factor with the binned-temperature representation. The comparison separates the predictive contribution of operating-regime control from the contribution of the temperature representation. Equation (1) defines the expected load for building b at hour t under candidate model m.

$$\hat{E}_{b,t}^{(m)} = \alpha_b + g_b^{(m)}(T_t) + \gamma_b^{(m)T} r_{b,t} + \delta_b^T h_t + \eta_b W_t \quad (1)$$

In Equation (1), $\hat{E}_{b,t}^{(m)}$ denotes the expected load of building b at hour t under candidate model m; α_b is the building-specific intercept; $g_b^{(m)}(T_t)$ is the model-dependent temperature-response component; $r_{b,t}$ is the operating-regime indicator vector; h_t is the 23-element hour-of-day indicator vector; and W_t is the weekend indicator. The regime-coefficient vector $\gamma_b^{(m)}$ is set to zero for M1 and M2 and estimated for M3 and M4. Hour 0 and the first operating regime are the reference categories.

For M1 and M3, the temperature-response component takes the two-sided change-point form defined in Equation (2).

$$g_b^{CP}(T_t) = \beta_{c,b} max(T_t - T_{c,b}, 0) + \beta_{h,b} max(T_{h,b} - T_t, 0) \tag{2}$$

In Equation (2), $T_{c,b}$ and $T_{h,b}$ are the building-specific cooling and heating change points, while $\beta_{c,b}$ and $\beta_{h,b}$ are the corresponding slopes. The maximum operators impose zero contribution below the cooling hinge and above the heating hinge. Cooling change points were selected from 16 to 26 °C and heating change points from 8 to 18 °C, both in 2 °C increments, using training observations only.

For M2 and M4, the change-point component is replaced by the fixed binned-temperature representation in Equation (3).

$$g_b^{BIN}(T_t) = \sum_{k \in K_T} \theta_{b,k} I(T_t \in B_k) \tag{3}$$

In Equation (3), each B_k is a fixed temperature band with boundaries $[-\infty, 0)$, [0, 3), …, [39, 42), and [42, $+\infty$) °C; $\theta_{b,k}$ is the building-specific coefficient; and K_T is the set of non-reference bins. The omitted reference is the band centered nearest 18 °C. A bin was considered supported when it contained at least 24 training hours. Unsupported or previously unseen bins were not interpreted in the response-curve analysis. Binning was used to characterize the response curves and their support, not to select the predictive baseline.

3.2.2. Leakage-Resistant Operating-Regime Proxy

To avoid current-day target leakage, the primary regime assigned to day d was based on the 10th percentile of the previous day's 24 hourly loads. A training day required at least 18 valid hours on both day d and day d − 1. The proxy was standardized using training days only and clustered into K = 3 ordered groups by k-means, with 20 replicates and a maximum of 500 iterations. Cluster labels were ordered from low to high proxy load and assigned to every hour of day d. The resulting variable is an operating-state proxy that may reflect occupancy, scheduling, and control conditions; it is not a direct occupancy measurement. Current-day P10 and the median load from 00:00 to 05:59 were retained only for sensitivity analysis. A regime was considered supported when it contained at least 168 training hours.

For comparisons across buildings, each binned-temperature coefficient was normalized by the building's mean training load, as defined in Equation (4).

$$\begin{aligned} \theta_{\widetilde{b,k}} &= \theta_{b,k} / \overline{E}_{b,tr} \\ \overline{E}_{b,tr} &= (1/|T_{b,tr}|)\ \Sigma_{t \in T(b,tr)}\ E_{b,t} \end{aligned} \tag{4}$$

Here, $\theta_{\widetilde{b,k}}$ is the estimated coefficient for temperature band k, and $\bar{E}_{b,tr}$ is the mean observed load over the corresponding training observations. All comparisons of temperature-response coefficients across buildings and building-use groups use the normalized coefficients; raw coefficients are not compared across buildings. Predictive baseline performance is evaluated separately using CV-RMSE on the original load scale.

https://doi.org/10.3390/su18168289

3.2.3. Repeated Whole-Week Validation and Diagnostics

The non-event record was divided into calendar weeks and evaluated using five offset whole-week validation folds. In fold f, every fifth eligible week beginning at offset f was held out, so each non-event week served once as validation. Every candidate model was refitted independently within each fold. The heat-event window and its exclusion buffer were omitted from training and validation. For each building, the primary predictive score was the mean CV-RMSE across the five folds; the fold-level standard deviation and interquartile range described validation variability. Whole weeks were held out because adjacent hourly observations are strongly autocorrelated. Equation (5) defines the fold-level score.

$$CV\text{-}RMSE_{b,m}{}^{(f)} = \surd\{(1/|V_{b,f}|)\ \Sigma_{t\in V(b,f)}(E_{b,t} - \hat{E}_{b,t}^{(m,-f)})^2\}/\overline{E}_{b,f}{}^{val}$$
$$CV\text{-}RMSE_{b,m}^{-} = (1/5)\ \Sigma_{f=1}{}^{5}\ CV\text{-}RMSE_{b,m}{}^{(f)} \tag{5}$$

In Equation (5), $V_{b,f}$ is the validation set for building b in fold f, $\overline{E}_{b,f}^{val}$ is its mean observed load, and $\hat{E}_{b,t}^{(m,-f)}$ is the prediction from model m fitted without the weeks assigned to fold f. The primary building-level score is the arithmetic mean across the five offset folds. Model complexity was assessed using the Bayesian information criterion [47].

The BIC definition used for this secondary complexity diagnostic is given in Equation (6).

$$BIC_{b,m}^{(f)} = n_{b,f}^{tr} ln\left(\frac{RSS_{b,m}^{(f)}}{n_{b,f}^{tr}}\right) + k_{b,m}^{(f)} ln\left(n_{b,f}^{tr}\right) \tag{6}$$

In Equation (6), $n_{b,f}^{tr}$ is the number of training observations, $RSS_{b,m}^{(f)}$ is the training residual sum of squares, and $k_{b,m}^{(f)}$ is the effective model rank, including the two selected change-point parameters where applicable. BIC was calculated on identical training observations for all candidate models within each fold and was treated as a secondary complexity diagnostic. Model adequacy was also examined through residual autocorrelation at 1, 24, and 168 h and Ljung–Box tests at 24 and 168 lags. Because temporal dependence remained in all models, blocked CV-RMSE was the primary basis for model selection. Event hours entered neither training nor validation.

3.3. Event and Placebo Diagnostics

Event over-demand and under-demand were measured as the mean positive and negative normalized deviations across hours with both an observed load and a valid M3 prediction, as defined in Equation (7).

$$D_b = (1/\mid \varepsilon_b{}^{valid} \mid)\ \ \Sigma_{t\in\varepsilon(b,valid)}\ (E_{b,t} - \hat{E}_{b,t})/\bar{E}_b \tag{7}$$

The excluded fraction was reported, and a site and building-use cell would have been treated as unreliable if more than 20% of event hours were excluded. The 24 h temperature-bin support rule does not govern event predictions because the selected M3 baseline uses continuous change-point terms. That rule applies only to the binned response curves in Sections 3.4 and 3.5. Event-hour exclusions instead arose from missing observations, invalid predictions, or unsupported operating-regime assignments. Equation (8) defines post-event load-shape similarity using the mean-centered Pearson correlation with normal days.

$$\rho = \frac{\sum_h (x_h - \overline{x})(y_h - \overline{y})}{\sqrt{\sum_h (x_h - \overline{x})^2 \sum_h (y_h - \overline{y})^2}} \tag{8}$$

Cosine similarity and other scale-invariant measures were not used because they discriminate poorly among all-positive daily load profiles.

All event and recovery diagnostics were recomputed after baseline selection using M3 with the previous-day P10 operating-regime proxy. Positive rebound and load-shape disruption were compared with non-event windows whose mean temperature was within ± 3 °C of the corresponding post-event window. For an upper-tail statistic, the empirical p-value was calculated as $p = [1 + \Sigma \text{ I}(S\_j \geq S\_obs)]/(B + 1)$, so ties remained in the tail and could not increase apparent significance. The displayed placebo percentile was $\Sigma \text{ I}(S\_j \leq S\_obs)/B$. Rebound and load-shape tests used B = 500 draws, giving a tail-probability resolution of 0.002. The duration diagnostic used 2000 draws from contiguous 72 h non-event blocks. Its upper- and lower-tail p-values used the same plus-one correction, and the two-sided value was $\min\{1, 2 \min(p\_high, p\_low)\}$. Contiguous blocks were used because scattered hours would understate the strong serial dependence on the data. Recovery-placebo tests were limited to buildings with mean blocked M3 CV-RMSE below 25%; the duration diagnostic included all buildings with a valid event-residual series. These building-level thresholds were nominal and were not adjusted across metrics. Individual findings were treated as exploratory, and the aggregate counts were considered relative to the number expected at a 5% false-positive rate.

The controls were defined independently of their effect on the results. Baseline selection used out-of-sample prediction and BIC without reference to event residuals. Placebo tests calibrated event specificity against each metric's own null distribution. The meter check was based on the physical location of cooling measurement. This separation prevents the workflow from favoring a null result by construction.

3.4. Academic-Calendar Stratification

The load response to temperature was stratified by academic term and break. Because BDG2 sites are anonymized, the BASE calendar approximated standard US semester dates and was treated as an operating-calendar proxy rather than a direct measure of occupancy. Robustness was assessed under 13 prespecified definitions as follows: BASE; boundary shifts of −7, +7, −14, and +14 days; break expansion and contraction by 14 and 21 days; 8- and 12-week summer sessions classified as term; exclusion of an 8-week summer session; and exclusion of the final 14 days of term as an examination-period sensitivity. These alternatives were fixed before the coefficients were inspected.

For every building, stratum, and calendar definition, the 18 °C reference band was required to contain at least 24 training hours. Operating-regime and hour-of-day reference levels were selected only from levels present in the fitted stratum. Nuisance columns were retained only when they increased the design-matrix rank. A temperature bin was interpreted only when term and break each contained at least 24 training hours and coefficients were available for at least five buildings. Agreement between the calendar and the previous-day P10 operating-regime proxy was assessed using Cramér's V and normalized mutual information (NMI).

3.5. Meter-Coverage Check

The Fox meter analysis was limited to buildings that passed the same-building and common-hour audit as follows: 13 education and 10 office buildings. Every carrier specification used the same timestamps with valid electricity, chilled water, and weather data, while the operating-regime definition remained the previous-day electricity P10. Electricity-only and chilled-water-only curves were treated as the primary carrier endpoints. Because electricity and chilled water have different units and conversion efficiencies, their raw values were not added. The equal-weight normalized index in Equation (9) was retained only as a sensitivity specification.

$$y_{b,t}^{idx} = \frac{1}{2}\frac{E_{b,t}}{\overline{E}_b} + \frac{1}{2}\frac{C_{b,t}}{\overline{C}_b} \tag{9}$$

The electricity weight varied from 0 to 1 in increments of 0.1, with the chilled-water weight set to one minus the electricity weight. Each carrier was scaled by its training mean, training 90th percentile (P90), or training root-mean-square (RMS) value. Weighting and normalization can alter both the magnitude and shape of a combined index. Neither equal weighting nor a non-significant result was treated as evidence of equivalence. Differences between building-use groups were assessed with the Mann–Whitney rank-sum test [48] and Cliff's delta [49].

3.6. Robustness Analysis

Sensitivity to analytical choices was assessed by resampling the observed data; no physical behavior was simulated. Equation (10) defines the hot-bin difference between office and education buildings. For the matched carrier analysis, the same resampled building identifiers were used for every carrier endpoint, weight, and normalization method.

$$\Delta^{(m)} = med_b\left\{\theta_{b,hot}^{(m)}\right\}_{office} - med_b\left\{\theta_{b,hot}^{(m)}\right\}_{education} \tag{10}$$

Buildings, rather than individual hours, were resampled with replacement because consecutive observations within a building are strongly correlated and hour-level resampling would produce intervals that are too narrow [50]. The primary analyses used B = 10,000 common resamples and percentile-based 95% intervals defined by the 2.5th and 97.5th percentiles. Comparisons between office and education buildings included the observed difference, Mann–Whitney p-value, Cliff's delta, bootstrap median and 95% interval, and the probability that the difference exceeded zero. Benjamini–Hochberg false-discovery-rate adjustment [51] was applied within the prespecified test families, and the 99 combinations of carrier weight, normalization, and temperature bin were also assessed as one global family. The stage decomposition compared electricity in all selected buildings (stage A), electricity in the matched meter sample (stage B), and the multi-meter specification in that same sample (stage C). These intervals quantify stability under building resampling. They do not fully propagate uncertainty in the within-building coefficients, dependence induced by common weather, or uncertainty associated with sampling a single site.

3.7. Use of AI-Assisted Tools in Manuscript Preparation

ChatGPT 5.6 (OpenAI; accessed July 2026) and Claude Opus 4.8 (Anthropic (San Francisco); accessed July 2026) were used during manuscript preparation for language editing, assistance with MATLAB2023b code debugging, and support in the literature searching. The authors executed all analyses, checked every numerical output, verified the cited sources against publisher records, and made all methodological and interpretive decisions. The tools were not used to generate or alter the underlying data, execute the final analyses, or create the reported figures and numerical results.

4. Results

4.1. Repeated Blocked Validation Identifies the Contribution of Operating-Regime Control

Across the 60 buildings, the leakage-resistant operating-regime change-point model (M3) had a lower mean blocked-validation error than the conventional change-point model

(M1) in 58 cases. The median paired reduction in CV-RMSE was 1.055 percentage points (Wilcoxon signed-rank $p = 2.10 \times 10^{-11}$; two-sided sign-test $p = 3.18 \times 10^{-15}$; and rank-biserial effect size = 0.995). Median building-level CV-RMSE fell from 17.13% with M1 to 15.95% with M3. M3 improved all 15 Fox education buildings and all 15 Fox office buildings, as well as 14 of 15 Panther education buildings and 14 of 15 Panther office buildings. Its mean training-fold BIC was lower than that of M1 in all 60 buildings.

The ablation analysis attributed the predictive gain to operating-regime control rather than to temperature binning. Adding the regime factor to the change-point model improved 58 of 60 buildings. Replacing the change-point terms with temperature bins without regime control improved only 21 of 60 buildings and produced no significant median benefit (Wilcoxon $p = 0.060$). Once regime control was included, M4 did not improve on M3 (median M4 − M3 difference = +0.007 percentage points; $p = 0.877$). M3 was selected because it achieved comparable held-out accuracy with fewer parameters (median 31 versus 40), a lower median full-fit BIC (84,046 versus 84,057), and valid fold-level results for every building. Event residuals also depended on the baseline as follows: replacing M1 with M3 reduced the absolute mean event deviation in 42 of 60 buildings, although the positive and negative components did not change uniformly.

Substantial residual dependence remained. Median autocorrelation at lags of 1, 24, and 168 h was 0.886, 0.580, and 0.564 for M1 and 0.869, 0.419, and 0.469 for M3. Operating-regime control reduced daily and weekly structure but did not remove it; Ljung–Box tests remained significant in every building at both 24 and 168 lags. Repeated whole-week validation provides the primary evidence of predictive improvement, while BIC serves as a secondary complexity diagnostic. Figure 2 summarizes the paired validation and ablation results. The corresponding site- and building-use-level results for baseline validation, event residuals, and placebo diagnostics are summarized in Table 3.

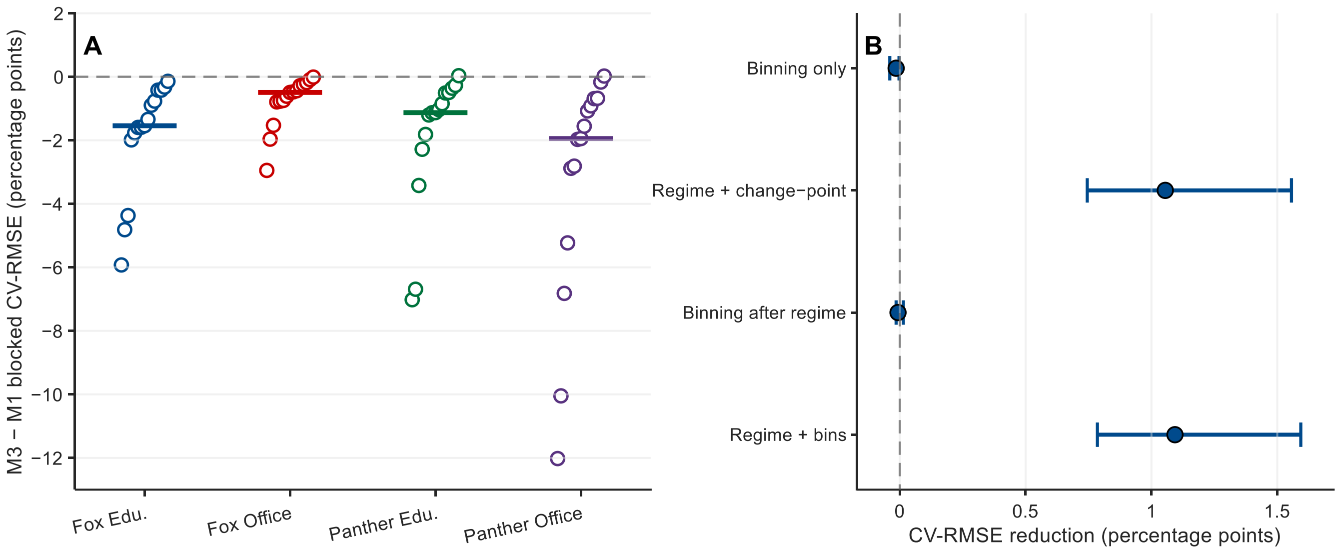

Figure 2. Repeated whole-week blocked validation and baseline ablation. (**A**) Paired building-level differences in mean CV-RMSE between the selected load-derived operating-regime change-point baseline (M3) and the conventional change-point baseline (M1) across the four site and building-use cells. Negative values favor M3, and horizontal segments show cell medians. (**B**) Median building-level changes in CV-RMSE are attributable to temperature binning and operating-regime control. Error bars show 95% building-level bootstrap confidence intervals.

Table 3. Repeated whole-week baseline validation, event residuals, and placebo diagnostics by site and building-use cell. Columns are grouped under validation, event residuals, and placebo diagnostics. Baseline columns compare the conventional change-point model (M1) with the selected previous-day P10 operating-regime model (M3). Rebound and load-shape entries give the number of significant buildings among those meeting the prespecified M3 baseline-quality threshold. Duration entries give the median two-sided block-null p-value and the number of significant buildings among all valid cases.

	Validation				Event Residuals			Placebo Diagnostics			
Site and Building-Use Cell	n	CV-RMSE (%) M1/M3	Median Training-Fold BIC M1/M3	M3 Better CV/BIC	Over-Demand M1/M3	Under-Demand M1/M3	Absolute Deviation Shrank	Rebound Sig.	Load-Shape Sig.	Signed Rebound (Median)	Duration p (Median; Sig.)
Fox, education	15	15.7 13.7	84,272 80,495	15/15 15/15	0.055 0.036	0.020 0.025	12/15	1/14	0/14	+0.002	0.81; 1/15
Fox, office	15	18.2 17.0	77,597 77,252	15/15 15/15	0.020 0.033	0.062 0.065	9/15	1/12	0/12	+0.013	0.29; 1/15
Panther, education	15	17.4 16.1	66,776 62,864	14/15 15/15	0.042 0.036	0.034 0.040	10/15	0/11	0/11	−0.016	0.55; 0/15
Panther, office	15	19.8 17.6	57,027 50,458	14/15 15/15	0.060 0.039	0.020 0.049	11/15	0/12	1/12	−0.029	0.50; 1/15
All	60	17.1 16.0	69,043 67,247	58/60 60/60	0.044 0.035	0.031 0.038	42/60	2/49	1/49	−0.011	0.50; 3/60

4.2. Recovery-Related Metrics Show Limited Event Specificity in Placebo Tests

Under M3, 49 of the 60 buildings met the prespecified mean blocked CV-RMSE threshold of 25% and were eligible for the recovery placebo tests. At the nominal, unadjusted 0.05 level, positive rebound was significant in 2 of 49 eligible buildings and post-event load-shape disruption in 1 of 49. The median temperature-matched percentile was 0.462 for rebound and 0.466 for load shape, indicating that the typical event value lay near the center of its matched non-event distribution. The median signed 48 h post-event residual integral was −0.0106, providing no evidence of systematic net over-consumption after the event.

The independent 72 h block-null duration diagnostic was significant in 3 of 60 buildings, with a median two-sided p-value of 0.499. This number of isolated findings is consistent with the number expected at a nominal 5% threshold. The tested recovery-related signals were not reproducibly distinguishable from temperature-matched or duration-matched non-event variation. This conclusion applies only to the specified metrics, windows, and metered-load data and does not imply that physical recovery behavior is absent. Figure 3 presents the three placebo diagnostics and their prespecified thresholds.

4.3. Calendar Definition Bounds the Supported Interpretation

Under the BASE calendar at Fox, the education and office curves satisfied the joint support rule through 39–42 °C. No building-level fit was supported above 42 °C, so that range remained inconclusive (Figure 4). Within the supported hot bins, no comparison between term and break remained significant after false-discovery-rate adjustment within the site and building-use family. For education, q = 0.670 at 36–39 °C and q = 0.962 at 39–42 °C; for office, q = 0.808 and q = 0.112, respectively. The BASE calendar provided no reproducible evidence of a difference between term and break in either group.

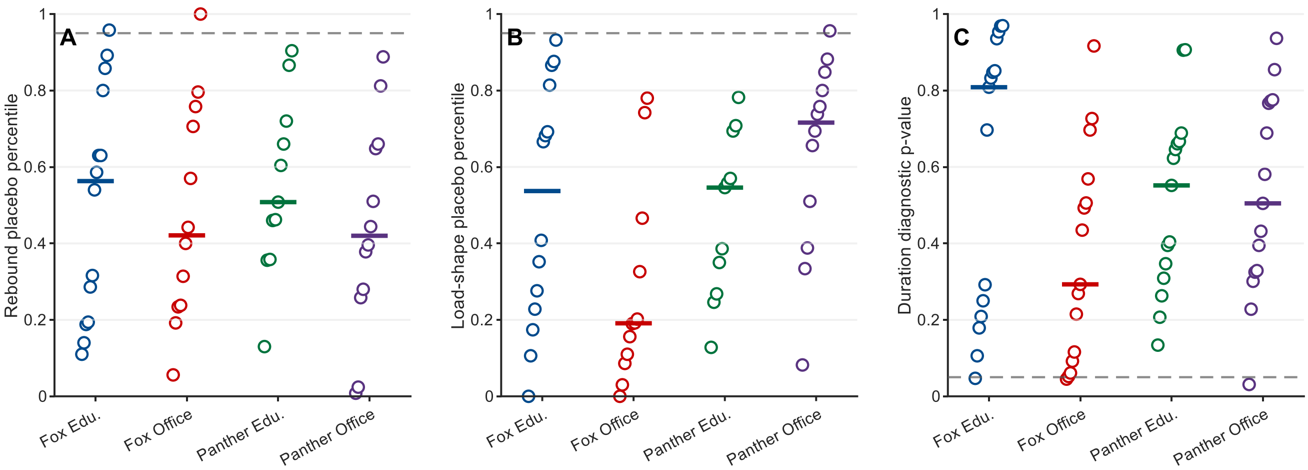

Figure 3. M3-based placebo diagnostics by site and building-use cell. Panels (**A**,**B**) show positive rebound and post-event load-shape disruption as the percentile of each event value within its temperature-matched null distribution; the dashed line marks the 0.95 significance threshold. Panel (**C**) shows two-sided p-values from the contiguous 72 h block-null duration diagnostic; the dashed line marks p = 0.05. Horizontal colored segments show cell medians. Panels (**A**,**B**) include the 49 buildings that met the prespecified M3 baseline-quality threshold, with 2 and 1 significant cases, respectively. Panel (**C**) includes all 60 buildings, of which 3 were significant.

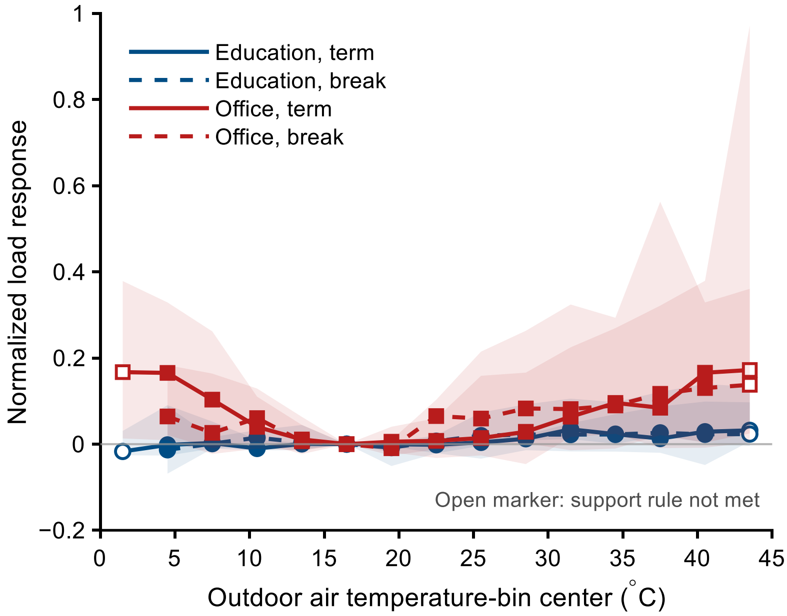

Figure 4. Load-response curves for term and break at Fox under the BASE academic-calendar definition and the previous-day P10 operating-regime specification. Lines show the building-level median normalized response, and shaded bands show 95% bootstrap confidence intervals. Filled markers identify temperature bins that satisfy the joint support rule: at least 24 training hours in both term and break, with coefficients available for at least five buildings. Open markers identify bins that do not meet this rule. They are shown descriptively and are not used for inferential interpretation.

Across the 13 prespecified calendar definitions, joint support at Fox was retained in 11 of 13 scenarios at 36–39 °C, 10 of 13 at 39–42 °C, and only 3 of 13 above 42 °C. Education coefficients retained their BASE sign in 81.8% of supported 36–39 °C scenarios and 80.0% of supported 39–42 °C scenarios. Office coefficients remained positive in every supported scenario. When an 8- or 12-week summer session was classified as term, the education coefficients changed from positive to negative in the hot bins. These sign changes indicate sensitivity to calendar assumptions but do not identify one alternative calendar as uniquely correct.

Calendar labels and load-derived operating regimes overlapped only partly. Under the BASE definition, median Cramér's V was 0.234 for education and 0.179 for office, while median NMI was 0.037 and 0.023. Calendar stratification and the previous-day P10 proxy were retained as complementary controls rather than treated as interchangeable measures of occupancy. Table 4 summarizes support, sign stability, and calendar-regime agreement.

Table 4. Matched-sample, carrier, and calendar robustness results for Fox. In Panels (**A**,**B**), Δ denotes the office minus education difference in the mean response across 36–42 °C. Intervals are 95% building-level bootstrap intervals based on B = 10,000 resamples, and q is the BH-FDR-adjusted *p*-value. Panel (**A**) separates electricity in all selected buildings, electricity in the matched sample, and the multi-meter specification in the same matched sample. Panel (**B**) reports the carrier endpoints and equal-weight normalization sensitivity. Panel (**C**) reports support and sign stability across 13 prespecified calendar definitions.

Panel A: Sample and carrier decomposition (mean 36–42 °C)						
Specification	**n (E/O)**	**Observed Δ**	**95% bootstrap interval**	**Mann–Whitney *p***	**Cliff's δ**	**BH-FDR q**
A: electricity, all buildings	15/15	0.104	[−0.022, 0.747]	0.051	0.422	0.223
B: electricity, matched buildings	13/10	0.048	[−0.105, 0.180]	0.556	0.154	0.828
C: multi-meter, matched buildings	13/10	0.094	[−0.233, 0.342]	0.780	0.077	0.828
Panel B: Carrier endpoints and equal-weight normalization sensitivity (matched buildings)						
Specification	**n (E/O)**	**Observed Δ**	**95% bootstrap interval**	**Mann–Whitney *p***	**Cliff's δ**	**BH-FDR q**
Electricity only, mean normalization	13/10	0.048	[−0.105, 0.172]	0.515	0.169	0.975
Chilled water only, mean normalization	13/10	0.162	[−0.474, 0.725]	0.877	0.046	0.975
Equal weight, mean normalization	13/10	0.107	[−0.232, 0.373]	0.598	0.138	0.975
Equal weight, P90 normalization	13/10	0.046	[−0.173, 0.252]	0.598	0.138	0.975
Equal weight, RMS normalization	13/10	0.102	[−0.197, 0.305]	0.926	0.031	0.985
Panel C: Academic-calendar robustness						
Building-use group	**BASE top supported bin**	**Support 36–39 °C**	**Support 39–42 °C**	**Support > 42 °C**	**Sign preserved 36–39/39–42**	**BASE Cramér's V/NMI**
Education	39–42 °C	11/13	10/13	3/13	81.8%/80.0%	0.234/0.037
Office	39–42 °C	11/13	10/13	3/13	100%/100%	0.179/0.023

Figure 4 shows the term and break response curves under the BASE calendar.

4.4. Building-Use Contrasts Are Not Robust to Sample and Carrier Specification

In the electricity-only analysis of all selected Fox buildings, the observed difference between office and education buildings across 36–42 °C was 0.104. This estimate was statistically unstable: the Mann–Whitney p-value was 0.051, the BH-FDR-adjusted q-value was 0.223, and the 95% bootstrap interval [−0.022, 0.747] included zero. The 39–42 °C bin had a nominal p-value of 0.038, but its q-value was also 0.223 and its 95% interval included zero. Restricting the electricity-only analysis to buildings retained in the matched meter sample reduced the observed mean hot-bin difference to 0.048 ($p = 0.556$; 95% interval [−0.105, 0.180]). The change from stage A to stage B indicates that meter availability also altered the composition of the analyzed sample.

Applying the matched multi-meter specification to the same 13 education and 10 office buildings produced an observed mean hot-bin difference of 0.094 ($p = 0.780$; 95% interval [−0.233, 0.342]). The change from matched electricity to the matched multi-meter result had a bootstrap median of 0.035 and a 95% interval of [−0.265, 0.281]. The probability of attenuation was 0.425. These results provide no consistent evidence of attenuation, amplification, or reversal after carrier inclusion.

The carrier-separated endpoints led to the same inference. In the matched common-hour sample, the electricity-only mean hot-bin difference was 0.048 (95% interval [−0.105, 0.172]; $p = 0.515$; Cliff's $\delta = 0.169$). The chilled-water-only difference was 0.162 (95% interval [−0.474, 0.725]; $p = 0.877$; Cliff's $\delta = 0.046$) (Figure 5). Equal weighting gave observed differences of 0.107, 0.046, and 0.102 under mean, P90, and RMS normalization. All three intervals included zero. Across all 99 combinations of carrier weight, normalization, and temperature bin, every 95% interval included zero, no unadjusted Mann–Whitney p-value was below 0.05, and no result remained significant after BH-FDR adjustment.

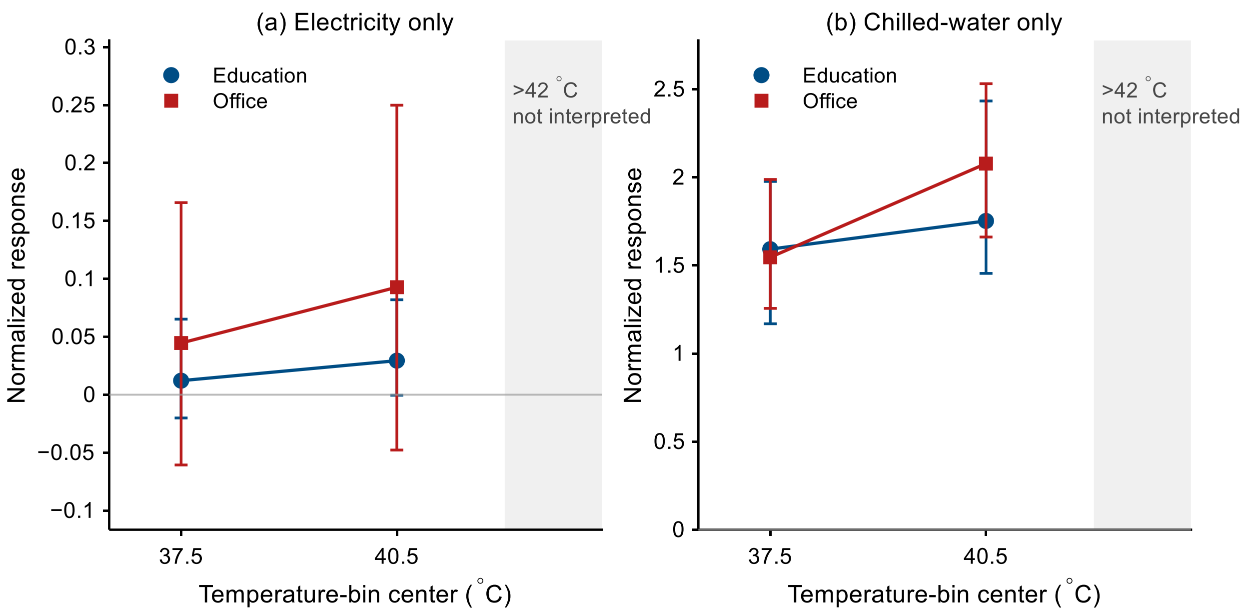

Figure 5. Carrier-separated heat-response curves for matched Fox buildings. Panel (**a**) shows electricity-only responses and panel (**b**) chilled-water-only responses for the same sample (education n = 13; office n = 10), using the previous-day electricity P10 operating-regime proxy and mean normalization. Points show building-level medians and vertical bars show 95% bootstrap confidence intervals. Only bins with joint support are interpreted. The region above 42 °C is marked as not interpreted because it does not meet the support threshold.

Figure 6 separates electricity in all selected buildings (stage A), electricity in the matched sample (stage B), and the multi-meter specification in that same sample (stage C).

The point estimate decreased after sample restriction, whereas the change associated with carrier inclusion remained too uncertain to support a directional conclusion. Table 4 reports numerical decomposition and carrier sensitivity.

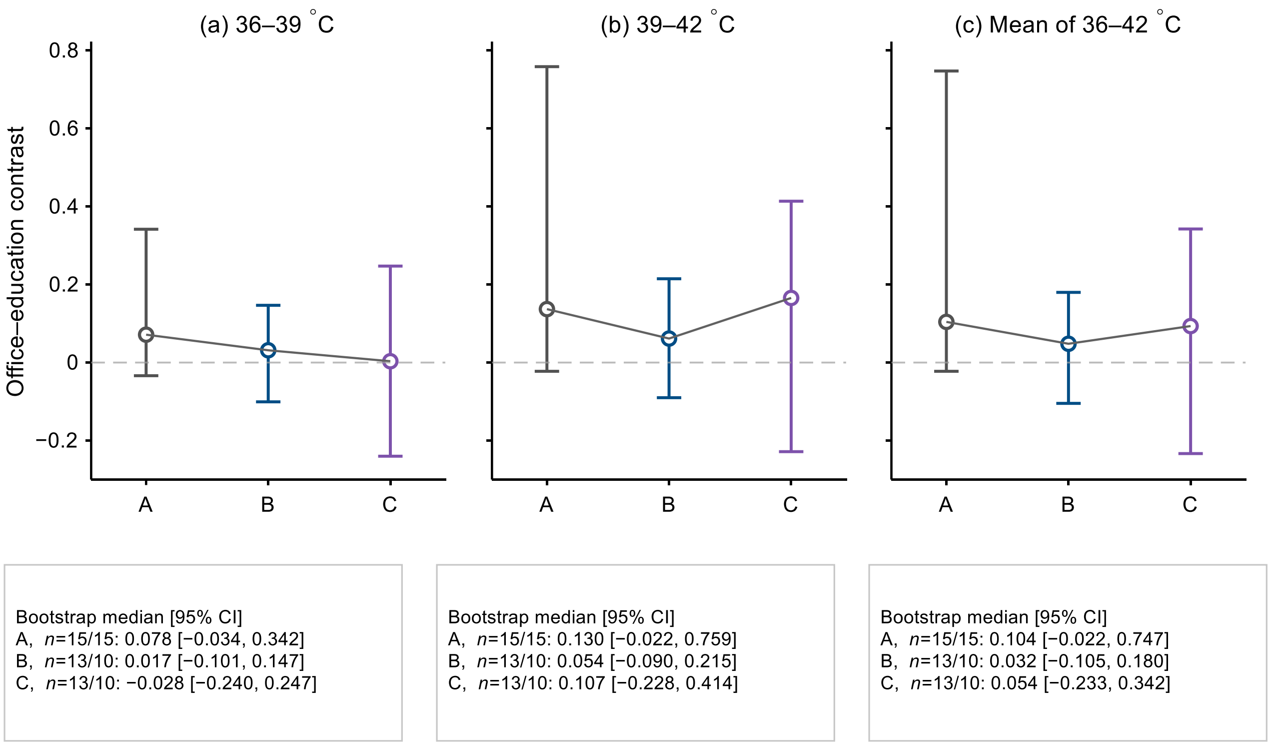

Figure 6. Decomposition of the difference between office and education buildings at Fox. Stage A uses electricity-only data for all selected buildings (education/office n = 15/15). Stage B restricts the electricity-only analysis to buildings retained in the matched meter sample (n = 13/10). Stage C applies the matched multi-meter specification to that same sample. Points show the observed differences, vertical bars show percentile-based 95% bootstrap intervals, and the boxes below the panels report the bootstrap median and 95% interval for each stage. Panel (**a**) covers 36–39 °C, panel (**b**) 39–42 °C, and panel (**c**) the mean of the two hot bins. Every bootstrap interval includes zero.

The remaining campuses provide information about temperature support rather than an independent replication of the Fox analysis at 36–42 °C. Under the BASE calendar, Panther and Rat were supported only through 33–36 °C, Bull had limited support at 36–39 °C, and Eagle and Hog were supported only at lower temperatures. The implication is therefore methodological: meter coverage, common-hour availability, sample composition, and index construction must be documented before differences between building-use groups are interpreted.

Figure 5 shows the matched carrier-separated response curves.

5. Discussion

5.1. What a Meter Records and What It Does Not

The distinction between a building as designed and a building as operated is familiar in building-performance research. The results show that the same distinction applies to operational meter data. A meter records the outcome of a particular operating regime, meter architecture, and pattern of data availability; it does not describe the building in isolation. At Fox, the estimated difference between office and education buildings changed after restriction to buildings with common carrier coverage and changed again under alternative carrier weights and normalization methods. Every corresponding 95% interval included zero, so the data do not support a stable ranking of the two building-use groups.

Open energy datasets remain valuable because they provide scale and reproducibility that are difficult to obtain through individual audits. Their interpretation, however, depends on a clear definition of what was observed. The baseline, operating calendar, meter coverage, common-hour availability, and carrier construction determine which buildings and which forms of cooling enter the comparison. These choices define the measurement basis and should not be treated as secondary reporting details.

5.2. Meter Availability, Sample Composition, and Carrier Visibility

District cooling matters because cooling demand may appear on a chilled-water meter rather than on the building's electricity meter. The analysis also shows that carrier visibility cannot be separated from sample composition. Restricting the all-building electricity analysis to the matched meter sample reduced the mean hot-bin point estimate from 0.104 to 0.048. Applying the multi-meter specification to those same buildings changed the point estimate to 0.094, but the interval for the carrier-related change was wide and included zero. Accordingly, the results do not support the conclusion that chilled-water inclusion causes the difference to disappear or reverse.

The weighting and normalization analysis leads to the same conclusion. Electricity-only, chilled-water-only, and equal-weight results all had intervals that included zero, and none of the 99 prespecified tests were significant either before or after multiplicity adjustment. Although the numerical value of a combined index changed with its construction, the inferential result did not: the data did not identify a stable difference between office and education buildings.

A meter audit should therefore precede any campus ranking used to prioritize retrofits, establish performance targets, or structure energy contracts. When carrier availability affects both what is measured and which buildings remain in the comparison, the resulting ranking may reflect the data architecture more strongly than the buildings' response to heat.

5.3. Building Use and Calendar as First-Order Operational Variables

Education and office buildings differ in both their physical use and annual operating schedules. The BASE calendar supported comparisons through 39–42 °C at Fox, yet neither group showed a statistically significant difference between term and break after FDR adjustment. Across the prespecified alternatives, office coefficients remained positive whenever support was available, whereas education coefficients changed sign under different summer-session assumptions. The estimated education response therefore depends on how term and break are defined.

Calendar labels showed only limited agreement with the previous-day P10 operating-regime proxy. The following two controls contain different information: the calendar identifies institutional periods, whereas the load-derived regime reflects the operating state realized in the meter data. Neither is a direct measure of occupancy. Used together, they show whether an interpretation depends on the nominal schedule, the observed load state, or both.

The data also impose a clear upper boundary on interpretation. Support above 42 °C was available in only 3 of the 13 calendar scenarios, and the BASE fit was unsupported. The dataset cannot determine the term-time response above this threshold with adequate reliability, so the response should be reported as inconclusive rather than extrapolated from the lower-temperature trend.

5.4. Implications for Performance-Based Evaluation

The predictive gain cannot be attributed solely to greater model flexibility. Adding a load-derived operating-regime factor improved repeated blocked-validation performance in 58 of 60 buildings, whereas temperature binning provided no detectable predictive benefit after regime control. The predictive baseline and the descriptive temperature-bin curves serve different purposes as follows: M3 defines the event counterfactual, while the binned model makes support limits and response patterns visible.

Performance-based evaluation uses measured evidence to inform design and operational decisions. Each control in the present workflow answers a distinct interpretive question. Repeated validation assesses whether the baseline generalizes to held-out periods. Placebo analysis determines whether an event metric is unusual relative to matched non-event periods. Calendar sensitivity assesses whether the response changes under plausible definitions of the operating period. Matched carrier analysis determines whether a comparison remains consistent after accounting for meter availability, sample restriction, and index construction.

Applied in sequence, these controls require analysts to document how expected consumption was defined and whether the baseline performed adequately on held-out weeks. They also require the event signal to be compared with matched non-event variation, the supported temperature range to be identified under plausible calendar definitions, and building-use differences to be tested using matched buildings, common hours, separate carriers, alternative weights, and multiplicity-adjusted uncertainty. Claims that do not satisfy these checks should be narrowed to match the evidence.

These checks do not require new instrumentation or proprietary automation data. They require a transparent meter inventory, reproducible calendar definitions and exclusions, common-hour matching, and explicit reporting of uncertainty and unsupported ranges. Rather than guaranteeing a single answer, the checks indicate when open meter data are sufficient for a conclusion and when the evidence remains inconclusive.

6. Limitations

Several limitations define the scope of the results. Fox was the only campus to reach the extreme temperature range examined here, so its findings require replication before they can be extended to other climates or operating structures. The records date from 2016 to 2017 and may not represent current schedules, controls, retrofit conditions, or climate exposure. Selecting the 15 best-covered buildings on each site and building-use cell improves comparability but may favor well-metered buildings; the strict carrier analysis retained only 13 education and 10 office buildings. The academic calendars are reconstructed proxies because the sites are anonymized, and the summer-session alternatives cannot identify the actual institutional calendar.

The carrier-weight analysis uses normalized indices rather than a source-energy conversion, which would require plant-efficiency assumptions that are not available in BDG2. Building-level bootstrap captures sensitivity to building selection but does not fully propagate uncertainty in the within-building coefficients, dependence caused by common weather, or site-level sampling uncertainty. Finally, meter data cannot establish indoor thermal conditions, occupant exposure, causal mechanisms, or physical resilience.

Independent validation requires a site where the operating context and meter architecture can be observed directly. A suitable campaign would compare the reconstructed calendar with institutional timetables and measured occupancy and compare the previous-day P10 regimes with building-automation records for air-handling units, setpoints, valve positions, and equipment schedules. Building-level chilled-water demand should be checked against measured flow and supply-return temperature differences, together with district-plant production and distribution logs. Outdoor conditions should also be paired with indoor air temperature and relative humidity. Meter commissioning records and on-site inspection would be needed to confirm sensor mapping, units, and the causes of missing data. These observations would test whether the inferred regimes and carrier-specific responses correspond to actual operation and indoor conditions, thereby providing physical validation beyond the sensitivity analyses reported here.

7. Conclusions

A heat response inferred from a single electricity meter is conditional on the analytical and metering context. In this study, a previous-day load-derived operating-regime factor improved repeated whole-week prediction in 58 of 60 buildings, while temperature binning added no predictive benefit after regime control. Recovery-related event signals were uncommon under temperature-matched placebo and block-null tests. Calendar sensitivity showed that the supported temperature range and the education coefficients depended on boundary and summer-session assumptions; temperatures above 42 °C remained inconclusive. The estimated difference between office and education buildings was unstable under matched-sample restriction, common-hour analysis, carrier separation, alternative weighting and normalization, 95% bootstrap uncertainty, and BH-FDR adjustment.

Taken together, these results show that open meter data can support a defensible assessment of heat response only when baseline quality, calendar definition, meter coverage, sample composition, and carrier construction are reported together. In this dataset, the available evidence does not identify one building-use group as more responsive to heat; instead, it shows why a stable ranking cannot be justified.

Author Contributions: Conceptualization, C.C. and L.Ö.; methodology, L.Ö. and C.C.; software, Y.C.; validation, L.Ö., C.C. and Y.C.; formal analysis, Y.C. and L.Ö.; investigation, C.C. and Y.C.; resources, C.C.; data curation, Y.C.; writing, original draft preparation, C.C., L.Ö. and Y.C.; writing, review and editing, C.C. and L.Ö.; visualization, Y.C.; supervision, C.C. and L.Ö.; project administration, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study analyzed publicly available data. The Building Data Genome Project 2 is openly available and is cited as reference [12]. The data presented in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.21355524, reference number [52].

Acknowledgments: During preparation of this manuscript, the authors used ChatGPT (OpenAI; accessed July 2026) and Claude (Anthropic (San Francisco); accessed July 2026) for language editing, MATLAB code debugging assistance, and literature search support. The authors reviewed and edited all output, verified the analyses and cited sources, and take full responsibility for the content of the publication. No AI-assisted tool was used to create or alter the underlying data, execute the final analyses, or generate the reported numerical results.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BDG2	Building Data Genome Project 2
BH-FDR	Benjamini–Hochberg false-discovery-rate adjustment
BIC	Bayesian information criterion

CV-RMSE	Coefficient of variation in the root-mean-square error
CW	Chilled water
HVAC	Heating, ventilation and air conditioning
NMI	Normalized mutual information
P10	10th percentile
P90	90th percentile
RMS	Root mean square

References

1. United Nations Environment Programme. *Global Status Report for Buildings and Construction 2025–2026: As Climate Risks Rise and Cities Grow, We Must Rethink How We Build to Create Better Lives for All*; UNEP: Paris, France, 2026. [CrossRef] [PubMed]
2. International Energy Agency. *Global Energy Review 2025*; IEA: Paris, France, 2025. Available online: https://www.iea.org/reports/global-energy-review-2025 (accessed on 28 July 2026).
3. Su, M.A.; Ngarambe, J.; Santamouris, M.; Yun, G.Y. Empirical evidence on the impact of urban overheating on building cooling and heating energy consumption. *iScience* **2021**, *24*, 102495. [CrossRef] [PubMed]
4. Santamouris, M. Recent progress on urban overheating and heat island research: Integrated assessment of the energy, environmental, vulnerability and health impact. *Energy Build.* **2020**, *207*, 109482. [CrossRef]
5. Liang, J.; Qiu, Y.; Wang, B.; Shen, X.; Liu, S. Impacts of heatwaves on electricity reliability: Evidence from power outage data in China. *iScience* **2025**, *28*, 111855. [CrossRef] [PubMed]
6. Baniassadi, A.; Heusinger, J.; Sailor, D.J. Energy efficiency vs. resiliency to extreme heat and power outages: The role of evolving building energy codes. *Build. Environ.* **2018**, *139*, 86–94. [CrossRef]
7. Ibrahim, A.; Alsukkar, M.; Wang, L.; Hermaputi, R.L.; Dong, Y.; Yin, S.; Xiao, Y.; Hu, P. Targeting integrated daylighting comfort in deep office by multi-objective optimization using partial vertical deployment of adaptive corrugated shadings. *Energy Built Environ.* **2026**, *in press*. [CrossRef]
8. Ibrahim, A.; Alsukkar, M.; Eltaweel, A.; Hermaputi, R.L. Multi-objective optimization of daylighting for an office space in a very hot climate: A comparison of corrugated, perforated, and separated shadings. *Buildings* **2026**, *16*, 2625. [CrossRef]
9. Rezaie, B.; Rosen, M.A. District heating and cooling: Review of technology and potential enhancements. *Appl. Energy* **2012**, *93*, 2–10. [CrossRef]
10. Lake, A.; Rezaie, B.; Beyerlein, S. Review of district heating and cooling systems for a sustainable future. *Renew. Sustain. Energy Rev.* **2017**, *67*, 417–425. [CrossRef]
11. Werner, S. International review of district heating and cooling. *Energy* **2017**, *137*, 617–631. [CrossRef]
12. Miller, C.; Kathirgamanathan, A.; Picchetti, B.; Arjunan, P.; Park, J.Y.; Nagy, Z.; Raftery, P.; Hobson, B.W.; Shi, Z.; Meggers, F. The Building Data Genome Project 2, energy meter data from the ASHRAE Great Energy Predictor III competition. *Sci. Data* **2020**, *7*, 368. [CrossRef] [PubMed]
13. Villa, D.L. Institutional heat wave analysis by building energy modeling fleet and meter data. *Energy Build.* **2021**, *237*, 110774. [CrossRef]
14. Khalilnejad, A.; French, R.H.; Abramson, A.R. Evaluation of cooling setpoint setback savings in commercial buildings using electricity and exterior temperature time series data. *Energy* **2021**, *233*, 121117. [CrossRef]
15. Khalilnejad, A.; French, R.H.; Abramson, A.R. Data-driven evaluation of HVAC operation and savings in commercial buildings. *Appl. Energy* **2020**, *278*, 115505. [CrossRef]
16. Guo, M.; Choi, Y.; Cheong, S.-M.; O'Neill, Z. Current and future residential electricity demand using large-scale smart meter data in a changing climate. *Sustain. Cities Soc.* **2025**, *130*, 106623. [CrossRef]
17. Li, X.; Zhou, Y.; Yu, S.; Jia, G.; Li, H.; Li, W. Urban heat island impacts on building energy consumption: A review of approaches and findings. *Energy* **2019**, *174*, 407–419. [CrossRef]
18. Santamouris, M.; Papanikolaou, N.; Livada, I.; Koronakis, I.; Georgakis, C.; Argiriou, A.; Assimakopoulos, D.N. On the impact of urban climate on the energy consumption of buildings. *Sol. Energy* **2001**, *70*, 201–216. [CrossRef]
19. Hirano, Y.; Fujita, T. Evaluation of the impact of the urban heat island on residential and commercial energy consumption in Tokyo. *Energy* **2012**, *37*, 371–383. [CrossRef]
20. Miller, C.; Arjunan, P.; Kathirgamanathan, A.; Fu, C.; Roth, J.; Park, J.Y.; Balbach, C.; Gowri, K.; Nagy, Z.; Fontanini, A.D.; et al. The ASHRAE Great Energy Predictor III competition: Overview and results. *Sci. Technol. Built Environ.* **2020**, *26*, 1427–1447. [CrossRef]
21. Kreider, J.F.; Haberl, J.S. Predicting hourly building energy use: The Great Energy Predictor Shootout—Overview and discussion of results. *ASHRAE Trans.* **1994**, *100*, 1104–1118.

22. Peplinski, M.; Dilkina, B.; Chen, M.; Silva, S.J.; Ban-Weiss, G.; Sanders, K.T. A machine learning framework to estimate residential electricity demand based on smart meter electricity, climate, building characteristics, and socioeconomic datasets. *Appl. Energy* **2024**, *357*, 122413. [CrossRef]
23. Zhang, S.; Guo, Q.; Smyth, R.; Yao, Y. Extreme temperatures and residential electricity consumption: Evidence from Chinese households. *Energy Econ.* **2022**, *107*, 105890. [CrossRef]
24. Wang, Y.; Jin, A.S.; Sanders, K.T. A systematic review of literature utilizing residential smart meter data. *Renew. Sustain. Energy Rev.* **2026**, *225*, 116130. [CrossRef]
25. Homaei, S.; Hamdy, M. Thermal resilient buildings: How to be quantified? A novel benchmarking framework and labelling metric. *Build. Environ.* **2021**, *201*, 108022. [CrossRef]
26. Attia, S.; Levinson, R.; Ndongo, E.; Holzer, P.; Kazanci, O.B.; Homaei, S.; Zhang, C.; Olesen, B.W.; Qi, D.; Hamdy, M.; et al. Resilient cooling of buildings to protect against heat waves and power outages: Key concepts and definition. *Energy Build.* **2021**, *239*, 110869. [CrossRef]
27. Sengupta, A.; Al Assaad, D.; Bastero, J.B.; Steeman, M.; Breesch, H. Impact of heatwaves and system shocks on a nearly zero energy educational building: Is it resilient to overheating? *Build. Environ.* **2023**, *234*, 110152. [CrossRef]
28. Sheng, M.; Reiner, M.; Sun, K.; Hong, T. Assessing thermal resilience of an assisted living facility during heat waves and cold snaps with power outages. *Build. Environ.* **2023**, *230*, 110001. [CrossRef]
29. Wang, Z.; Hong, T.; Li, H. Informing the planning of rotating power outages in heat waves through data analytics of connected smart thermostats for residential buildings. *Environ. Res. Lett.* **2021**, *16*, 074003. [CrossRef]
30. Pickering, E.M.; Hossain, M.A.; French, R.H.; Abramson, A.R. Building electricity consumption: Data analytics of building operations with classical time series decomposition and case-based subsetting. *Energy Build.* **2018**, *177*, 184–196. [CrossRef]
31. Kim, M.K.; Kim, Y.-S.; Srebric, J. Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. *Sustain. Cities Soc.* **2020**, *62*, 102385. [CrossRef]
32. Gui, X.; Gou, Z.; Lu, Y. Reducing university energy use beyond energy retrofitting: The academic calendar impacts. *Energy Build.* **2021**, *231*, 110647. [CrossRef]
33. Gaspar, K.; Gangolells, M.; Casals, M.; Pujadas, P.; Forcada, N.; Macarulla, M.; Tejedor, B. Assessing the impact of the COVID-19 lockdown on the energy consumption of university buildings. *Energy Build.* **2022**, *257*, 111783. [CrossRef] [PubMed]
34. Zhao, X.; Yin, Y.; Zhang, S.; Xu, G. Data-driven prediction of energy consumption of district cooling systems based on weather forecast data. *Sustain. Cities Soc.* **2023**, *90*, 104382. [CrossRef]
35. Raymand, F.; Najafi, B.; Haghighat Mamaghani, A.; Moazami, A.; Rinaldi, F. Machine learning-based estimation of building characteristics employing electrical and chilled-water consumption data: Pipeline optimisation. *Energy Build.* **2023**, *295*, 113327. [CrossRef]
36. Fu, H.; Lee, S.; Baltazar, J.-C.; Claridge, D.E. Occupancy analysis in commercial building cooling energy modelling with domestic water and electricity consumption. *Energy Build.* **2021**, *253*, 111534. [CrossRef]
37. Pickering, E.M.; Hossain, M.A.; Mousseau, J.P.; Swanson, R.A.; French, R.H.; Abramson, A.R. A cross-sectional study of the temporal evolution of electricity consumption of six commercial buildings. *PLoS ONE* **2017**, *12*, e0187129. [CrossRef] [PubMed]
38. Ahn, K.U.; Shin, H.S.; Park, C.S. Energy analysis of 4625 office buildings in South Korea. *Energies* **2019**, *12*, 1114. [CrossRef]
39. Granderson, J.; Price, P.N. Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models. *Energy* **2014**, *66*, 981–990. [CrossRef]
40. Granderson, J.; Price, P.N.; Jump, D.; Addy, N.; Sohn, M.D. Automated measurement and verification: Performance of public-domain whole-building electric baseline models. *Appl. Energy* **2015**, *144*, 106–113. [CrossRef]
41. ASHRAE. *Guideline 14-2023: Measurement of Energy, Demand, and Water Savings*; ASHRAE: Atlanta, GA, USA, 2023.
42. Efficiency Valuation Organization. *International Performance Measurement and Verification Protocol (IPMVP): Core Concepts*; EVO: Washington, DC, USA, 2022.
43. Kissock, J.K.; Haberl, J.S.; Claridge, D.E. Inverse modeling toolkit: Numerical algorithms. *ASHRAE Trans.* **2003**, *109*, 425–434.
44. Liang, X.; Hong, T.; Shen, G.Q. Improving the accuracy of energy baseline models for commercial buildings with occupancy data. *Appl. Energy* **2016**, *179*, 247–260. [CrossRef]
45. Ding, Y.; Wang, Q.; Wang, Z.; Han, S.; Zhu, N. An occupancy-based model for building electricity consumption prediction: A case study of three campus buildings in Tianjin. *Energy Build.* **2019**, *202*, 109412. [CrossRef]
46. Kim, H.G.; Lee, S.E.; Kim, D.W. Impact of calendarization on change-point models. *Energy Build.* **2024**, *303*, 113803. [CrossRef]
47. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
48. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [CrossRef]
49. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol. Bull.* **1993**, *114*, 494–509. [CrossRef]
50. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]

https://doi.org/10.3390/su18168289

51. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
52. Cihangir, C. Meter-Aware Interpretation of Extreme-Heat Load Response in Campus Buildings: The Role of Academic Calendar and Chilled-Water Metering [Dataset]. Zenodo. 2026. Available online: https://about.zenodo.org/contact/ (accessed on 29 July 2026). [CrossRef]