forecasting

MDPI

Article

Does Synthetic Data Help Building Energy Analytics with Short Metering Histories? A Matched-Origin Evaluation with Structure-Free Controls and Simple Baselines

Pedro Martins [1,*], Tiago Esteves [2], Paulo Váz [3], José Silva [3], Filipe Madeira [2,4] and Maryam Abbasi [2,4,*]

[1] Coimbra Business School (ISCAC), Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal
[2] School of Management and Technology, Santarém Polytechnic University, 2001-904 Santarém, Portugal; 170100194@esg.ipsantarem.pt (T.E.); filipe.madeira@esg.ipsantarem.pt (F.M.)
[3] School of Technology and Management of Viseu, Polytechnic University of Viseu, Campus Politécnico, 3504-510 Viseu, Portugal; paulovaz@estgv.ipv.pt (P.V.); jsilva@estgv.ipv.pt (J.S.)
[4] Research Center for Arts and Communication (CIAC), Santarém Polytechnic University, 2001-904 Santarém, Portugal
* Correspondence: pmmartins@iscac.pt (P.M.); maryam.abbasi@esg.ipsantarem.pt (M.A.)

Highlights

What are the main findings?

- With a matched forecast origin, the reduction in Prophet's forecasting error obtained from a synthetic supplement fell from 205.1 to 51.6 percentage points at 5% real history; no statistically supported difference from the reduction obtained with structure-free noise was detected, although equivalence was established in only 4 of 18 comparisons.
- A time-of-week profile fitted on the scarce real data alone was more accurate than every supplemented model at all five forecast origins.

What are the implications of the main findings?

- Gains attributed to synthetic supplements can arise from the evaluation design and from regularisation of the model, so they should be tested with matched forecast origins, structure-free controls and simple baselines fitted on real data.
- In anomaly detection, a supplement can lower the F1-score through the decision threshold without lowering average precision, so its usefulness has to be established separately for each task and model.

Abstract

Synthetic data are often used to complete short building metering histories. We re-evaluate a transparent hour-of-week, temperature and AR(1) generator, calibrated on held-out buildings, completing 5–20% real training windows of 82 education buildings from the Building Data Genome Project 2 over 20 random splits. With the real window over a year before the test period, the supplement cut Prophet's median paired error by 205.1 percentage points at 5%. With a matched forecast origin the reduction fell to 51.6, 15.1 and 3.1 points at 5, 10 and 20%, reversed over the first week, and did not differ detectably from white-noise, dispersion-matched or fully permuted complements (equivalence within ±3 points in only 4 of 18 comparisons). Real-only Prophet improved sharply when the same calendar weeks moved a year nearer the test period; season effects were imprecise. Flat-trend Prophet and a time-of-week profile on real data alone beat every supplemented model at five origins, suggesting the supplement mainly regularises Prophet's trend. Supplements lowered Isolation Forest F1 through the contamination threshold, not average precision, and gave a profile-residual detector no supported advantage over regularising it without added

Check for updates

Academic Editor: Alessandro Niccolai

Received: 8 October 2026
Revised: 18 November 2026
Accepted: 19 November 2026
Published: 21 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

data. Supplements should be tested with matched origins, structure-free controls, simple baselines and regularised models, per task and model.

Keywords: synthetic data; data augmentation; building energy; load forecasting; anomaly detection; data scarcity; forecast evaluation; prophet; isolation forest; BDG2

1. Introduction

The buildings and construction sector accounts for about 32% of global energy use and 34% of global CO_2 emissions [1]. Much of that energy is used by existing buildings whose operation is increasingly managed through metered data. Two analytical tasks sit at the centre of that monitoring. Load forecasting supports scheduling, demand response and measurement and verification (M&V) of savings [2–4]; anomaly detection flags faults, waste and abnormal use before they become expensive [5,6]. Both are usually trained on the building's own metered history, and that history is often short. Meters are installed after a renovation, data platforms are commissioned midway through a year, communication failures leave long holes, and a new building has no past at all [7,8]. The question practitioners then face is not whether data exist, but whether a few weeks of data are enough, and what to do when they are not.

A popular answer is to augment the scarce history with synthetic observations [9–11]. Generators range from simple perturbations and statistical models to physical simulation and deep generative networks conditioned on building metadata [12–15]. In a large comparison on the Building Data Genome Project 2 (BDG2), data generation was the most effective of four strategies under the most severe data shortage [7]. The way such gains are measured deserves scrutiny, however. Synthetic series are still most often judged by how closely they resemble real data, not by what they do for the downstream task [11,16], and when downstream effects are measured, three design choices can make a supplement look useful for reasons unrelated to its content.

The first is the forecast origin. If a model trained on a short real window is compared with a model whose training set has been filled with synthetic data up to the start of the test period, the two models forecast from different origins and over different horizons. Any difference then mixes the value of the added data with the value of being closer to the target period [17,18], and with whatever else differs between an old and a recent window, such as its season and consumption level. The second is the absence of a structure-free control: without a supplement that restores the length of the training window but carries no temporal information, one cannot say whether the generator's structure matters. The third is the absence of simple baselines. In forecasting competitions and in building energy practice, seasonal naive forecasts and time-of-week regressions are hard to beat [19–21], and a sophisticated pipeline that does not beat them offers no evidence that it is needed.

This paper re-examines a transparent statistical–temporal generator (hour-of-week profile, outdoor temperature and an AR(1) residual) for supplementing short histories of education buildings in BDG2. The generator is calibrated only on held-out calibration buildings and is used to complete the training window of separate evaluation buildings, whose testing always relies on real data. An earlier, unpublished evaluation of the same generator placed the scarce real window at the start of the training period (a leading-window design) and found that synthetic data cut Prophet's forecasting error dramatically, that a white-noise supplement gave a similar reduction, and that the same supplement degraded Isolation Forest anomaly detection. That design confounded the supplement with a change of forecast origin, and neither task had simple baselines or complete controls.

Here we re-run the full pipeline from the public data over twenty random building splits and add a matched-origin design, a seasonally matched lagged window, four further calendar origins, two further structure-free controls (dispersion-matched noise and a permuted synthetic complement), a partly structured control (an hour-of-week-permuted complement), horizon-resolved errors, three simple baselines, flat-trend, changepoint and temperature-free Prophet variants, a profile-residual anomaly detector with real-only variants, threshold-free detection metrics and site-clustered inference.

The research questions are:

RQ1. Does supplementing a short real history with synthetic data reduce forecasting error once all scenarios share the same forecast origin?

RQ2. Does the generator's temporal structure add anything beyond restoring the length of the training window, relative to structure-free controls?

RQ3. How do the supplemented models compare with simple baselines that use only the short real window?

RQ4. Does the same supplement help or harm unsupervised anomaly detection, and does the answer depend on the detector?

None of the ingredients of this evaluation is new: matched forecast origins, structure-free controls, naive benchmarks, threshold-free metrics and cluster-aware inference are established practice. The contribution is to combine them in one controlled benchmark on public data and to show how much each of them changes the conclusion about a synthetic supplement. Four findings result. First, we show that most of the forecasting gain observed under the leading-window design is associated with the information available at the forecast origin rather than with the supplement: under a matched origin the median paired gain shrinks from 205.1 to 51.6 pp at 5% and from 65.4 to 3.1 pp at 20%, reverses over the first week, and in a seasonally matched lagged window real-only Prophet improved markedly when the same calendar weeks were moved one year closer to the test period, whereas the contrasts between seasons were too imprecise to size any seasonal contribution. Second, no statistically supported advantage of the generator over white noise, dispersion-matched noise or a fully permuted synthetic complement was detected at the primary origin, nor over white noise at four further origins (an absence of evidence rather than a demonstration of equivalence, which was established for only 4 of the 18 comparisons at the primary origin), and a Prophet model with a flat trend trained on the scarce real data alone is more accurate than any supplemented model; flat-trend, changepoint and regularisation checks are consistent with the supplement acting mainly as a regulariser of Prophet's trend rather than as a source of information. Third, we show that a time-of-week profile computed from a few weeks of real data outperforms every Prophet variant fitted on the scarce window at every origin, and at the primary origin also Prophet trained on the full history, without any temperature information. Fourth, we show that the anomaly-detection result depends on the detector and on how it sets its threshold: supplements lower the F1-score of Isolation Forest by moving its contamination threshold, not by degrading its ranking of anomalies, and a real-only profile-residual detector with a regularised scale and a cross-fitted threshold reaches a mean F1 at least as high as that of every supplemented detector in the main anomaly setting of the matched-origin design. From these results we derive an evaluation protocol for synthetic data in building energy analytics.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the data, generator, evaluation designs, models and statistics. Section 4 reports the results, Section 5 discusses their implications and limitations, and Section 6 concludes.

https://doi.org/10.3390/forecast8060113

2. Related Work

2.1. Building Load Forecasting with Short Histories

Building load forecasting spans engineering simulation, statistical time-series models, classical machine learning and deep architectures [2,3,22]. Model complexity does not guarantee accuracy: a support vector regression outperformed a Transformer on a public building [23], and a comparison on 1045 monthly series of the M3 competition found that standard statistical methods were more accurate than a range of machine learning methods [19]. In M&V practice, time-of-week-and-temperature (TOWT) regressions remain reference baselines because they capture the dominant weekly occupancy cycle with few parameters [4,21]. Granderson et al. [4] reported median CV(RMSE) below 25% for all ten tested baseline models when twelve months of training data were available, and the error grows when the training window shrinks. Prophet [24] decomposes a series into a piecewise-linear trend, Fourier seasonalities and regressors; its trend component is known to extrapolate poorly when fitted to short windows, a point that becomes central in our results.

Several strategies address short histories. Transfer learning reuses models from similar source buildings but depends on reliable similarity measures and suffers from negative transfer when buildings differ [8,25,26]. Incremental learning assumes that new real data keep arriving. Physics-informed methods require thermal and operational metadata that open datasets rarely provide. Li et al. [7] compared the four families on BDG2 and found data generation best under extreme scarcity; their study did not include a supplement without temporal structure, so the source of the advantage was not isolated.

2.2. Synthetic Energy Time Series and Their Evaluation

Generation methods include perturbation of real series (jittering, scaling, window warping) [9,10], statistical and autoregressive models with controllable characteristics [11,27], physical simulation [12], decomposition-based augmentation [15], generative adversarial networks [13] and conditional diffusion models trained on BDG2 meters and metadata [14]. Reviews agree that evaluation is the weak point. Turowski et al. [11] found that most evaluations of synthetic energy series measure statistical fidelity rather than usefulness, and Stenger et al. [16] place train-on-synthetic, test-on-real utility at the centre of their taxonomy. Fidelity and utility can diverge in both directions: a supplement can look realistic and still not help, or look crude and still help by restoring something the model needs, such as the extent of the training window.

2.3. Anomaly Detection in Building Energy Data

Labelled anomalies are scarce, so unsupervised detectors dominate [5,6,28]. Isolation Forest isolates rare points with few random splits, scales to large data and is trained on presumed-normal data only [29]; autoencoders and density methods are common alternatives [30]. Benchmark studies warn that detector rankings are fragile and that simple rules frequently rival complex models on point anomalies [31–33]. Synthetic data enter anomaly detection mostly as synthetic faults: Han and Chong combine simulation and conditional GANs to generate fault examples [34], and Tian et al. oversample rare anomalous patterns [35]. Reinforcing normal behaviour, which is what happens when a forecasting augmentation is reused for detection, has received less attention.

2.4. Evaluation Pitfalls

Out-of-sample evaluation of forecasts requires a fixed origin and horizon for all compared methods [17,18], scale-free accuracy measures [36], and benchmarks against naive methods [20]. In machine learning more broadly, failure to compare against tuned

simple baselines has repeatedly produced illusory progress [37,38]. Table 1 positions the present study against these strands. Relative to prior work on synthetic data for buildings, it combines a matched forecast origin, several structure-free controls, simple baselines and a cross-task comparison under the same supplement.

Table 1. Positioning of this study relative to representative prior work.

Line of Work	Typical Requirement or Gap	This Study
Transfer learning [8,25,26]	Similar source buildings and similarity metrics; risk of negative transfer	No model is fitted on source buildings, but information from held-out calibration buildings is transferred through the generator's hour-of-week profile and residual scale
Data generation on BDG2 [7]	No structure-free control; origin of the gain not isolated	White-noise, dispersion-matched and permuted controls; matched origin, lagged-window and multiple-origin diagnostics
Deep generative models [13,14]	Large calibration corpora and metadata; harder to audit	Three-component generator with known structure
Fault injection for detection [34,35]	Synthetic anomalies added	Synthetic normal behaviour added; two detectors
Synthetic-data evaluation surveys [11,16]	Fidelity measured more often than utility	Utility measured on two tasks against controls and baselines
Forecast evaluation practice [17,20]	Fixed origin, naive benchmarks	Both designs reported; three simple baselines and a flat-trend variant

3. Materials and Methods

Figure 1 summarises the pipeline. For each random building split, buildings are selected and divided into calibration and evaluation sets, the generator is calibrated on the calibration buildings, and the evaluation buildings are used for the forecasting and anomaly-detection experiments under the evaluation designs described in Section 3.4. Results are then pooled across splits for statistical analysis. Each step is described below.

3.1. Data

We use the Building Data Genome Project 2 (BDG2) [39], an open collection of hourly readings from 3053 energy meters in 1636 non-residential buildings across 19 sites in North America and Europe, covering 2016 and 2017 (17,544 hourly records per meter). Each building identifier has the form `site_use_name` (for example `Fox_education_Nilda`), which encodes the site and the primary use and allows each electricity series to be joined with the weather of its site. Two files are used: `electricity_cleaned.csv`, with hourly electricity consumption in kWh per building, and `weather.csv`, from which only outdoor air temperature is retained, because it is the weather variable most directly linked to consumption and is available for every site. The electricity file is the cleaned version distributed by the BDG2 authors, in which zero readings of electricity meters, runs of zero readings longer than 24 h and outliers flagged by an anomaly-detection routine had already been removed (set to missing) over the full two-year series [39]. We applied no further outlier treatment. Because this upstream cleaning used the whole series, the retained values carry a small amount of information from outside any training window; this affects which hours are present but not their values, and it is shared by all scenarios. The exact files used are identified by their SHA-256 checksums in the Supplementary Materials.

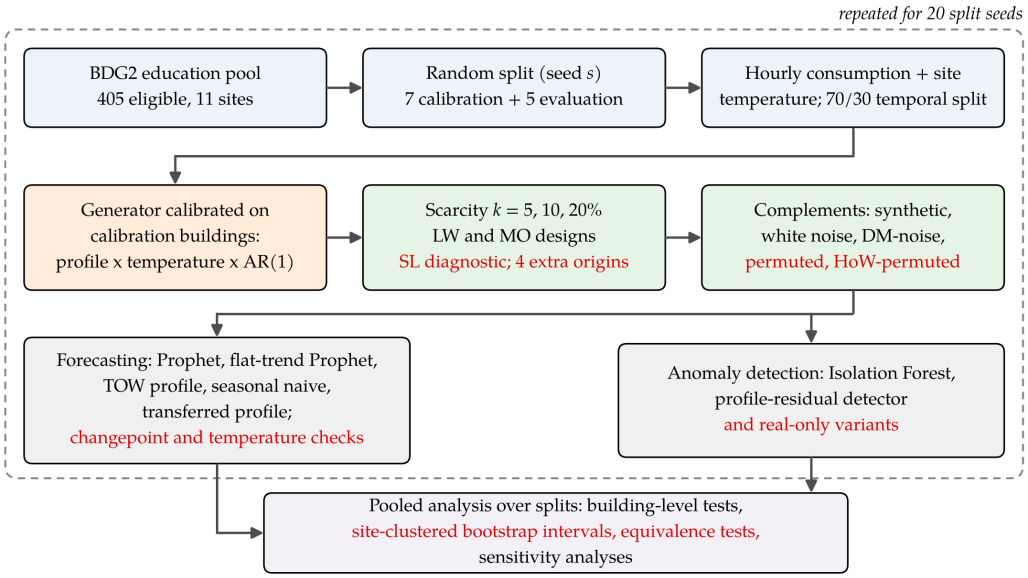


Figure 1. Experimental pipeline. Building selection, calibration of the generator, construction of the scarcity scenarios and both downstream tasks are repeated for each of the 20 random building splits (dashed box); statistics are computed on the pooled paired results. LW: leading window; MO: matched origin; SL: seasonally matched lagged window; DM-noise: dispersion-matched noise; HoW: hour of week; TOW: time of week.

The analysis is restricted to the education typology to reduce functional heterogeneity: school and university buildings share a strong weekday and daytime occupancy pattern, which makes an hour-of-week profile a meaningful modelling primitive. Table 2 summarises the filtering funnel.

Table 2. Dataset and filtering summary (BDG2 electricity meters, education typology).

Item	Value
Hourly period	1 January 2016 to 31 December 2017
Records per meter	17,544
Electricity meters in `electricity_cleaned.csv`	1578
with less than 5% missing values	1019
of the education typology	405 (11 sites)
Buildings per split	12 (7 calibration + 5 evaluation)
Valid split seeds	20 (Section 3.2)
Short-gap interpolation limit	6 h
Post-interpolation missing threshold	2%
Rows removed for missing temperature or consumption	0.61 to 0.95% per split
Train/test split per building	first 70% / last 30% of rows (temporal)
Training window (median)	12,253 h (≈511 days, January 2016 to May 2017)
Test window (median)	5251 h (≈219 days, May to December 2017)
First test hour (forecast origin), range	19 May to 28 May 2017

3.2. Building Selection and Temporal Split

For each split seed, buildings are drawn from the eligible education buildings so as to diversify sites: one building per site is sampled first, the sample is completed to twelve buildings, shuffled with the seed and partitioned into seven calibration and five evaluation buildings. The preprocessing steps are applied in the following order. (i) Gaps of up to six hours in the two-year consumption series of each sampled building are filled by linear interpolation in time; (ii) a building is kept only if fewer than 2% of its values remain missing, and a split in which a sampled building fails this check is treated as

invalid; (iii) the consumption series is joined with the site temperature and hours without consumption or temperature are removed; (iv) each building's series is split temporally into the first 70% of its remaining rows for training and the last 30% for testing; (v) the scarce windows of Section 3.4 are cut from the training rows. Because interpolation precedes steps (iv) and (v), a value interpolated at the edge of a window can, in principle, use an endpoint outside it. We audited this: of the 1,660 interpolated values that fall inside the scarce real windows, only 1 lies in a gap that extends beyond the window, so the effect on the results is negligible. The same reasoning applies at the boundary between training and testing: a gap that straddles it is filled with a test-period endpoint, but the six-hour limit confines any such value to the last six training hours of a building (at most 0.05% of its training rows). Removing hours in step (iii) leaves an irregular time index; every time-dependent operation therefore uses timestamps rather than row positions: Prophet is fitted on the hourly calendar, horizons are elapsed time from the forecast origin, hour-of-week profiles are indexed by clock time, the seasonal naive forecast uses the last 168 clock hours, and the lags of the anomaly detector are taken at $t - 1\,\mathrm{h}$ and $t - 24\,\mathrm{h}$ (Section 3.8).

With BDG2's two-year coverage the temporal split places the training window between January 2016 and late May 2017 (about 511 days) and the test window between late May and December 2017 (about 219 days); the first test hour of the evaluation buildings ranges from 19 May 2017 to 28 May 2017. Calibration and evaluation buildings never overlap within a split. The generator's parameters are estimated without any consumption of the evaluation buildings; evaluation-building consumption enters the supplement only through its scale, which is computed from the real observations available in each scenario (Section 3.3). Because the 70% cut is made per building, the training period of a calibration building can end slightly after the earliest forecast origin among the evaluation buildings of the same split: by at most 211 h, which concerns at most 1.4% of the calibration rows. This bounds the amount of later data used in calibration, but not the resulting change in the calibrated profile, which we did not quantify; we therefore do not claim that the overlap leaves the profile unchanged. The overlap is removed entirely in the multiple-origin analysis of Section 3.7, where the generator is recalibrated on data strictly before each origin, and the main qualitative findings hold there (Section 4.3).

The earlier evaluation of this generator used five seeds (42, 100, 200, 400 and 500). To reduce the dependence of the conclusions on a few draws, we extend this to twenty seeds by adding the sequence 600, 700, … in steps of 100. Three candidate seeds of the sequence (700, 1100 and 1500) drew a building that failed the 2% check and were replaced by the next seeds (2100, 2200 and 2300), so the twenty valid seeds are 42, 100, 200, 400, 500, 600, 800, 900, 1000, 1200, 1300, 1400, 1600, 1700, 1800, 1900, 2000, 2100, 2200 and 2300. With five evaluation buildings per split this gives 100 building–split cases.

3.3. Statistical–Temporal Generator

The generator produces normalised synthetic consumption for each evaluation building from parameters estimated only on the training portion of the calibration buildings (Algorithm 1). Its three multiplicative components are fixed in advance. Let $P(h_t, d_t)$ be the mean normalised consumption of the calibration buildings for hour h_t and weekday d_t, smoothed by a centred moving average of width three, $\mathrm{Temp}(t)$ the site's outdoor temperature, and $\varphi = 0.85$ a fixed AR(1) coefficient chosen to inject hourly persistence without tuning it to the calibration buildings. With innovations $\eta_t \sim \mathcal{N}\big(0, \sigma_{\mathrm{res}}^2(1-\varphi^2)\big)$,

$$e_0 = \eta_0, \qquad e_t = \varphi\, e_{t-1} + \eta_t \quad (t > 0), \tag{1}$$

$$R(t) = \max(1 + e_t,\, 0.1), \qquad T(t) = \max\big(\beta\,\mathrm{Temp}(t) + \alpha,\, 0.1\big), \tag{2}$$

$$C_{\mathrm{syn}}(t) = P(h_t, d_t)\, T(t)\, R(t). \tag{3}$$

The stationary standard deviation of e_t is σ_{res}. Because the process starts from $e_0 = \eta_0$ rather than from the stationary distribution, $\text{Var}(e_t) = \sigma^2_{\text{res}}\left(1 - \varphi^{2(t+1)}\right)$; with $\varphi = 0.85$ the variance is within 1% of its stationary value after 14 hours. No burn-in is used, because this transient affects only the first hours of series that are about 17,000 hours long and lie at the beginning of the training period. The thermal coefficients (α, β) come from a least-squares fit of the profile-corrected normalised consumption on temperature, with β clipped to $[-0.05, 0.05]$ so that temperature cannot dominate the profile; σ_{res} is the standard deviation of the multiplicative residual left after the profile and temperature terms. Each evaluation building receives a freshly simulated AR(1) residual. The normalised series is converted to the building's scale in each scenario using only the real data available in that scenario: the scale s_k is the median, over the 24 h of the day, of the hourly medians of the available real window. Because consumption is right-skewed, this median-based scale is slightly lower than the mean of the real window (median ratio 0.987 in the matched-origin design at 5%), and the mean of the scaled complement is correspondingly slightly below that of the real data it extends (median ratio 0.985); the synthetic series is normalised to a mean close to one only before this scaling (Section 4.1).

Algorithm 1. Statistical–temporal generator: calibrate on calibration buildings, then generate for each evaluation building.

Input: training data of the calibration buildings; evaluation buildings; site temperatures
Output: normalised synthetic series for each evaluation building

Normalise each calibration series by its training mean
$P \leftarrow$ mean normalised value per (hour, weekday), smoothed with a centred window of width 3
$(\alpha, \beta) \leftarrow$ least-squares fit of (normalised consumption$/P$) on temperature; clip β to $[-0.05, 0.05]$
$\sigma_{\text{res}} \leftarrow$ standard deviation of the multiplicative residual after P and T
foreach *evaluation building* **do**
    $T(t) \leftarrow \max(\beta\,\text{Temp}(t) + \alpha,\ 0.1)$
    simulate e_t with Equation (1) and $\varphi = 0.85$
    $C_{\text{syn}}(t) \leftarrow P(h_t, d_t)\,T(t)\ \max(1 + e_t, 0.1)$

3.4. Scarcity Levels and Evaluation Designs

Scarcity is simulated by keeping $k \in \{5, 10, 20\}\%$ of the training rows of each evaluation building, which corresponds to roughly 25, 51 and 102 days of hourly data; a 100% scenario serves as the full-history reference. Two evaluation designs place the scarce window differently (Figure 2).

Leading-window (LW) design. The real history is the first k% of the training window. For the real-only scenario the last observation is therefore a median of 409 to 486 days before the first test hour, depending on k, while the supplemented scenarios are filled up to the test boundary. The design mimics a building whose history is short but old, and it confounds the supplement with a change of forecast origin. It also compares unequal information sets: the complement of a supplemented model is built from the calibration buildings' consumption and from site temperatures over the whole training period, up to the test boundary. Those records would be available to an analyst forecasting from the test boundary, which is the origin of the supplemented models, but not at the end of the leading window, where the real-only model stops; the real-only model uses nothing dated after its window except the test-period temperature supplied as a regressor.

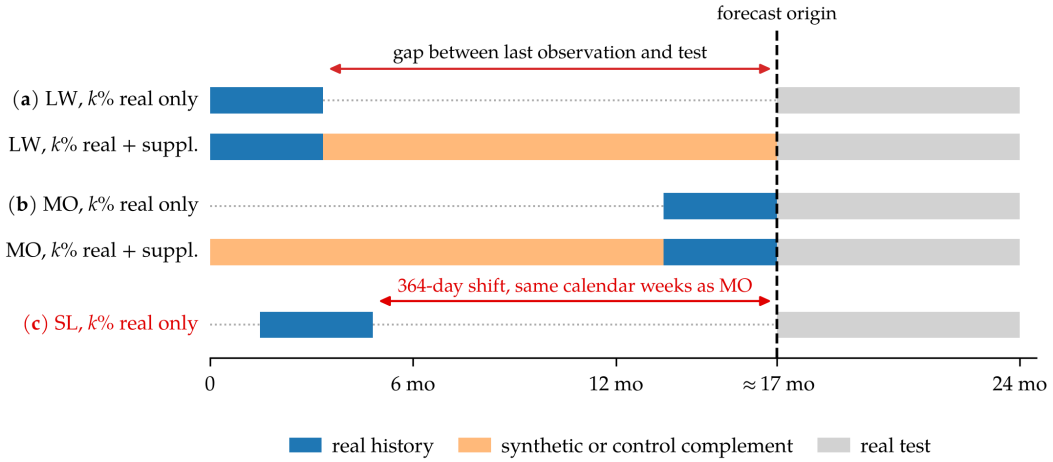


Figure 2. Timeline of the evaluation designs for a $k = 20\%$ scarcity level (drawn to scale on the two-year BDG2 series). (**a**) Leading-window (LW) design: the real history is the first k% of the training window, so the real-only model is fitted far from the test period, while the supplemented models are filled up to the test boundary. (**b**) Matched-origin (MO) design: the real history is the last k% of the training window (a recently metered building) and the complement fills the earlier timestamps, so every scenario shares the same forecast origin and the same test period. (**c**) Seasonally matched lagged-window (SL) diagnostic: the real-only models are fitted on the calendar weeks of the MO window one year (364 days) earlier, which contrasts the season of the window with its recency.

Matched-origin (MO) design. The real history is the last k% of the training window, which mimics a recently metered building, and the complement fills the earlier timestamps. Every scenario, real-only or supplemented, ends at the same hour and forecasts the same test period, so within this design differences between scenarios can be attributed to the training data alone. It is the design we treat as primary. A common forecast origin is not the same as identical data availability. The supplemented scenarios and the transferred profile also draw on the calibration buildings, and the supplemented scenarios on site temperatures over the filled timestamps; because the 70% cut is made per building, the training period of a calibration building can end up to 211 h after the origin of an evaluation building (Section 3.2). In the main analysis all methods therefore share the origin and the test period, but not exactly the same records. Only at the additional origins of Section 3.7 does no method use any record dated after the origin.

Moving from the LW to the MO design changes more than the forecast origin of the real-only model: the real window also moves from January 2016 to the weeks before the test period, so its season, recency, consumption level and the scale s_k derived from it all change together with the extrapolation distance. Comparisons between the two designs are therefore contrasts between two information settings, not estimates of a pure origin effect. To examine these factors, we add two diagnostics. The seasonally matched lagged-window (SL) diagnostic fits the real-only models on the same calendar weeks as the MO window, shifted back by 364 days (52 weeks, which keeps the weekday alignment); relative to MO it changes recency and extrapolation distance (by about one year), together with any change in the building's consumption level between the two years, which it cannot separate from them, and relative to LW it changes mainly the season of the window. The multiple-origin analysis (Section 3.7) repeats the matched-origin comparison at four further calendar origins. The two designs are thus reported as distinct scenarios, and all inference about the effect of the supplement is drawn within the matched-origin design and its additional origins, where every method shares the forecast origin and the test period (with the qualification on data availability given above); the leading-window results show what that protocol yields and are not used to estimate the value of augmentation.

Forecasting target. In both designs the forecast is an open-loop projection from a fixed origin over the whole test period (about seven months), conditional on the observed site temperature. This corresponds to the baseline projection used in measurement and verification of savings, in which a model fitted before an intervention is projected over the post-intervention period; it is not a forecast updated as new readings arrive. Shorter lead times are examined through horizon-resolved errors (Section 3.9) and through the 90-day test windows of the multiple-origin analysis.

3.5. Training Supplements and Controls

In every supplemented scenario the timestamps of the training window that are not covered by the real fraction are filled with one of five complements (Table 3). The synthetic complement is the generator output scaled as described above. The white-noise control draws independent values from a normal distribution centred on the building's scale with a standard deviation of 30% of that scale, with negative values set to zero; it restores the length of the window but carries no hourly or weekly structure. Because the white noise is less dispersed than the synthetic series, we add a dispersion-matched noise (DM-noise) control that draws independent normal values with the same mean and standard deviation as the scaled synthetic complement of that building and scenario. The moments are matched before negative draws are set to zero; truncation affects on average 1.4% of the draws and changes the mean by less than 0.5% and the standard deviation by less than 1.5%. DM-noise does not, however, reproduce the right skew and the long upper tail of the synthetic series (median skewness 0.69 for the synthetic complement and 0.04 for DM-noise), so it differs from the synthetic complement in more than temporal order. Two further controls address this. The permuted complement is a random permutation of the scaled synthetic complement over its timestamps: it keeps the empirical marginal distribution exactly (including its support, skewness and tails) and removes all temporal structure, namely the calendar profile, the AR(1) autocorrelation and the alignment with temperature. The hour-of-week-permuted complement permutes the synthetic values only within each of the 168 h-of-week cells: it keeps the marginal distribution and the calendar profile and removes the AR(1) autocorrelation and the hour-by-hour alignment between synthetic consumption and temperature. Comparing the synthetic complement with the permuted complement isolates the contribution of temporal structure as a whole, and comparing it with the hour-of-week-permuted complement isolates autocorrelation and load–temperature dependence from the calendar profile. In all supplemented scenarios the real site temperature is used for the filled timestamps, so only the added consumption differs.

Table 3. Training scenarios. The complement fills the training timestamps not covered by the real fraction. Testing always uses the real test period of the building.

Scenario	Real Fraction	Complement
100% real	full window	none (reference)
k% real	5, 10, 20%	none
k% real + synthetic	5, 10, 20%	generator output, Equation (3), scaled with the real fraction
k% real + noise	5, 10, 20%	i.i.d. normal, mean = scale, s.d. = 0.3 × scale, truncated at zero
k% real + DM-noise	5, 10, 20%	i.i.d. normal with the mean and s.d. of the scaled synthetic complement, truncated at zero

Table 3. *Cont.*

Scenario	Real Fraction	Complement
k% real + permuted	5, 10, 20%	random permutation of the scaled synthetic complement
k% real + HoW-permuted	5, 10, 20%	scaled synthetic complement permuted within hour-of-week cells (MO design)

3.6. Forecasting Models and Baselines

Prophet. Prophet [24] is fitted per building and scenario with daily and weekly seasonality, no annual seasonality (the scarce windows are far shorter than a year) and outdoor temperature as an external regressor. Temperature gaps are interpolated over at most six hours, separately within the training data and within the test period, and otherwise imputed with the training mean. All other settings are the library defaults, including the piecewise-linear trend with 25 candidate changepoints placed uniformly over the first 80% of the training history (`changepoint_range` = 0.8) and a changepoint prior scale of 0.05 [40]. The model forecasts every hour of the test period.

Weather information. The temperature regressor is the observed site temperature over the test period. The reported errors are therefore conditional on perfect knowledge of test-period weather (an ex post forecast), as is usual for M&V baselines but not available in operational forecasting, where a weather forecast would be used. Section 4.7 reports a temperature-free sensitivity analysis. The transferred profile described below takes the observed test-period temperature through its thermal term $T(t)$ and is conditional on realised weather in the same way, although that term varies little (Section 4.1); the time-of-week profile and the seasonal naive forecast use no temperature.

Flat-trend Prophet. To test whether the supplements act through the trend component, Prophet is also fitted with a constant trend (`growth = 'flat'`) on the real-only scarce data and, in the MO design, on the real data plus the synthetic complement.

Changepoint placement and regularisation. With the default `changepoint_range`, the candidate changepoints of a supplemented model lie in the first 80% of the concatenated history. In the MO design the real observations occupy only the last 5 to 20% of that history, so the candidates fall in the complement or at its boundary, whereas the real-only model places its candidates within the real observations. Any apparent regularisation by the supplement could therefore stem from where the trend is allowed to change rather than from the added data. We test this in three ways: (i) we record the dates of the default candidate changepoints; (ii) we refit the supplemented models (synthetic and white noise) with the candidate changepoints set explicitly to the dates that Prophet places for the real-only model, that is, inside the real segment; and (iii) we refit the real-only model with smaller changepoint prior scales $\{0.001, 0.005, 0.01, 0.05\}$, choosing the value that minimises the RMSE on the last 168 hours of the real window when the model is fitted to the preceding hours, so that no test data are used in the selection.

Simple baselines. Three baselines use only the real data available in each scenario and require no fitting beyond averaging. The time-of-week profile (TOW) predicts each test hour by the mean consumption of the same hour of the week in the available real window, a temperature-free variant of the TOWT reference model [21]. The seasonal naive forecast repeats the last observed week (168 clock hours) of real data [20]. The transferred profile predicts with the deterministic part of the generator, $P(h_t, d_t)\,T(t)$, rescaled with the building's scale; it uses no information from the evaluation building other than that scale and asks how far the calibration buildings alone can go.

3.7. Additional Forecast Origins

To examine season and lead time with no method using any record dated after the origin, the matched-origin comparison is repeated at four calendar origins common to all buildings: 1 September 2016, 1 December 2016, 1 March 2017 and 1 September 2017. At each origin O, every method uses only real data before O; the generator is recalibrated on the calibration buildings' data before O; the real window is the last 612 or 2450 hours before O (the lengths of the 5% and 20% windows of the main analysis); the complements fill all earlier timestamps from 1 January 2016; and the test period is the 90 days after O, evaluated over the first 7, the first 30 and all 90 days. NRMSE is normalised by the mean real consumption before O. Prophet (real only, real + synthetic, real + noise), flat-trend Prophet (real only), the three simple baselines and Prophet with the full history before O are compared. Together with the main origin (late May 2017, about 219-day test period), this gives five origins spread over the year, two of them at the start of September, one year apart.

3.8. Anomaly Detection

Artificial point anomalies are injected in about 1% of the test hours of each evaluation building (about 52 anomalies per building), each shifting the consumption by ± 3 standard deviations of the building's real consumption over its full training window (the same value in every scenario, scarcity level and design), with the sign chosen at random and values truncated at zero. Locations, signs and magnitudes are generated once per building from a fixed seed, so every detector, scenario, scarcity level and design is evaluated on exactly the same injected test series. Because the cleaned BDG2 electricity data contain no zero readings, no injected spike falls on an already-zero observation; truncation does, however, shorten negative spikes when the consumption is below three standard deviations, which happened for 44% of the negative spikes (22% of all injections). The realised shift therefore averages 2.70 training standard deviations rather than 3.

Isolation Forest. Isolation Forest [29] uses five features: consumption, hour of day, weekday, and the 1-hour and 24-hour lags; consumption and lags are standardised with statistics of the training data only, and the contamination parameter is set to 0.01, with a sensitivity check at 0.008 and 0.012. The lags are taken at clock times $t - 1\,\mathrm{h}$ and $t - 24\,\mathrm{h}$; hours whose lagged values are missing are not scored. (In an earlier implementation the lags were taken as row shifts, which across the 0.6–0.95% of removed hours could pair an hour with a non-adjacent one; all Isolation Forest results reported here use timestamp-based lags, and the change altered per-building F1 by 0.012 on average.) Test features are computed from the contaminated test series, as they would be in operation: an injected spike at hour t also enters the lag features of hours $t + 1$ and $t + 24$, and these hours keep the label "normal", so a detector that flags them is penalised. The first 24 test hours are not scored because their 24-hour lag is unavailable; this excludes on average 0.6 of the about 52 injected anomalies per building. The detector is trained on the real-only window, on the 100% history, and on the real window completed with each of the complements.

Profile-residual detector. As a simple baseline, each training window defines a robust hour-of-week profile: the median m_c and the scaled median absolute deviation $s_c = 1.4826\,\mathrm{MAD}_c$ of each of the 168 hour-of-week cells c. A test hour is flagged when its absolute robust z-score $|x_t - m_{c(t)}|/s_{c(t)}$ exceeds the 0.99 quantile of the training z-scores, the same 1% contamination logic used for Isolation Forest. Quantiles are computed with linear interpolation between order statistics. A cell whose MAD is zero, which can happen when a cell holds only three or four observations, takes the scaled MAD of the whole training window as its scale. In this rule the same observations estimate the profile, the scale and the threshold. Because that in-sample threshold may be what fails when

cells are sparse, we also evaluate three real-only variants that need no supplement: a pooled scale (one scaled MAD of the residuals $x_t - m_{c(t)}$ over all cells), a shrunk scale ($\tilde{s}_c^2 = (n_c s_c^2 + n_0 s_{\text{pool}}^2)/(n_c + n_0)$ with a prior weight $n_0 = 10$ fixed in advance), and a cross-fitted threshold (the training weeks are split into two folds by alternating ISO week; the z-scores of each fold are computed with the profile and scale estimated on the other fold, and the threshold is the 0.99 quantile of these out-of-fold scores), alone and combined with the shrunk scale. The detector is evaluated with the real window and with the synthetic, white-noise, DM-noise and permuted complements. The hour-of-week-permuted complement is not reported for this detector because it leaves every cell's median and MAD, and therefore the detector, unchanged by construction.

Metrics and sensitivity. Precision, recall and F1 are computed per building against the injected labels at the fixed thresholds above. Because F1 depends on the threshold, we also report the average precision (AP), the area under the precision–recall curve of the anomaly score ($-$score for Isolation Forest and the robust z-score for the profile detector), which does not depend on a threshold. In the MO design we vary the anomaly magnitude (2, 3 and 5 training standard deviations) and prevalence (0.5, 1 and 2% of test hours); the profile detector is evaluated on the full 3×3 grid and Isolation Forest on the five settings that vary one factor at a time. Only point anomalies are studied; persistent faults, drifts and collective anomalies are outside the scope of this benchmark.

3.9. Metrics

The primary forecasting metric is the normalised RMSE,

$$\text{NRMSE} = \frac{100}{\bar{c}_{\text{train}}} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(y_t - \hat{y}_t\right)^2}, \tag{4}$$

where $\bar{c}_{\text{train}}$ is the mean real consumption of the building over its full training window. The denominator is fixed for all scenarios of a building and is used only to express the error on a common scale; it does not enter any model. NRMSE in this form is an out-of-sample coefficient of variation of the RMSE. It is related to, but not interchangeable with, the CV(RMSE) of ASHRAE Guideline 14 [41], which is computed on the fit of a baseline model over its own baseline period with a degrees-of-freedom correction and is used as a calibration acceptance criterion; comparisons with that criterion are indicative only. RMSE (kWh) and MAE are recorded as auxiliary metrics. To expose horizon effects, NRMSE is also computed over the first 7 days and the first 30 days of the test period, defined as elapsed clock time (168 h and 720 h) from the forecast origin, not as a number of retained rows.

3.10. Statistical Analysis

All comparisons are paired: two scenarios are compared on the same building of the same split. *Units and hierarchy.* The 100 building–split cases come from 82 distinct evaluation buildings nested in 11 sites (5 to 12 distinct buildings per site); buildings of the same site share weather, operating calendars and often calibration buildings, and 30 of the evaluation buildings serve as calibration buildings in other splits. Repeat appearances of a building are therefore not independent, and neither are buildings of the same site. For each comparison we first average the paired differences of each distinct building over its appearances, and treat the building as the unit of analysis, with sites as clusters.

Estimands. For forecasting, the primary estimand is the median over distinct buildings of the building-level paired difference in NRMSE, Δ, in percentage points (pp); we also report the Hodges–Lehmann pseudomedian (HL), the median of all pairwise averages of the building-level differences, which is the location parameter targeted by the Wilcoxon signed-rank test. The two coincide when the differences are symmetric and diverge when a

minority of buildings has much larger differences, as happens when real-only Prophet fails badly in a few buildings; a significant Wilcoxon test with a median interval that includes zero is then not a contradiction but a statement about different parameters. Differences between marginal medians of two scenarios (reported descriptively in Section 4.2) are a third, different quantity and are not used for inference. For anomaly detection, where many buildings have identical F1, the estimand is the mean building-level difference in F1 or AP.

Uncertainty and tests. Confidence intervals are 95% percentile intervals from a two-stage cluster bootstrap [42] that resamples sites and then buildings within the sampled sites (5000 resamples). With only 11 sites these intervals are approximate and tend to be wide. An interval that excludes zero does not show that the result would survive the removal of each individual site; no leave-one-site-out analysis was performed. The intervals describe the variation expected if sites, and buildings within sites, were drawn again from the population that the sample represents: BDG2 education buildings with nearly complete two-year electricity records (405 buildings on 11 sites). They are not intervals for education buildings in general. They also condition on the calibration fits used in the twenty splits: the calibration buildings drawn and the generator parameters estimated from them are held fixed in the resampling, so the variability that recalibrating the generator would add is not carried into the intervals. The building-level Wilcoxon signed-rank test [43] (p_B) is applied across distinct buildings, and, as a cluster-level sensitivity analysis, the Wilcoxon test is also applied to the 11 site means of the building-level differences (p_S; with 11 sites the smallest attainable two-sided p-value is 0.001). Holm's step-down correction [44] is applied to p_B within pre-specified families: (1) the main forecasting comparisons of both designs (54 tests); (2) the horizon-specific comparisons (30 tests); (3) the mechanism checks of Section 4.5; (4) the multiple-origin comparisons; (5) the anomaly-detection comparisons of F1 and (6) of AP. Per-split Wilcoxon tests combined with Fisher's method [45] are not used for inference: the splits share buildings, sites, weather and calibration buildings, so the chi-square reference distribution is not calibrated. As a purely descriptive consistency measure we report the share of the 20 splits in which the mean paired difference favours the first scenario.

Equivalence. For the comparisons between the synthetic supplement and the structure-free controls, a nonsignificant difference is not taken as evidence of equal performance. Equivalence is declared only when the 90% cluster-bootstrap interval of Δ lies entirely within a margin of $\pm\delta$, which corresponds to two one-sided tests at $\alpha = 0.05$ [46]. For forecasting the margin is $\delta = 3$ pp, fixed before these tests were run. For reference, it is one eighth of the median error of Prophet with the full history (24.0%), less than half of the smallest difference between methods that we treat as practically meaningful in this study (the 6.8 to 8.8 pp by which the time-of-week profile improves on supplemented Prophet in the matched-origin design), and small relative to the 30% hourly tolerance that is often quoted for M&V baselines [41]. Results for $\delta = 2$ and $\delta = 5$ pp are also given. For the F1 of the profile-residual detector we use $\delta = 0.05$, about one ninth of its full-history F1 and less than half of the gain that a complement brings at 5%; for Isolation Forest, whose F1 values are much smaller, $\delta = 0.02$. Outside these explicit tests we state only whether a statistically supported difference was detected. Throughout the paper, expressions such as "no statistically supported difference", "not detectably different" or "gives errors close to" report the absence of evidence for a difference and are not claims of equivalence, which is claimed only where the test above is passed.

3.11. Settings of the Sensitivity Analyses

On the five original splits we repeat the experiments with the AR(1) coefficient set to $\varphi = 0.70$ and $\varphi = 0.95$, which changes the persistence of the synthetic residual and

hence the temporal structure that the generator adds, and with Prophet's changepoint prior scale raised from 0.05 to 0.5, which makes the trend more flexible. The mechanism checks (changepoint placement, validated prior scale, flat trend, temperature-free models), the SL diagnostic, the multiple-origin analysis and the anomaly-detection sensitivity analyses described above use all twenty splits.

3.12. Implementation and Reproducibility

The pipeline is implemented in Python 3.13 with pandas 3.0.5, NumPy 2.5.3, scikit-learn [47] 1.9.1 (Isolation Forest, metrics), statsmodels (autocorrelation), SciPy 1.18.1 (tests) and Prophet 1.5.0 with the CmdStan backend. Seeds are fixed per stage: the split seed s controls building selection, $10000 + s$ the generator, $20000 + s$ the forecasting controls and $30000 + s$ the anomaly injection. The full pipeline was re-executed from the public BDG2 files for this article; Appendix A compares the re-execution with the outputs of the earlier evaluation and reports optimiser diagnostics. The code, an environment lock file, the checksums of the input files, the building and site memberships of every split, the timestamps that delimit every training window, all model settings and the per-building paired results of every experiment are provided as Supplementary Materials.

4. Results

4.1. Sample, Generator Calibration and Fidelity

The twenty splits yield 100 building–split evaluation cases drawn from 82 distinct education buildings on 11 sites; the most frequently drawn building appears in 4 splits. Across splits, 116 distinct buildings served as calibration buildings at least once, and 30 of the 82 evaluation buildings also served as calibration buildings in other splits. Hours removed for missing temperature or consumption amounted to 0.61–0.95% of the rows per split.

Table 4 lists the generator parameters estimated on the calibration buildings of the five original splits and their range over all twenty splits. The thermal slope stays well inside the clipping bound in every split ($|\beta| \leq 0.0095$ against 0.05; the clip was never active). Over the training-period temperatures of the evaluation sites, the temperature term $T(t)$ has a median coefficient of variation of 0.9% across building–split cases (maximum 12%, in split 1800), which is small next to the hour-of-week profile and the AR(1) residual. In practice the generator is therefore a profile multiplied by an AR(1) residual. The residual standard deviation, which sets the dispersion of the synthetic series, varies by a factor of three between splits (0.193 to 0.617), depending on how heterogeneous the drawn calibration buildings are. The 99.9th percentile in Table 4 is that of the normalised training consumption of the calibration buildings; it describes the upper tail of the calibration data and is not a parameter of the generator.

Table 4. Generator parameters estimated on the calibration buildings. Rows give the five original splits and the range over all twenty splits. $\varphi = 0.85$ in all splits. 99.9th pct.: 99.9th percentile of the calibration buildings' training consumption, each series normalised by its training mean (descriptive, not used by the generator).

Split Seed	Thermal Slope β	Intercept α	σ_{res}	99.9th pct.
42	0.00245	0.967	0.311	3.78
100	0.00129	0.984	0.373	5.64
200	0.00012	0.997	0.327	3.78
400	−0.00073	1.009	0.193	2.07
500	−0.00197	1.027	0.402	4.60
All 20 splits (min–max)	−0.0049 to 0.0095	0.881–1.070	0.193–0.617	1.94–5.92

Table 5 and Figure 3 compare the synthetic and real series of the evaluation buildings over the training window, both on the normalised scale. The real series are divided by their own training mean, so their normalised mean is exactly one by construction. The synthetic series has a mean close to, but not exactly, one (0.998 in split 500 and 1.004 over all splits), because the profile is an average of mean-normalised calibration series and the temperature and residual factors have means close to one; the truncations at 0.1 in Equation (2) add a small upward bias. The fidelity comparison therefore precedes the scenario scaling: in the forecasting scenarios the synthetic series is multiplied by the median-based scale s_k, which is 0.99 times the mean of the real window at 5% in the matched-origin design (0.98 at 20%; medians over cases), so the scaled complement sits slightly below the real mean of the window it extends. The average hour-of-week profile is reproduced closely (Figure 3a): weekday daytime peaks, night-time base load and the weekend drop are all present, because they are shared by most education buildings. The individual series are another matter. The point-wise normalised RMSE between a building's synthetic and real series is about 0.50, the Kolmogorov–Smirnov test rejects equality of distributions in 99% of the building–split cases (a descriptive figure only: the test assumes independent observations and is not calibrated for serially dependent hourly series, even on the 1000-point subsamples used here), and the synthetic series are more dispersed than the real ones on average (standard deviation 0.420 vs. 0.358; in split 500, 0.488 vs. 0.328, a 49% increase). The synthetic autocorrelation (Figure 3c) keeps the daily cycle inherited from the profile, but the AR(1) residual adds a smooth short-lag persistence that differs from the real one (mean absolute ACF error 0.283 over 48 lags). These properties are consistent with the generator's intended role as a supplement that conveys the generic shape of an education building, not a replica of the target building.

Table 5. Generator fidelity on the training window of the evaluation buildings, before scenario scaling (real series normalised by their training mean; synthetic series as generated). "Split 500" columns give the mean over its five buildings; the last column gives the mean over all twenty splits with the range of split means in brackets (or real vs. synthetic values). The KS rejection rate is descriptive (Section 4.1).

Metric	Real (Split 500)	Synthetic (Split 500)	All 20 Splits
Normalised mean	1.000	0.998	1.004 [0.991, 1.022]
Normalised standard deviation	0.328	0.488	0.358 vs. 0.420
Normalised 99th percentile	1.705	2.415	1.921 vs. 2.201
Point-wise normalised RMSE	–	0.479	0.500 [0.372, 0.661]
Mean KS statistic	–	0.199	0.254 [0.189, 0.351]
KS rejection rate ($\alpha = 0.05$)	–	100%	99%
Mean abs. ACF error (48 lags)	–	0.229	0.283 [0.159, 0.422]

4.2. Forecasting Under the Leading-Window Design

Table 6 summarises the forecasting error of every scenario and method in both designs, and Figure 4a shows the Prophet scenarios of the leading-window design. This design reproduces the earlier evaluation (Appendix A). Prophet trained on the first 5, 10 or 20% of the training window produced median NRMSE values of 233.0, 168.6 and 97.9%, against 24.0% for the full-history reference. Completing the window with synthetic data brought the median down to 32.3, 33.2 and 32.0%; white noise gave 34.9, 33.9 and 32.9%, and dispersion-matched noise 33.6, 35.3 and 32.7%. The real-only errors are also extremely variable: the split means range from 47% to 1252%, whereas all supplemented scenarios stay within a band of a few percentage points. Under this design every supplement appears to remove 66 to 86% of the median error; the permuted complement behaves like the other controls.

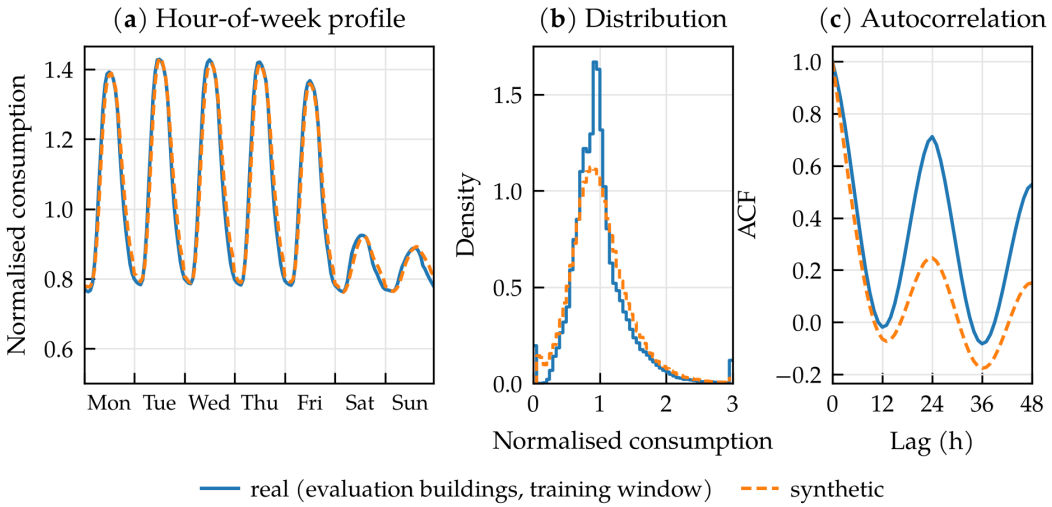


Figure 3. Fidelity of the synthetic series, averaged over the five evaluation buildings of all 20 splits (normalised by each building's training mean). (**a**) Mean hour-of-week profile. (**b**) Density of normalised hourly consumption (values above 3 pooled in the last bin). (**c**) Mean autocorrelation function up to 48 h.

Table 6. Forecasting error (NRMSE, %) over the full test period. For each design the first column gives the median over the 100 building–split cases with the interquartile range in brackets; the second gives the mean over splits of the per-split means ± their standard deviation. Prophet uses default settings; TOW: time-of-week profile; DM: dispersion matched; HoW: hour of week (MO design only). Differences between the medians of two rows are differences between marginal medians and are not the paired estimates of Tables 7 and 8.

	Leading-Window Design		**Matched-Origin Design**	
Method	**Median [IQR]**	**Split Mean ± s.d.**	**Median [IQR]**	**Split Mean ± s.d.**
5% real history				
Prophet, real only	233.0 [78.8, 588.8]	477.0 ± 297.5	78.6 [40.0, 169.2]	121.7 ± 49.1
Prophet, real + synthetic	32.3 [24.3, 44.6]	38.9 ± 9.0	26.2 [20.8, 40.1]	32.5 ± 6.8
Prophet, real + noise	34.9 [20.6, 48.9]	38.6 ± 10.7	31.9 [17.3, 42.4]	32.9 ± 8.0
Prophet, real + DM-noise	33.6 [21.8, 50.1]	38.7 ± 9.8	31.5 [17.2, 42.6]	32.7 ± 8.1
Prophet, real + permuted	33.1 [20.6, 50.3]	38.6 ± 10.2	31.6 [16.9, 42.7]	32.9 ± 8.0
Prophet, real + HoW-permuted	–	–	26.0 [20.0, 38.5]	31.4 ± 7.1
Flat-trend Prophet, real only	28.8 [14.9, 40.9]	37.0 ± 15.4	21.8 [13.0, 36.5]	26.2 ± 8.0
TOW profile	26.9 [15.8, 40.1]	36.0 ± 15.3	20.6 [12.4, 35.9]	24.9 ± 7.6
Seasonal naive	29.1 [16.3, 45.1]	39.0 ± 16.8	21.3 [13.4, 36.5]	27.3 ± 9.1
Transferred profile	28.5 [23.2, 47.1]	37.4 ± 10.0	26.8 [21.6, 40.3]	33.4 ± 8.3
10% real history				
Prophet, real only	168.6 [79.5, 372.3]	444.3 ± 293.3	44.1 [27.9, 112.9]	79.4 ± 39.0
Prophet, real + synthetic	33.2 [24.3, 44.6]	38.5 ± 8.7	28.1 [19.9, 41.2]	33.8 ± 8.5
Prophet, real + noise	33.9 [20.0, 46.5]	37.4 ± 9.6	30.9 [15.9, 42.1]	33.9 ± 9.5
Prophet, real + DM-noise	35.3 [21.3, 46.8]	37.8 ± 9.6	31.5 [16.2, 43.0]	33.9 ± 9.6
Prophet, real + permuted	34.4 [19.4, 46.9]	37.6 ± 9.6	31.2 [16.5, 42.6]	33.8 ± 9.5
Prophet, real + HoW-permuted	–	–	27.3 [19.5, 40.9]	33.3 ± 9.0
Flat-trend Prophet, real only	26.6 [14.7, 38.7]	36.1 ± 16.0	21.3 [13.0, 35.4]	25.8 ± 8.3
TOW profile	24.4 [15.0, 37.2]	34.2 ± 15.0	19.8 [12.2, 35.6]	25.0 ± 7.9
Seasonal naive	28.4 [19.7, 44.7]	32.9 ± 8.8	21.3 [13.4, 36.5]	27.3 ± 9.1
Transferred profile	30.0 [23.3, 47.4]	36.9 ± 9.8	28.3 [21.8, 42.2]	33.5 ± 7.5
20% real history				
Prophet, real only	97.9 [40.3, 194.4]	176.8 ± 106.1	38.1 [17.7, 76.6]	86.6 ± 60.5
Prophet, real + synthetic	32.0 [22.4, 43.0]	36.9 ± 8.0	27.3 [20.0, 39.9]	36.2 ± 10.8
Prophet, real + noise	32.9 [18.9, 45.0]	35.3 ± 9.6	30.3 [17.9, 43.3]	35.8 ± 11.3
Prophet, real + DM-noise	32.7 [20.1, 45.0]	35.6 ± 9.5	30.4 [16.9, 43.1]	35.8 ± 11.3
Prophet, real + permuted	31.7 [20.2, 45.6]	35.5 ± 9.3	30.0 [17.1, 43.0]	35.8 ± 11.3
Prophet, real + HoW-permuted	–	–	26.8 [19.1, 39.6]	35.4 ± 10.6

Table 6. *Cont.*

Method	Leading-Window Design		Matched-Origin Design	
	Median [IQR]	**Split Mean ± s.d.**	**Median [IQR]**	**Split Mean ± s.d.**
Flat-trend Prophet, real only	25.3 [15.1, 40.1]	33.8 ± 11.9	22.5 [13.3, 35.6]	26.4 ± 8.2
TOW profile	24.4 [15.2, 37.3]	32.2 ± 13.6	20.7 [12.9, 35.4]	25.5 ± 7.9
Seasonal naive	27.8 [16.3, 42.8]	34.5 ± 11.3	21.3 [13.4, 36.5]	27.3 ± 9.1
Transferred profile	29.6 [23.1, 41.7]	35.9 ± 9.0	27.8 [21.8, 39.9]	33.4 ± 7.6
100% real history (reference, identical in both designs)				
Prophet, real only	24.0 [16.5, 42.9]	33.8 ± 12.5	24.0 [16.5, 42.9]	33.8 ± 12.5
Flat-trend Prophet, real only	22.9 [13.8, 31.2]	26.3 ± 8.5	22.9 [13.8, 31.2]	26.3 ± 8.5
TOW profile	20.5 [12.8, 31.9]	26.2 ± 9.6	20.5 [12.8, 31.9]	26.2 ± 9.6
Seasonal naive	21.3 [13.4, 36.5]	27.3 ± 9.1	21.3 [13.4, 36.5]	27.3 ± 9.1

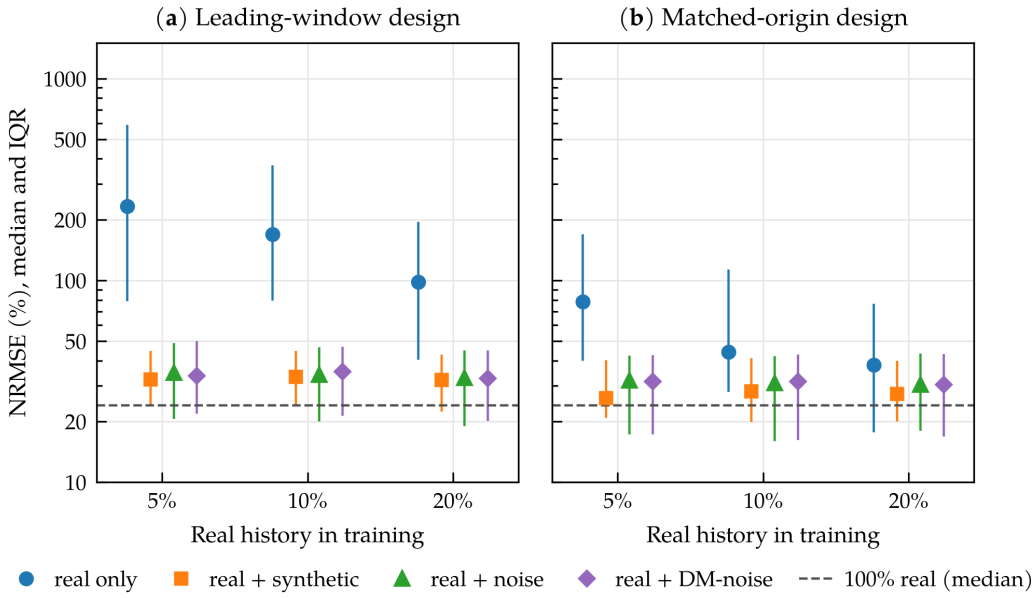


Figure 4. Prophet forecasting error (NRMSE over the full test period, logarithmic scale) by scenario. Markers show the median over the 100 building–split cases and bars the interquartile range. (**a**) Leading-window design: the real-only model is fitted on the first k% of the training window. (**b**) Matched-origin design: the real-only model is fitted on the last k% and all scenarios share the forecast origin.

4.3. Forecasting Under a Matched Forecast Origin

When the scarce real window is moved to the end of the training period, so that all scenarios forecast from the same origin, the picture changes (Figure 4b). The median error of Prophet trained on real data only falls from 233.0 to 78.6% at 5%, from 168.6 to 44.1% at 10% and from 97.9 to 38.1% at 20%. The real window now covers the most recent weeks instead of January 2016, which changes its recency, season and level as well as the extrapolation distance (Section 3.4); Section 4.3 examines these factors. The supplemented scenarios change much less (synthetic: 26.2, 28.1 and 27.3%). The gap that the supplement closes therefore shrinks from a median paired building-level difference of 205.1, 124.2 and 65.4 percentage points in the leading-window design (Table 7) to 51.6, 15.1 and 3.1 pp under the matched origin (Table 8). At 5% the reduction is clear (95% site-clustered interval [−83.5, −23.6] pp; better in 85% of buildings; $p_S < 0.001$). At 10% and 20% it is small and uncertain. At 20% the median difference is −3.1 pp with an interval of [−30.9, 1.9] pp that includes zero, while the Hodges–Lehmann estimate is −14.1 pp and the building-level Wilcoxon test is significant (Holm-adjusted $p = 0.032$). The two do not contradict each other: the paired differences are strongly skewed, because the supplement rescues a minority of buildings in which real-only Prophet fails badly and makes little difference or does slight harm in

the rest (the synthetic supplement is better in only 57% of buildings). The Wilcoxon test targets the pseudomedian, which these large improvements pull away from zero, whereas the median describes the typical building. We therefore read the 20% result as a benefit concentrated in a minority of buildings rather than as a typical improvement.

Figure 5 shows the same shift case by case. In the leading-window design almost all points lie far below the diagonal, with real-only errors spread over two orders of magnitude. Under the matched origin the points at 20% cluster around the diagonal (the supplement helps in 54% of the cases, against 83% in the leading-window design), while at 5% a cloud of real-only failures above 50% remains. Those failures are the subject of Sections 4.4 and 4.5.

Table 7. Paired comparisons of forecasting error over the full test period, leading-window design. Δ: median over distinct buildings of the building-level paired NRMSE difference (first minus second scenario, pp) with 95% site-clustered bootstrap interval; negative values favour the first scenario. HL: Hodges–Lehmann pseudomedian of the same differences, the location parameter of the Wilcoxon test. "Better": share of distinct buildings in which the first scenario has the lower error. "Splits": share of the 20 splits in which the mean paired difference favours the first scenario (descriptive). p_B: building-level Wilcoxon p-value; p_H: Holm-adjusted p_B (family of 54 tests, both designs); p_S: Wilcoxon test on the 11 site means.

Comparison	k	Δ [95% CI]	HL	Better (%)	Splits (%)	p_B	p_H	p_S
synthetic vs. real only	5%	−205.1 [−457.7, −104.1]	−322.3	91	100	<0.001	<0.001	<0.001
	10%	−124.2 [−312.6, −62.4]	−198.0	95	100	<0.001	<0.001	<0.001
	20%	−65.4 [−134.2, −22.3]	−86.1	83	100	<0.001	<0.001	<0.001
noise vs. real only	5%	−204.4 [−451.0, −105.7]	−320.0	90	100	<0.001	<0.001	<0.001
	10%	−127.8 [−310.2, −71.2]	−198.4	95	100	<0.001	<0.001	<0.001
	20%	−70.3 [−135.9, −25.5]	−87.6	84	100	<0.001	<0.001	<0.001
synthetic vs. noise	5%	−1.4 [−3.8, 3.1]	−0.4	59	45	0.748	1.000	0.831
	10%	−2.0 [−3.4, 4.3]	0.3	56	25	0.787	1.000	0.966
	20%	0.1 [−2.4, 4.9]	1.1	50	25	0.251	1.000	0.638
synthetic vs. DM-noise	5%	−2.2 [−4.5, 3.1]	−0.5	61	35	0.639	1.000	0.700
	10%	−2.2 [−3.8, 3.2]	−0.1	57	30	0.913	1.000	0.966
	20%	0.0 [−2.5, 4.2]	0.8	50	35	0.317	1.000	0.577
synthetic vs. permuted	5%	−2.2 [−4.7, 3.1]	−0.3	59	40	0.780	1.000	0.898
	10%	−1.5 [−3.6, 3.4]	0.1	57	30	0.928	1.000	0.966
	20%	0.0 [−2.7, 5.1]	0.9	50	35	0.304	1.000	0.700
TOW profile vs. synthetic	5%	−5.9 [−8.7, −1.5]	−5.7	74	70	<0.001	<0.001	0.067
	10%	−6.8 [−10.6, −3.0]	−6.7	82	75	<0.001	<0.001	0.054
	20%	−7.0 [−10.0, −3.8]	−6.8	83	75	<0.001	<0.001	0.054
seasonal naive vs. synthetic	5%	−4.1 [−8.0, 1.5]	−3.0	56	60	0.039	1.000	0.365
	10%	−4.3 [−9.7, 0.9]	−4.9	66	85	<0.001	0.001	0.010
	20%	−3.7 [−8.6, 2.7]	−3.0	62	70	0.028	0.782	0.278
transferred vs. synthetic	5%	−0.9 [−2.0, 1.1]	−1.0	55	80	0.043	1.000	0.083
	10%	−0.6 [−2.3, 1.1]	−0.9	56	75	0.084	1.000	0.147
	20%	0.2 [−1.2, 1.8]	−0.2	49	55	0.672	1.000	0.520
synthetic vs. 100% real	5%	5.9 [2.8, 10.6]	6.2	26	30	0.001	0.041	0.175
	10%	6.6 [1.9, 11.3]	6.1	24	35	0.001	0.041	0.206
	20%	6.0 [0.4, 10.1]	4.8	27	35	0.003	0.102	0.365

Table 8. Paired comparisons of forecasting error over the full test period, matched-origin design. Δ: median over distinct buildings of the building-level paired NRMSE difference (first minus second scenario, pp) with 95% site-clustered bootstrap interval; negative values favour the first scenario. HL: Hodges–Lehmann pseudomedian of the same differences, the location parameter of the Wilcoxon test. "Better": share of distinct buildings in which the first scenario has the lower error. "Splits": share of the 20 splits in which the mean paired difference favours the first scenario (descriptive). p_B: building-level Wilcoxon p-value; p_H: Holm-adjusted p_B (family of 54 tests, both designs); p_S: Wilcoxon test on the 11 site means.

Comparison	k	Δ [95% CI]	HL	Better (%)	Splits (%)	p_B	p_H	p_S
synthetic vs. real only	5%	−51.6 [−83.5, −23.6]	−64.2	85	100	<0.001	<0.001	<0.001
	10%	−15.1 [−31.4, −0.4]	−27.2	66	95	<0.001	<0.001	<0.001
	20%	−3.1 [−30.9, 1.9]	−14.1	57	80	<0.001	0.032	0.019
noise vs. real only	5%	−49.5 [−83.5, −21.2]	−65.0	84	100	<0.001	<0.001	<0.001
	10%	−10.8 [−32.2, −0.3]	−27.6	67	100	<0.001	<0.001	<0.001
	20%	−4.0 [−30.8, 2.5]	−14.8	62	85	<0.001	0.010	0.054
synthetic vs. noise	5%	−0.7 [−4.4, 3.3]	−0.7	55	45	0.467	1.000	0.465
	10%	−0.4 [−3.0, 2.8]	−0.2	54	45	0.766	1.000	0.577
	20%	−0.5 [−2.5, 2.4]	0.1	52	40	0.939	1.000	0.898
synthetic vs. DM-noise	5%	−0.3 [−4.5, 3.4]	−0.5	54	45	0.562	1.000	0.413
	10%	−0.2 [−3.4, 2.0]	−0.2	55	45	0.801	1.000	0.765
	20%	−0.5 [−2.5, 2.6]	0.0	57	45	0.939	1.000	1.000
synthetic vs. permuted	5%	−1.1 [−4.5, 3.2]	−0.7	55	45	0.433	1.000	0.365
	10%	−0.4 [−3.5, 3.0]	−0.2	54	45	0.826	1.000	0.765
	20%	0.0 [−2.7, 2.9]	0.3	50	45	0.703	1.000	0.577
TOW profile vs. synthetic	5%	−6.8 [−10.3, −3.6]	−7.1	91	100	<0.001	<0.001	<0.001
	10%	−7.3 [−10.5, −5.6]	−8.1	93	100	<0.001	<0.001	<0.001
	20%	−8.8 [−11.8, −4.9]	−9.1	84	95	<0.001	<0.001	<0.001
seasonal naive vs. synthetic	5%	−5.7 [−9.2, −1.3]	−5.8	84	95	<0.001	<0.001	0.042
	10%	−5.8 [−10.0, −3.6]	−6.7	83	95	<0.001	<0.001	0.024
	20%	−6.8 [−9.8, −3.9]	−7.2	84	100	<0.001	<0.001	<0.001
transferred vs. synthetic	5%	0.6 [−0.5, 1.4]	0.4	46	35	0.278	1.000	0.520
	10%	0.9 [−1.0, 2.3]	0.2	41	50	0.656	1.000	0.700
	20%	0.9 [−1.4, 3.1]	0.0	41	60	0.965	1.000	0.365
synthetic vs. 100% real	5%	1.9 [−1.6, 6.0]	1.7	39	45	0.151	1.000	0.831
	10%	3.4 [−2.4, 7.0]	2.4	39	40	0.083	1.000	0.520
	20%	2.6 [−0.5, 5.3]	2.5	35	30	0.012	0.339	0.320

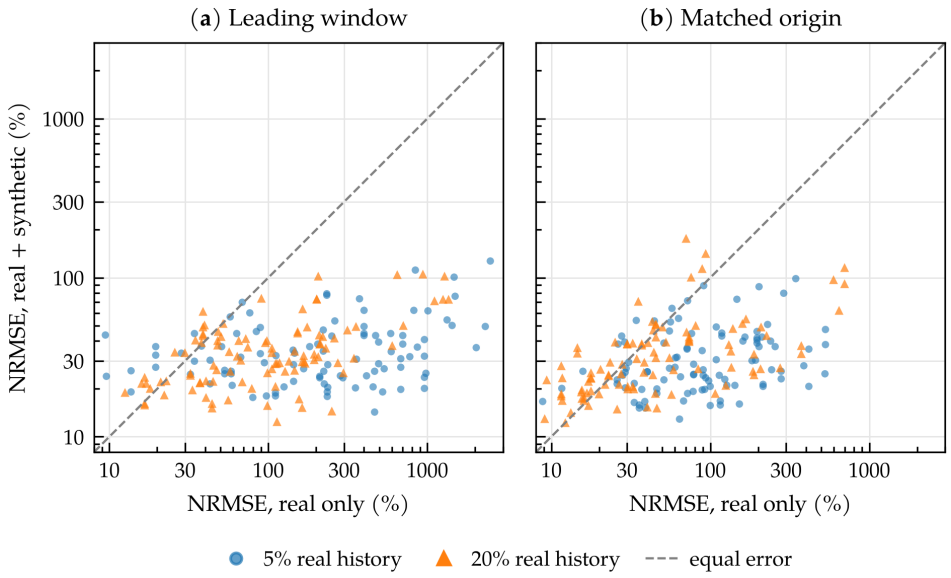


Figure 5. Per-case NRMSE of Prophet trained on real data only (x) and on real plus synthetic data (y), logarithmic axes. Points below the diagonal are cases in which the synthetic supplement reduced the error. (**a**) Leading-window design. (**b**) Matched-origin design.

Season, Recency and Additional Origins

Table 9 separates the factors that change between the two designs for the models fitted on the real window alone. For Prophet, moving the window from January 2016 (LW) to the same calendar weeks as the MO window but one year earlier (SL) gives imprecise changes: the median paired change is −4.8, −42.3 and 13.0 pp at 5, 10 and 20%, with intervals that run from large improvements to large deteriorations (at 10%, [−241.4, 59.0] pp). Moving the same calendar weeks one year closer, from SL to the weeks immediately before the origin (MO), reduces the error by 104.0, 89.9 and 84.5 pp, with intervals that exclude zero at every level. The observed pattern is therefore a large improvement of real-only Prophet when the window is moved one year closer to the test period. This contrast combines recency, a shorter extrapolation distance and any change in a building's consumption level between the two years, and the season contrasts are too imprecise to bound a seasonal contribution, whose size we leave open. The time-of-week profile and the flat-trend model, which do not extrapolate a trend, change far less: by −0.4 to −1.7 pp from LW to SL (no Holm-adjusted $p < 0.05$) and by −1.2 to −2.0 pp from SL to MO for the profile. Recent data help every model somewhat, but only the trend model depends on them heavily.

Table 9. Season and recency contrasts for the models fitted on the real window alone. Median NRMSE (%) over the full test period when the real window is the first k% of the training period (LW), the calendar weeks of the MO window one year earlier (SL; the window ends 364 days before the forecast origin) or the last k% (MO). Paired differences: median over distinct buildings with 95% site-clustered bootstrap interval. The MO − SL contrast also includes the shorter extrapolation distance and any change in consumption level between the two years.

	5% Real			10% Real			20% Real		
Method	**LW**	**SL**	**MO**	**LW**	**SL**	**MO**	**LW**	**SL**	**MO**
Prophet, real only	233.0	181.4	78.6	168.6	148.8	44.1	97.9	123.1	38.1
Flat-trend Prophet, real only	28.8	26.5	21.8	26.6	27.7	21.3	25.3	26.4	22.5
TOW profile	26.9	22.7	20.6	24.4	23.8	19.8	24.4	23.7	20.7
Seasonal naive	29.1	24.7	21.3	28.4	24.7	21.3	27.8	24.7	21.3
Paired differences, Δ [95% CI] (pp)									
Prophet, real only: SL − LW (season)	−4.8 [−164.4, 84.4]			−42.3 [−241.4, 59.0]			13.0 [−36.5, 78.9]		
Prophet, real only: MO − SL (recency)	−104.0 [−210.6, −33.3]			−89.9 [−172.5, −55.5]			−84.5 [−123.5, −45.6]		
TOW profile: SL − LW (season)	−1.7 [−3.5, 0.6]			−0.7 [−4.5, 0.3]			−0.4 [−2.4, 0.1]		
TOW profile: MO − SL (recency)	−2.0 [−4.7, −0.9]			−1.5 [−3.9, −0.6]			−1.2 [−3.3, −0.4]		

Table 10 repeats the matched-origin comparison at four further calendar origins, with no method using any record dated after the origin and the generator recalibrated before each origin (its parameters were similar to those of Table 4: $|\beta| \leq 0.0102$ and σ_{res} between 0.19 and 0.63). Over a 90-day horizon the benefit of the synthetic supplement over real-only Prophet depends on the origin: it is large at 1 December 2016 with the 5% window (−23.6 pp, [−57.8, −10.1]), and its intervals include zero in the other seven origin–level combinations (median differences between −5.1 and +1.6 pp). Over the first 7 days real-only Prophet is more accurate than the supplemented model at every origin and level (median differences between +3.0 and +7.3 pp, all intervals excluding zero). Against white noise the synthetic supplement had a slightly lower error at every origin (median differences between −2.8 and −1.4 pp over 90 days), but no interval excludes zero and no Holm-adjusted p-value is below 0.431; equivalence within ±3 pp is established in only 1 of the eight cases. A small advantage of the generator over white noise at these origins is therefore plausible but has not been demonstrated. Flat-trend Prophet (−7.3 to −4.1 pp) and the time-of-week profile (−8.4 to −6.1 pp) are more accurate than the supplemented model at every origin and level, with intervals excluding zero. The main

qualitative findings of the primary origin therefore hold across seasons, while the size of the supplement's benefit over real-only Prophet depends on the origin and the lead time.

Table 10. Additional forecast origins. At each origin all methods use only data before the origin; the generator is recalibrated before the origin; the real window is the last 612 h (5%) or 2450 h (20%); the test period is 90 days. Medians of NRMSE (%) over the 90 days and paired differences (median over distinct buildings, pp, with 95% site-clustered bootstrap interval) over 90 and 7 days. Flat: flat-trend Prophet, real only. Negative differences favour the first method.

Origin	*k*	Real	+Syn.	+Noise	Flat	TOW	Syn.−Real, 90 d	Syn.−Real, 7 d	Syn.−Noise, 90 d	TOW−syn., 90 d
1 September 2016	5%	34.4	29.3	32.8	22.3	20.1	−0.4 [−9.6, 4.9]	4.0 [0.9, 6.8]	−2.8 [−5.8, 2.3]	−8.0 [−10.7, −4.6]
	20%	30.0	26.1	29.5	23.7	19.9	0.8 [−4.0, 5.4]	3.0 [1.3, 5.5]	−2.0 [−3.5, 0.3]	−6.7 [−14.2, −4.5]
1 December 2016	5%	73.1	25.4	25.6	21.3	18.1	−23.6 [−57.8, −10.1]	6.2 [1.0, 10.0]	−2.0 [−4.9, 3.0]	−6.5 [−9.9, −4.1]
	20%	30.5	28.2	29.0	24.2	21.4	−1.7 [−11.0, 4.1]	4.5 [1.7, 8.1]	−1.4 [−2.7, 1.6]	−8.4 [−14.6, −6.0]
1 March 2017	5%	29.2	23.5	25.6	19.4	16.7	−5.0 [−16.0, 1.9]	5.7 [3.3, 8.9]	−2.4 [−6.1, 2.9]	−6.4 [−11.4, −3.2]
	20%	21.8	23.3	25.4	18.6	15.6	1.6 [−1.2, 5.4]	5.9 [3.7, 9.1]	−1.9 [−3.9, 1.0]	−6.3 [−10.1, −3.5]
1 September 2017	5%	37.2	24.3	28.8	22.2	19.6	−5.1 [−17.8, 2.2]	7.3 [4.1, 10.0]	−2.7 [−5.1, 0.5]	−6.1 [−9.8, −2.4]
	20%	27.8	26.1	29.4	20.3	18.2	0.9 [−5.0, 5.7]	5.8 [3.0, 7.9]	−1.7 [−5.0, 0.6]	−6.5 [−9.4, −4.0]

4.4. Error as a Function of Horizon

Figure 6 and Table 11 split the matched-origin error by horizon. Over the first week after the origin, Prophet trained on the real window alone is more accurate than the supplemented models: its median NRMSE is 17.1, 18.2 and 18.3% at 5, 10 and 20%, against 22.6, 22.0 and 21.3% with the synthetic supplement, and the real-only model is better in 80, 89 and 78% of buildings (median paired differences −5.6 [−8.7, −2.1], −5.8 [−9.3, −3.4] and −5.4 [−9.0, −2.2] pp; Holm-adjusted $p < 0.001$). Over the first 30 days the real-only model is still as good or better (median paired differences of 2.7, 3.8 and 4.9 pp in its favour). Only over the full horizon of about seven months does the real-only model fall behind, and the pattern points to the trend: with a few weeks of data, Prophet's piecewise-linear trend absorbs short-term level changes and extrapolates them over the remaining months.

Two features of the supplement plausibly explain why it costs accuracy at short lead times and helps over the full horizon. First, 80 to 95% of the training hours of a supplemented model come from the complement, so its level and its daily and weekly Fourier terms are estimated mostly from complement values. With the synthetic complement they are pulled towards the generic profile of the calibration buildings and towards the median-based scale, which lies slightly below the real mean (Section 3.3); with white noise, which has no daily or weekly shape, the estimated seasonal amplitude is diluted. Over the first days after the origin the real-only model reproduces the building's own recent level and weekly shape, and these pulls only add bias. The larger first-week error of the white-noise supplement (24.9 and 24.1% at 5 and 20%, against 22.6 and 21.3% with the synthetic one) and the closeness of the supplemented model to the transferred profile (Section 4.7) are consistent with this reading. Second, the cost of the real-only model's trend grows with the lead time, because a slope fitted to a few weeks is extrapolated linearly: it is small over a week and dominates after several months. The supplement thus exchanges a roughly constant bias for the removal of a growing trend error, and which of the two is larger depends on the horizon; the same reversal occurred at every additional origin

(Table 10). In practice, for short-term operational uses such as day-ahead or week-ahead scheduling and demand response, a short recent history is better used directly, through a seasonal naive forecast, a time-of-week profile or a real-only model, than completed with a supplement. For long open-loop projections such as M&V baselines, the problem that the supplement addresses is trend extrapolation, and a flat trend or a time-of-week profile handled it better than the supplement at every horizon studied (Table 11).

Table 11. Forecasting error by horizon in the matched-origin design: median NRMSE (%) over the first 7 days, the first 30 days and the full test period, and paired differences (median over distinct buildings, pp) with 95% site-clustered bootstrap intervals and Holm-adjusted p-values (horizon family for 7 and 30 days; main family for the full period). Negative differences favour the first method.

	First 7 Days		First 30 Days		Full Period	
Method	**5%**	**20%**	**5%**	**20%**	**5%**	**20%**
Prophet, real only	17.1	18.3	21.8	20.4	78.6	38.1
Prophet + synthetic	22.6	21.3	24.1	23.2	26.2	27.3
Prophet + noise	24.9	24.1	28.2	27.5	31.9	30.3
Flat-trend Prophet, real only	18.0	19.3	19.3	20.6	21.8	22.5
TOW profile	17.0	18.3	14.8	17.7	20.6	20.7
Seasonal naive	14.6	14.6	14.5	14.5	21.3	21.3
Paired differences, Δ [95% CI] (pp); p_H in brackets below						
Real only − synthetic	−5.6 [−8.7, −2.1] (<0.001)	−5.4 [−9.0, −2.2] (<0.001)	−2.7 [−5.3, 1.3] (0.625)	−4.9 [−7.2, −0.8] (0.009)	51.6 [23.6, 83.5] (<0.001)	3.1 [−1.9, 30.9] (0.032)
Flat-trend, real only − synthetic	−5.9 [−9.9, −2.5] (<0.001)	−4.1 [−7.7, −2.6] (<0.001)	−6.4 [−10.2, −3.7] (<0.001)	−4.3 [−7.0, −2.2] (<0.001)	−5.0 [−8.5, −3.0] (<0.001)	−7.0 [−10.8, −5.1] (<0.001)
TOW profile − synthetic	−6.9 [−10.8, −4.2] (<0.001)	−4.7 [−7.9, −1.3] (<0.001)	−6.1 [−11.2, −3.2] (<0.001)	−4.9 [−8.6, −1.9] (<0.001)	−6.8 [−10.3, −3.6] (<0.001)	−8.8 [−11.8, −4.9] (<0.001)

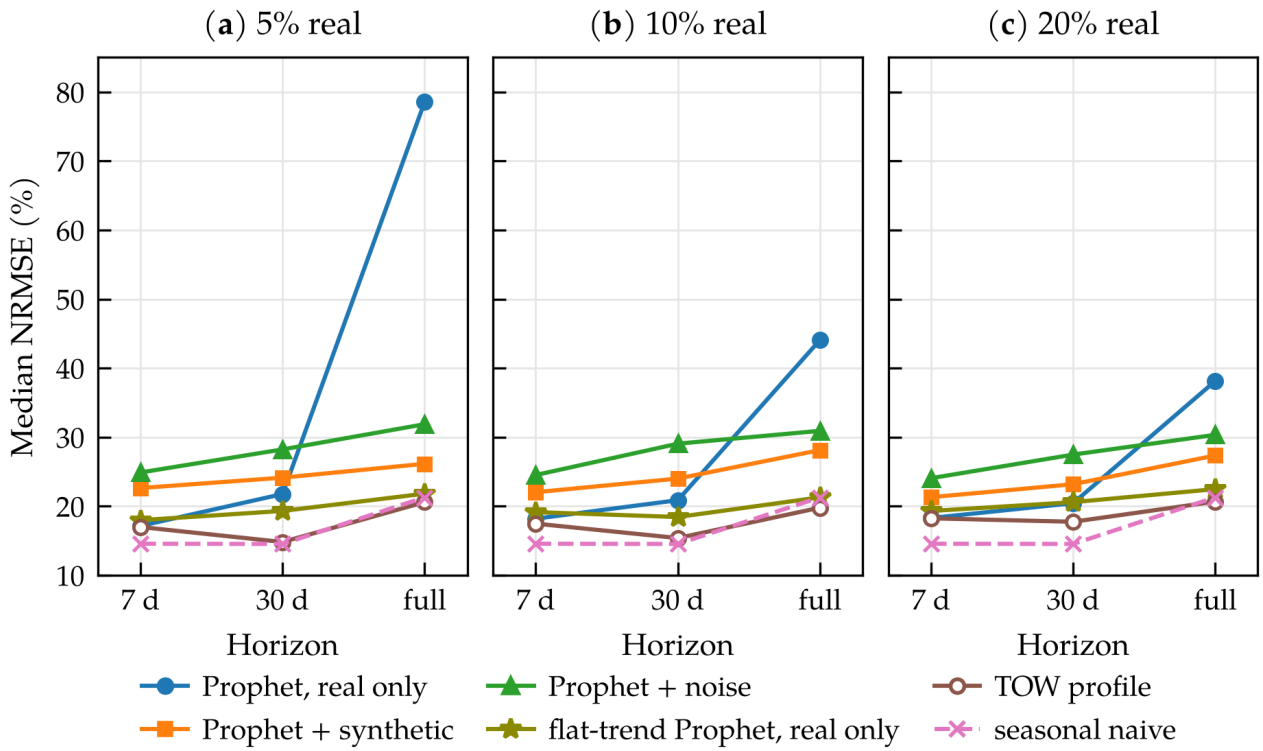


Figure 6. Median NRMSE as a function of the evaluation horizon (first 7 days, first 30 days and the full test period of about 219 days) in the matched-origin design. All curves share the same forecast origin.

4.5. What the Supplement Changes in Prophet

The flat-trend variant is consistent with this reading. With the trend fixed, Prophet trained on the real window alone reaches a median NRMSE of 21.8, 21.3 and 22.5% over the full horizon, below every supplemented model and below default Prophet trained on the full history (24.0%); it is better than Prophet with the synthetic supplement in 89, 93 and 89% of distinct buildings (median differences −5.0 [−8.5, −3.0], −6.5 [−8.8, −4.0] and −7.0 [−10.8, −5.1] pp; Table 12). Once the trend is flat, adding the synthetic complement makes the model worse: the median paired difference is +5.3, +5.7 and +5.0 pp at 5, 10 and 20%.

Prophet's default changepoint placement offers an alternative explanation. In every supplemented matched-origin model, the 25 candidate changepoints lay between 17 January and 16 February 2017; none fell inside the real segment at any level (at 20% the last candidate lay about two hours before it, at the end of the first 80% of the history). The trend of a supplemented model could therefore change only in the complement, whereas the real-only model places all its candidates in the real data. When the candidate changepoints of the supplemented models are placed in the real segment, at the dates Prophet uses for the real-only model, the supplemented models become less accurate than with the default placement (median paired differences +0.2, +1.0 and +1.1 pp; Hodges–Lehmann estimates +1.7 to +4.8 pp). The synthetic supplement then still reduces the error of real-only Prophet clearly at 5% (−28.7 [−60.8, −6.3] pp), but no longer detectably at 10% (−2.7 [−21.8, 5.3]) or 20% (−0.4 [−17.5, 4.1]), and no difference from white noise with the same placement was detected (+0.4, −0.1 and +1.1 pp, no Holm-adjusted $p < 0.05$). Changepoint placement therefore accounts for much of the remaining gain at 10 and 20%, but not for the gain at 5%: with only 25 days of real data, a long, level complement still fixes the initial slope of the trend, from which changepoints in the real segment can depart only under their Laplace prior.

Choosing a smaller changepoint prior scale for the real-only model on a validation week, without test data, changes its error little: the median paired difference from the default is zero at every level, and the marginal medians move from 78.6, 44.1 and 38.1% to 64.0, 38.0 and 37.7%. One validation week is a poor guide to a seven-month extrapolation, and the default 0.05 was selected in 26 to 44% of cases; even the smallest prior scale, 0.001, applied to every building, leaves median errors of 65.2, 31.3 and 27.5%. The validated model remains less accurate than the supplemented model at 5% (+19.4 [5.0, 68.2] pp) and than the flat-trend model at every level (+11.4 to +28.1 pp). A smaller prior scale is therefore a weak substitute for the supplement, whereas the strongest form of trend regularisation, a flat trend, is better than both.

These checks are consistent with the supplement acting on Prophet mainly through the trend, but they do not isolate every effect. Three others remain possible: the complement anchors the level of the series at the median-based scale (Section 3.3); it pulls the daily and weekly Fourier terms towards the generic profile of the calibration buildings; and it changes the data from which the temperature coefficient is estimated, because 80 to 95% of the training hours then pair the observed temperature with complement values. The last effect appears to be small here, since removing the temperature regressor changes the error of the supplemented model by a median of at most 0.04 pp (Section 4.7). The cost of the synthetic complement once the trend is flat (above) indicates that, with the trend fixed, these remaining effects are harmful on balance. The fitted trend components of the real-only and supplemented models were not compared directly, so the attribution to the trend rests on these indirect checks.

Table 12. Mechanism checks in the matched-origin design. Upper part: median NRMSE (%) with interquartile range over the full test period. Lower part: paired differences (median over distinct buildings, pp) with 95% site-clustered bootstrap interval and Holm-adjusted p-value (mechanism family). "Changepoints in real segment": candidate changepoints placed where Prophet places them for the real-only model. "Validated prior scale": changepoint prior scale chosen from $\{0.001, 0.005, 0.01, 0.05\}$ on the last week of the real window. "No temperature": Prophet without the temperature regressor.

Model or Comparison	5% Real	10% Real	20% Real
Prophet, real only (default)	78.6 [40.0, 169.2]	44.1 [27.9, 112.9]	38.1 [17.7, 76.6]
Prophet + synthetic (default)	26.2 [20.8, 40.1]	28.1 [19.9, 41.2]	27.3 [20.0, 39.9]
Prophet + synthetic, changepoints in real segment	30.0 [20.8, 53.2]	34.2 [20.7, 56.2]	30.7 [20.2, 49.5]
Prophet + noise, changepoints in real segment	32.8 [17.4, 57.8]	38.5 [16.3, 61.3]	34.1 [18.3, 53.5]
Prophet, real only, validated prior scale	64.0 [28.9, 131.8]	38.0 [24.2, 82.6]	37.7 [17.4, 64.9]
Prophet, real only, prior scale 0.001	65.2 [28.8, 152.6]	31.3 [18.9, 58.5]	27.5 [15.1, 41.3]
Flat-trend Prophet, real only	21.8 [13.0, 36.5]	21.3 [13.0, 35.4]	22.5 [13.3, 35.6]
Flat-trend Prophet + synthetic	26.4 [20.4, 41.7]	26.6 [20.3, 41.9]	26.3 [19.0, 39.3]
Prophet, real only, no temperature	90.4 [52.3, 193.9]	65.3 [33.5, 165.7]	37.6 [19.4, 93.4]
Prophet + synthetic, no temperature	26.5 [20.6, 39.5]	29.1 [20.0, 40.5]	28.5 [19.1, 41.4]
Flat-trend Prophet, real only, no temperature	22.3 [13.2, 36.8]	22.9 [13.2, 37.2]	23.3 [14.1, 36.7]
TOW profile (no temperature)	20.6 [12.4, 35.9]	19.8 [12.2, 35.6]	20.7 [12.9, 35.4]
Prophet, 100% real, no temperature		27.9 [17.8, 42.9]	
Paired differences, Δ [95% CI] (pp) and p_H			
synthetic: real-segment − default changepoints	0.2 [−0.2, 1.6]; 0.249	1.0 [0.1, 5.0]; 0.001	1.1 [0.1, 5.3]; 0.020
synthetic (real-segment changepoints) − real only	−28.7 [−60.8, −6.3]; <0.001	−2.7 [−21.8, 5.3]; 0.613	−0.4 [−17.5, 4.1]; 0.963
synthetic − noise, both with real-segment changepoints	0.4 [−4.0, 7.2]; 1.000	−0.1 [−4.3, 7.2]; 1.000	1.1 [−2.6, 5.6]; 1.000
real only: validated − default prior scale	0.0 [−8.7, 0.0]; 0.446	0.0 [−2.9, 0.0]; 0.657	0.0 [0.0, 0.0]; 0.963
real only, validated prior − synthetic	19.4 [5.0, 68.2]; <0.001	4.3 [−3.3, 20.2]; 0.026	2.1 [−4.1, 11.7]; 0.329
flat-trend real only − synthetic	−5.0 [−8.5, −3.0]; <0.001	−6.5 [−8.8, −4.0]; <0.001	−7.0 [−10.8, −5.1]; <0.001
flat-trend: synthetic − real only	5.3 [4.1, 9.4]; <0.001	5.7 [4.2, 9.0]; <0.001	5.0 [3.1, 7.4]; <0.001
synthetic − HoW-permuted	0.4 [0.1, 1.0]; 0.046	0.3 [−0.1, 0.9]; 0.502	0.1 [−0.2, 1.3]; 0.963
no temperature: synthetic − real only	−60.7 [−103.1, −32.9]; <0.001	−26.9 [−103.8, −4.9]; <0.001	−3.7 [−26.1, 0.1]; 0.012
TOW profile − synthetic (no temperature)	−6.1 [−10.2, −3.9]; <0.001	−7.4 [−10.2, −5.2]; <0.001	−8.5 [−12.4, −5.0]; <0.001
flat-trend real only − synthetic (no temperature)	−5.0 [−8.1, −3.3]; <0.001	−5.9 [−8.2, −3.8]; <0.001	−6.7 [−10.6, −3.2]; <0.001

4.6. Synthetic Data Versus Structure-Free Controls

If the generator's hour-of-week structure carried useful information, the synthetic supplement should beat supplements of the same length and dispersion that carry none. No such advantage was detected. Figure 7 shows the median paired differences between the synthetic supplement and each control with their 90% site-clustered bootstrap intervals. All 18 intervals of the comparisons with white noise, DM-noise and the permuted complement include zero, none of the building-level tests is significant after Holm correction (all Holm-adjusted $p = 1.000$), and none of the site-level tests is either (smallest $p_S = 0.365$). The absence of a demonstrated advantage is not, however, evidence of equal performance. Equivalence within ± 3 pp is established for 4 of the 18 comparisons with white noise, DM-noise and the permuted complement (white noise at 10%; white noise at 20%; DM-noise at 20%; permuted at 20%, all in the matched-origin design); with a ± 5 pp margin, for 18 of 18; with ± 2 pp, for none. The other intervals cross -3 pp, the side that would favour the synthetic supplement, except in the leading-window design at 20%, where they cross $+3$ pp. A small advantage of the generator therefore cannot be excluded in those settings, but none has been demonstrated. For these 14 comparisons the results show neither a difference nor equal performance. The dispersion-matched control, which has the same mean and spread as the synthetic complement but no temporal order, gives errors close to those of the white-noise control (the median paired differences of the two controls from the synthetic supplement differ by at most 0.8 pp at every level and in both designs), and so does the permuted complement, which keeps the synthetic marginal distribution exactly. Removing only the autocorrelation and the load–temperature alignment while keeping the calendar profile (the hour-of-week-permuted control) gave errors equivalent to those of the synthetic supplement within ± 3 pp at every level; at 5% the synthetic supplement was even slightly worse ($+0.44$ pp [0.06, 1.01], Holm-adjusted $p = 0.046$), in line with the sensitivity analysis in which stronger AR(1) persistence increased the error (Section 4.10).

These results hold under both designs, so the lack of a detected advantage over white noise under the leading-window design is not an artefact of the shifted forecast origin.

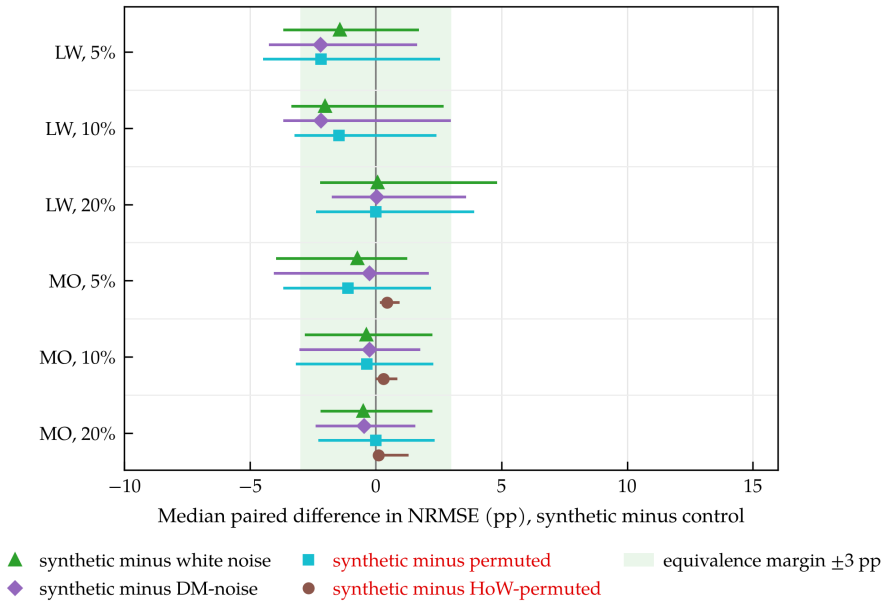


Figure 7. Difference between the synthetic supplement and the structure-free controls. Points are medians of the paired differences across distinct buildings and bars are 90% site-clustered bootstrap intervals, the intervals used for the equivalence test: a bar entirely inside the shaded band establishes equivalence within ±3 pp. Negative values favour the synthetic supplement. The HoW-permuted control was run in the matched-origin design only.

4.7. Simple Baselines and Weather Information

Figure 8 ranks all methods by their median error in the matched-origin design, and the lower part of Table 8 gives the paired tests. The time-of-week profile computed from the scarce real window alone is the most accurate method at every scarcity level, with median NRMSE of 20.6, 19.8 and 20.7%. It improves on Prophet with the synthetic supplement by 6.8, 7.3 and 8.8 pp (better in 91, 93 and 84% of buildings, Holm-adjusted $p < 0.001$, site-level $p_S < 0.001$), and it is also more accurate than default Prophet trained on the full 17-month history (24.0%). Split by split, the mean error of the profile is below that of the supplemented models in 39 of the 40 split–level combinations listed in Appendix B. The seasonal naive forecast, which only repeats the last observed week, follows closely (21.3%; it does not depend on k because every window contains the last week before the origin), and so does the flat-trend Prophet trained on the real window (21.8, 21.3 and 22.5%), which the TOW profile beats by 0.8 to 0.9 pp (median paired differences). The transferred profile, which uses nothing from the target building but its scale, gives errors close to those of Prophet with the synthetic supplement (median paired difference 0.6, 0.9 and 0.9 pp, all intervals including zero and none extending beyond 3.1 pp): the augmented Prophet model forecasts roughly what the calibration buildings would predict for a generic education building.

The same ranking holds under the leading-window design (Table 6), where the TOW profile computed from a January window still beats the supplemented Prophet models by 5.9 to 7.0 pp.

Weather information. All Prophet results above use the observed test-period temperature. Without the temperature regressor (Table 12), the median error of Prophet with the synthetic supplement is 26.5, 29.1 and 28.5%, that of real-only Prophet 90.4, 65.3 and 37.6%, and that of flat-trend Prophet 22.3, 22.9 and 23.3%; full-history Prophet without temperature reaches 27.9%. Removing temperature hardly changes the supplemented

model (median paired difference −0.02 to +0.04 pp), worsens real-only Prophet (+3.5, +4.8 and +1.1 pp), and leaves the flat-trend model more accurate than the supplemented one (−5.0 [−8.1, −3.3] to −6.7 [−10.6, −3.2] pp). The time-of-week profile and the seasonal naive forecast never used temperature, so their errors are unchanged and remain the lowest. The conclusions about the supplement do not depend on knowing the test-period weather.

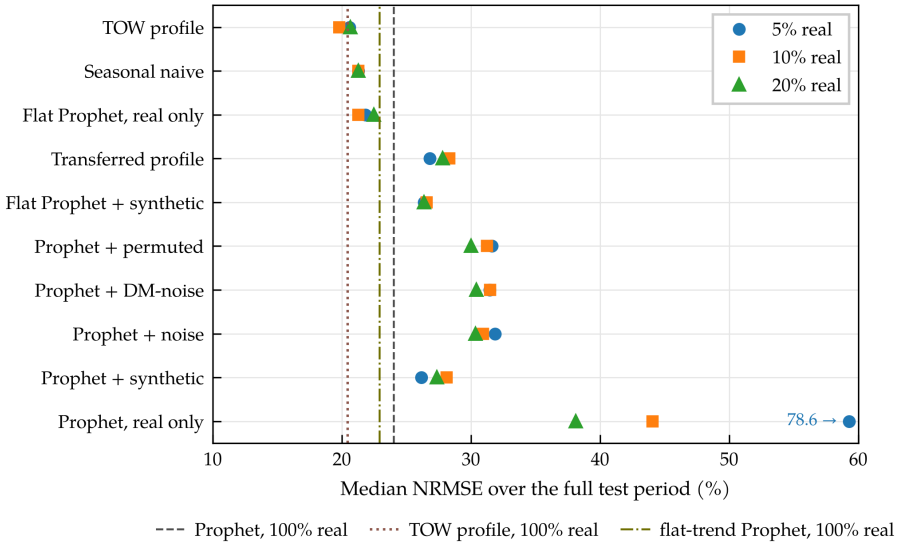


Figure 8. Median forecasting error of all methods in the matched-origin design. Points beyond the axis limit are drawn at its right edge and labelled with their value. Vertical lines mark full-history references. TOW: time of week; DM: dispersion matched.

4.8. Anomaly Detection with Isolation Forest

Table 13 and Figure 9 report the Isolation Forest results. The detector is weak even with the full history (precision 0.113, recall 0.148, F1 0.091, AP 0.114), so all differences are read against a low baseline. Training on a short real window flags many more test hours (on average 535 flags for about 52 injected anomalies at 5% under the matched origin, against 110 with the full history), which raises recall at the cost of precision; F1 under scarcity is close to the full-history value in the matched-origin design (0.101, 0.102 and 0.099).

Every supplement lowers F1, in both designs and at every scarcity level. With the synthetic complement, recall collapses to 0.092, 0.084 and 0.067 and F1 to 0.023, 0.026 and 0.027 (matched origin); in 49 to 52% of cases the detector finds no injected anomaly at all, against 3 to 5% with the real window alone. The controls show that this harm is not specific to the generator: white noise (0.037, 0.036 and 0.037), DM-noise (0.023, 0.023 and 0.023) and the permuted complement (0.025, 0.024 and 0.021) lower it as well. The paired reductions in F1 relative to the real window are between 0.061 and 0.074, with cluster-bootstrap intervals that exclude zero and Holm-adjusted $p < 0.001$ for every complement and level (Table 14); the site-level tests agree ($p_S \leq 0.002$). In F1 the synthetic complement is not better than any control: against white noise the mean difference is −0.011 [−0.023, 0.000] at 5%, and against DM-noise and the permuted complement it lies within the ±0.02 equivalence margin at every level. The contamination sensitivity check leaves the conclusion unchanged: at contamination 0.008 and 0.012 the F1 of every supplemented scenario remains below that of the corresponding real-only scenario (Table 13, footnote).

The threshold-free average precision changes this picture in an instructive way. The synthetic complement does not lower AP relative to the real window (mean differences +0.013, +0.013 and +0.018, all intervals including zero), and it yields a higher AP than white noise, DM-noise and the permuted complement (differences of +0.02 to +0.03, intervals excluding zero, Holm-adjusted $p \leq 0.014$). With a synthetic complement the

forest therefore still ranks injected hours roughly as well as with the real window alone, and slightly better than with unstructured complements; what fails is the decision threshold. The contamination rule sets the threshold at the 1% most anomalous training points, and when 80 to 95% of the training points come from a complement that is more dispersed than the building's own data, that quantile moves far out and few test hours exceed it (269 flags at 5%, against 535 with the real window). Under the timestamp-based lags used here (Section 3.8) the Isolation Forest values differ only slightly from those of the row-shifted implementation, and none of the conclusions changes.

Table 13. Isolation Forest results: mean over splits of the per-building means ± standard deviation over splits. The first four metric columns refer to the matched-origin design; the last column gives F1 for the leading-window design. F1 is not the harmonic mean of the averaged precision and recall because averaging is done per building and then across splits. AP: average precision (threshold-free). Lags are timestamp-based (Section 3.8).

Scenario	Precision (MO)	Recall (MO)	F1 (MO)	AP (MO)	F1 (LW)
100% real	0.113 ± 0.067	0.148 ± 0.071	0.091 ± 0.043	0.114	0.091 ± 0.043
5% real only	0.063 ± 0.026	0.398 ± 0.102	0.101 ± 0.038	0.074	0.090 ± 0.035
5% real + synthetic	0.043 ± 0.056	0.092 ± 0.083	0.023 ± 0.015	0.088	0.028 ± 0.020
5% real + noise	0.083 ± 0.119	0.168 ± 0.070	0.037 ± 0.020	0.065	0.039 ± 0.023
5% real + DM-noise	0.042 ± 0.051	0.101 ± 0.086	0.023 ± 0.015	0.058	0.025 ± 0.018
5% real + permuted	0.045 ± 0.058	0.106 ± 0.086	0.025 ± 0.014	0.058	0.023 ± 0.015
5% real + HoW-permuted	0.047 ± 0.045	0.143 ± 0.115	0.036 ± 0.022	0.072	0.030 ± 0.021
10% real only	0.080 ± 0.053	0.330 ± 0.086	0.102 ± 0.040	0.079	0.093 ± 0.044
10% real + synthetic	0.047 ± 0.063	0.084 ± 0.079	0.026 ± 0.021	0.094	0.027 ± 0.019
10% real + noise	0.067 ± 0.075	0.141 ± 0.059	0.036 ± 0.017	0.071	0.037 ± 0.017
10% real + DM-noise	0.052 ± 0.071	0.088 ± 0.074	0.023 ± 0.016	0.061	0.026 ± 0.022
10% real + permuted	0.036 ± 0.048	0.087 ± 0.075	0.024 ± 0.019	0.060	0.026 ± 0.018
10% real + HoW-permuted	0.050 ± 0.053	0.117 ± 0.097	0.035 ± 0.026	0.075	0.033 ± 0.026
20% real only	0.083 ± 0.052	0.286 ± 0.079	0.099 ± 0.036	0.082	0.091 ± 0.047
20% real + synthetic	0.049 ± 0.058	0.067 ± 0.059	0.027 ± 0.023	0.101	0.026 ± 0.023
20% real + noise	0.077 ± 0.086	0.113 ± 0.059	0.037 ± 0.019	0.079	0.038 ± 0.021
20% real + DM-noise	0.035 ± 0.039	0.065 ± 0.056	0.023 ± 0.019	0.074	0.029 ± 0.029
20% real + permuted	0.036 ± 0.048	0.063 ± 0.053	0.021 ± 0.014	0.071	0.031 ± 0.025
20% real + HoW-permuted	0.039 ± 0.039	0.092 ± 0.074	0.032 ± 0.021	0.084	0.037 ± 0.029

Contamination sensitivity (MO, mean F1 at 0.008 / 0.012, for 5% / 10% / 20%): real only 0.100/0.098, 0.098/0.105, 0.096/0.101; synthetic 0.021/0.025, 0.020/0.032, 0.023/0.030; noise 0.031/0.040, 0.032/0.040, 0.033/0.041; DM-noise 0.021/0.024, 0.020/0.025, 0.021/0.027; permuted 0.023/0.025, 0.022/0.027, 0.019/0.024.

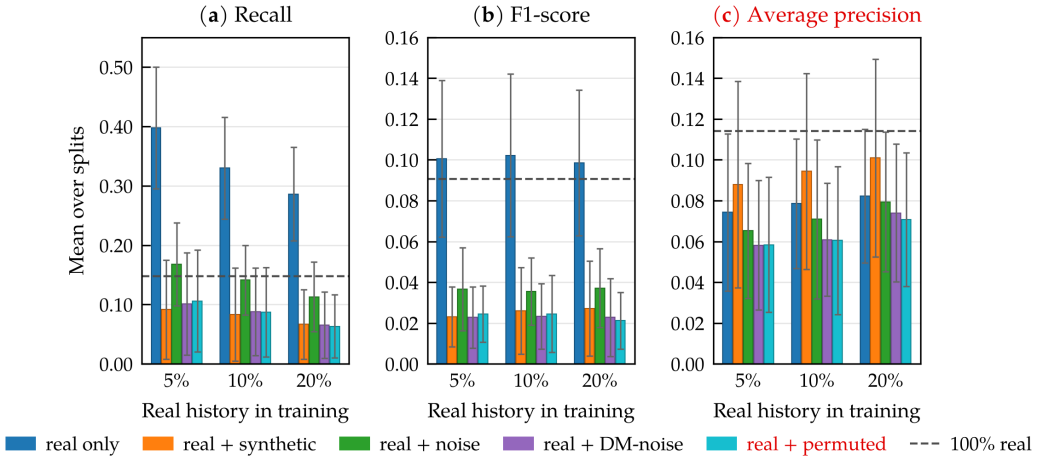


Figure 9. Isolation Forest on injected point anomalies, matched-origin design. (**a**) Recall. (**b**) F1-score. (**c**) Average precision. Bars are means over the 20 splits of the per-building values; error bars show ±1 standard deviation over splits.

Table 14. Paired differences in detection performance, matched-origin design: mean over distinct buildings of the building-level difference in F1 with 95% site-clustered bootstrap interval, and the mean difference in average precision (AP). Positive values favour the first scenario. †: equivalent within ±0.02 (Isolation Forest) or ±0.05 (profile detector) by the 90% interval.

	5% Real		10% Real		20% Real	
Comparison	**ΔF1 [95% CI]**	**ΔAP**	**ΔF1 [95% CI]**	**ΔAP**	**ΔF1 [95% CI]**	**ΔAP**
Isolation Forest						
synthetic − real only	−0.074 [−0.096, −0.054]	0.013	−0.069 [−0.095, −0.043]	0.013	−0.067 [−0.091, −0.043]	0.018
noise − real only	−0.063 [−0.082, −0.044]	−0.008	−0.063 [−0.086, −0.040]	−0.008	−0.061 [−0.082, −0.038]	−0.004
synthetic − noise	−0.011 [−0.023, 0.000]	0.021	−0.006 [−0.020, 0.009] †	0.021	−0.006 [−0.020, 0.008] †	0.022
synthetic − DM-noise	0.000 [−0.007, 0.006] †	0.027	0.005 [−0.005, 0.016] †	0.030	0.005 [−0.004, 0.015] †	0.025
synthetic − permuted	−0.002 [−0.009, 0.003] †	0.025	0.002 [−0.004, 0.011] †	0.030	0.006 [−0.003, 0.016] †	0.028
Profile-residual						
synthetic − real only	0.140 [0.061, 0.213]	0.343	0.057 [−0.018, 0.129]	0.228	0.030 [−0.056, 0.113]	0.122
noise − real only	0.107 [0.024, 0.195]	0.309	0.030 [−0.040, 0.106]	0.202	0.032 [−0.039, 0.101]	0.120
synthetic − noise	0.033 [−0.045, 0.116]	0.035	0.028 [−0.047, 0.101]	0.026	−0.002 [−0.065, 0.051]	0.002
synthetic − DM-noise	0.101 [0.059, 0.149]	0.052	0.088 [0.047, 0.134]	0.035	0.054 [0.012, 0.100]	−0.001
synthetic − permuted	0.134 [0.079, 0.194]	0.078	0.122 [0.068, 0.181]	0.048	0.096 [0.046, 0.149]	−0.017
Profile-residual detector, real-only variants − synthetic complement						
pooled scale	−0.066 [−0.146, 0.009]	0.098	−0.009 [−0.093, 0.074]	0.064	−0.001 [−0.094, 0.088]	0.022
shrunk scale	−0.081 [−0.160, −0.006]	0.060	−0.025 [−0.113, 0.065]	0.023	−0.010 [−0.096, 0.082]	−0.018
cross-fitted threshold	−0.147 [−0.225, −0.066]	−0.343	−0.128 [−0.206, −0.039]	−0.228	0.026 [−0.073, 0.122]	−0.122
shrunk + cross-fitted	0.016 [−0.060, 0.097]	0.060	0.097 [0.003, 0.189]	0.023	0.056 [−0.035, 0.148]	−0.018

4.9. Anomaly Detection with a Profile-Residual Detector

The weak Isolation Forest baseline raises the question of whether its behaviour generalises. Table 15 and Figure 10 report the profile-residual detector on the same injected anomalies. With the full history it reaches an F1 of 0.430 (precision 0.368, recall 0.675, AP 0.557), about 4.7 times the F1 of Isolation Forest. With a scarce real window the same rule is far less precise (precision 0.092 at 5% under the matched origin): with about 612 hours of data each hour-of-week cell holds only three or four observations, 11.6 of the 168 cells have a MAD of zero on average and fall back to the global scale, the cell medians and spreads are poorly estimated, and the 1% threshold calibrated on the same training residuals flags many normal test hours. A long complement corrects much of this. With the synthetic complement F1 rises to 0.271, 0.277 and 0.289 under the matched origin, and with white noise to 0.252, 0.264 and 0.304, mainly through higher precision. The gain of the synthetic complement over the real window alone is largest at 5% (mean F1 difference +0.140 [0.061, 0.213] in the matched-origin design and +0.183 [0.116, 0.255] in the leading-window design) and is not detectable at 20% under the matched origin (+0.030 [−0.056, 0.113]).

The new controls qualify the earlier reading that any long complement helps equally. Against white noise, no statistically supported difference was detected at any level (mean F1 differences +0.033, +0.028 and −0.002), but the intervals are too wide to establish equivalence within ±0.05, so the two complements cannot be said to perform equally either. Against the two controls that share the synthetic complement's dispersion, the synthetic complement is clearly better: DM-noise reaches F1 0.181 and the permuted complement 0.146 at 5% (differences +0.101 [0.059, 0.149] and +0.134 [0.079, 0.194], Holm-adjusted $p < 0.001$); the same holds at 10%, and at 20% for the permuted complement, whereas the DM-noise difference at 20% (+0.054) does not survive the Holm correction. The reason lies in how the detector uses the data: it estimates a median and a spread per hour of the week, so a complement whose values follow the calendar profile adds observations with a plausible level and spread to each cell, whereas a complement with the same overall dispersion but no calendar order places the full between-hour spread inside every cell, inflates the cell scales and lowers recall (0.255 with the permuted complement against 0.451 with the synthetic one). For this detector, then, the calendar structure of the complement matters relative to an equally dispersed unstructured complement, but it does

not beat a white-noise complement whose low, constant spread happens to suit a per-cell scale estimate.

Average precision shows that most of the supplements' effect is on the scale and the threshold rather than on the information available about the building. All complements raise AP strongly over the real window alone (from 0.131 to between 0.420 and 0.479 at 5%), but a real-only detector with a pooled scale reaches a higher AP than any supplemented detector (0.585 at 5%; difference from the synthetic complement +0.098 [0.023, 0.172]). Its F1 at the in-sample threshold is lower (0.217), because the pooled scale over-flags the high-variance daytime hours; combining a shrunk scale with a cross-fitted threshold gives F1 values of 0.305, 0.386 and 0.352 from the real window alone, at least as high as those of the synthetic complement (differences +0.016 [−0.060, 0.097], +0.097 [0.003, 0.189] and +0.056 [−0.035, 0.148]). A cross-fitted threshold on its own does not help at 5 and 10%, because with three to seven observations per cell the out-of-fold scales are even less stable; it helps only at 20%. The benefit of a long complement for this detector is therefore largely a remedy for unstable small-sample scale and threshold estimates, and a real-only detector with a regularised scale and an out-of-sample threshold obtains it without any added data.

Table 15. Profile-residual detector: mean precision, recall, F1 (±s.d. over splits) and average precision (AP) on the same injected anomalies as Table 13. With the full real history the detector reaches precision 0.368, recall 0.675, F1 0.430 ± 0.096 and AP 0.557. Real-only variants use no complement (Section 3.8).

	Leading-Window Design				Matched-Origin Design			
Scenario	**Prec.**	**Rec.**	**F1**	**AP**	**Prec.**	**Rec.**	**F1**	**AP**
5% real only	0.059	0.484	0.098 ± 0.030	0.078	0.092	0.502	0.138 ± 0.043	0.131
5% real + synthetic	0.391	0.436	0.272 ± 0.126	0.461	0.398	0.451	0.271 ± 0.105	0.479
5% real + noise	0.342	0.481	0.256 ± 0.110	0.420	0.321	0.500	0.252 ± 0.115	0.460
5% real + DM-noise	0.295	0.299	0.186 ± 0.116	0.410	0.292	0.333	0.181 ± 0.084	0.439
5% real + permuted	0.255	0.239	0.156 ± 0.126	0.404	0.260	0.255	0.146 ± 0.089	0.420
5% real only, pooled scale	0.192	0.787	0.271 ± 0.095	0.464	0.146	0.859	0.217 ± 0.059	0.585
5% real only, shrunk scale	0.141	0.807	0.211 ± 0.082	0.411	0.134	0.846	0.204 ± 0.063	0.542
5% real only, cross-fitted threshold	0.066	0.226	0.080 ± 0.028	0.078	0.128	0.258	0.125 ± 0.050	0.131
5% real only, shrunk + cross-fitted	0.296	0.593	0.268 ± 0.087	0.411	0.281	0.715	0.305 ± 0.098	0.542
10% real only	0.120	0.656	0.176 ± 0.063	0.195	0.160	0.665	0.232 ± 0.051	0.273
10% real + synthetic	0.424	0.420	0.282 ± 0.139	0.473	0.397	0.439	0.277 ± 0.114	0.505
10% real + noise	0.387	0.444	0.277 ± 0.120	0.464	0.348	0.469	0.264 ± 0.112	0.494
10% real + DM-noise	0.338	0.280	0.207 ± 0.127	0.443	0.313	0.319	0.198 ± 0.098	0.481
10% real + permuted	0.312	0.214	0.174 ± 0.130	0.460	0.255	0.251	0.164 ± 0.106	0.468
10% real only, pooled scale	0.199	0.776	0.277 ± 0.112	0.472	0.203	0.819	0.280 ± 0.076	0.576
10% real only, shrunk scale	0.167	0.786	0.243 ± 0.095	0.434	0.183	0.815	0.265 ± 0.063	0.530
10% real only, cross-fitted threshold	0.262	0.130	0.097 ± 0.053	0.195	0.328	0.159	0.149 ± 0.080	0.273
10% real only, shrunk + cross-fitted	0.348	0.639	0.319 ± 0.117	0.434	0.367	0.695	0.386 ± 0.096	0.530
20% real only	0.174	0.707	0.230 ± 0.071	0.321	0.211	0.719	0.276 ± 0.061	0.413
20% real + synthetic	0.405	0.397	0.280 ± 0.134	0.496	0.396	0.402	0.289 ± 0.109	0.535
20% real + noise	0.419	0.414	0.296 ± 0.119	0.502	0.397	0.435	0.304 ± 0.116	0.544
20% real + DM-noise	0.351	0.262	0.222 ± 0.121	0.495	0.375	0.278	0.239 ± 0.119	0.543
20% real + permuted	0.337	0.204	0.183 ± 0.139	0.521	0.365	0.223	0.197 ± 0.128	0.562
20% real only, pooled scale	0.201	0.790	0.282 ± 0.111	0.477	0.220	0.795	0.300 ± 0.088	0.566
20% real only, shrunk scale	0.203	0.773	0.267 ± 0.097	0.444	0.220	0.789	0.294 ± 0.073	0.520
20% real only, cross-fitted threshold	0.339	0.442	0.277 ± 0.075	0.321	0.366	0.491	0.321 ± 0.083	0.413
20% real only, shrunk + cross-fitted	0.308	0.695	0.322 ± 0.086	0.444	0.311	0.727	0.352 ± 0.102	0.520

Magnitude and prevalence. Table 16 varies the anomaly magnitude and prevalence in the matched-origin design. The ordering of the scenarios is stable across settings. For Isolation Forest every complement lowers F1 at every magnitude and prevalence. For the profile detector, the synthetic and white-noise complements improve on the in-sample real-only rule at 5% for all settings, DM-noise and the permuted complement do less well, and the real-only detector with a shrunk scale and a cross-fitted threshold has an F1 higher than, or within 0.01 of, that of the synthetic complement for anomalies of 2 and 3 standard

deviations, but not for the largest (5 standard deviations), where the supplemented detector is better at 5 and 20%. Larger anomalies are easier for all detectors (F1 of the profile detector with the full history 0.306, 0.430 and 0.507 at 2, 3 and 5 standard deviations), and the realised magnitude is smaller than nominal for negative spikes that are truncated at zero (Section 3.8). These results concern injected point anomalies only; Section 5.6 discusses what they imply for persistent faults and gradual degradation.



Figure 10. F1-score of the profile-residual detector on the same injected anomalies. (**a**) Leading-window design. (**b**) Matched-origin design. Bars are means over the 20 splits; error bars show ±1 standard deviation over splits.

Table 16. Sensitivity of detection F1 to anomaly magnitude (in training standard deviations) and prevalence, matched-origin design (mean over all building–split cases).

	Prevalence 1%			**Magnitude 3 s.d.**	
Detector, Scenario	**2 s.d.**	**3 s.d.**	**5 s.d.**	**0.5%**	**2%**
Profile-residual					
5% real only	0.106	0.138	0.179	0.082	0.216
5% + synthetic	0.163	0.271	0.416	0.223	0.326
5% + noise	0.143	0.252	0.415	0.199	0.305
5% + DM-noise	0.097	0.181	0.308	0.147	0.218
5% + permuted	0.079	0.146	0.282	0.118	0.181
5% real only, shrunk + cross-fitted	0.205	0.305	0.376	0.215	0.394
20% real only	0.209	0.276	0.334	0.182	0.383
20% + synthetic	0.177	0.289	0.436	0.243	0.334
20% + noise	0.159	0.304	0.469	0.254	0.353
20% + DM-noise	0.120	0.239	0.383	0.204	0.268
20% + permuted	0.100	0.197	0.348	0.176	0.213
20% real only, shrunk + cross-fitted	0.248	0.352	0.431	0.254	0.451
100% real	0.306	0.430	0.507	0.321	0.530
Isolation Forest					
5% real only	0.070	0.101	0.106	0.062	0.155
5% + synthetic	0.014	0.023	0.043	0.012	0.040
5% + noise	0.023	0.037	0.062	0.027	0.059
5% + DM-noise	0.016	0.023	0.040	0.012	0.037
5% + permuted	0.016	0.025	0.043	0.012	0.037
20% real only	0.059	0.099	0.115	0.065	0.144
20% + synthetic	0.016	0.027	0.050	0.015	0.038
20% + noise	0.022	0.037	0.064	0.026	0.060
20% + DM-noise	0.015	0.023	0.041	0.016	0.035
20% + permuted	0.015	0.021	0.045	0.015	0.034
100% real	0.048	0.091	0.128	0.070	0.122

4.10. Sensitivity to the Generator and Prophet Settings

Table 17 reports the sensitivity analyses on the five original splits. Changing the AR(1) coefficient from 0.85 to 0.70 or 0.95 changes the median forecasting error of the synthetic scenarios by at most 5.5 pp. The direction is informative: stronger persistence ($\varphi = 0.95$) tends to increase the error, while weaker persistence ($\varphi = 0.70$), which brings the synthetic residual closer to white noise, reduces it slightly in the leading-window design and leaves it practically unchanged under the matched origin. No statistically supported difference between the synthetic supplement and white noise was detected for any of the 18 design, level and coefficient combinations: the smallest uncorrected building-level p-value is 0.035, which does not survive Holm correction within this family. More temporal structure in the residual does not make the generator more useful. For Isolation Forest, the F1 of the synthetic scenarios stays below that of the real-only scenarios for every coefficient. Raising Prophet's changepoint prior scale to 0.5, which makes the trend more flexible, increases the error of the real-only scenarios further (matched origin, 5%: median 193.2% vs. 90.5% with the default), while the supplemented scenarios change little. This shows that more flexibility makes the trend failure worse; whether appropriate regularisation removes the need for a supplement is a different question, addressed by the validated prior scale and the changepoint ablation of Section 4.5.

Table 17. Sensitivity analyses on the five original splits (25 cases): median NRMSE (%) of Prophet over the full test period and mean Isolation Forest F1 (timestamp-based lags). φ: AR(1) coefficient of the generator; cps: changepoint prior scale of Prophet.

	Leading-Window Design			Matched-Origin Design		
Setting and Scenario	**5%**	**10%**	**20%**	**5%**	**10%**	**20%**
Prophet + synthetic, $\varphi = 0.70$	28.0	29.7	29.1	27.7	27.0	26.2
Prophet + synthetic, $\varphi = 0.85$	30.0	31.3	30.3	27.4	26.9	26.5
Prophet + synthetic, $\varphi = 0.95$	32.3	33.4	35.8	27.6	31.8	27.1
Prophet + noise (any φ)	35.9	34.6	33.8	32.1	32.4	30.5
IF F1, real + synthetic, $\varphi = 0.70$	0.046	0.046	0.045	0.039	0.037	0.047
IF F1, real + synthetic, $\varphi = 0.85$	0.044	0.038	0.039	0.030	0.030	0.037
IF F1, real + synthetic, $\varphi = 0.95$	0.029	0.038	0.026	0.023	0.022	0.033
IF F1, real only	0.102	0.114	0.107	0.141	0.140	0.129
Prophet cps 0.05 (default), real only	146.3	145.3	104.9	90.5	43.5	55.0
Prophet cps 0.05 (default), real + synthetic	30.0	31.3	30.3	27.4	26.9	26.5
Prophet cps 0.05 (default), real + noise	35.9	34.6	33.8	32.1	32.4	30.5
Prophet cps 0.5, real only	625.1	248.2	100.8	193.2	76.5	70.1
Prophet cps 0.5, real + synthetic	30.3	30.0	30.3	29.5	28.5	26.8
Prophet cps 0.5, real + noise	35.8	34.5	33.5	33.3	33.4	30.8

5. Discussion

5.1. Where the Forecasting Gain Came from

Under the leading-window design, synthetic data appeared to cut Prophet's error by a large factor, and white noise gave a similar reduction. The matched-origin design and the additional diagnostics show that this gain had several sources, none of which is the information content of the generator.

The first is the information available at the forecast origin. Under the leading-window design the real-only model was fitted on the first weeks of January 2016, a period that combines the winter break with the start of a semester, and was asked to forecast late May to December 2017, which includes the summer break and a full autumn semester. Its last observation was a median of 486 days (5%) to 409 days (20%) before the first test hour. Moving the same amount of real data next to the origin cut the real-only median error by a factor of 3.0 at 5% and 2.6 at 20% without any supplement. That move changes the season, recency, level and scale of the window together with the extrapolation distance,

so it is not a pure origin effect. The seasonally matched lagged window shows what happens when each of two changes is made on its own (Table 9). Moving the same calendar weeks one year closer to the test period reduced the error of real-only Prophet by 84.5 to 104.0 pp, with intervals that exclude zero, whereas changing only the season of the window gave imprecise estimates of either sign, with intervals spanning large gains and large losses. Models without a trend, such as the time-of-week profile, gained only one or two points from either change. Because the lagged window lies exactly one year before the matched-origin window, the first contrast cannot separate recency from extrapolation distance, nor either of them from a change in a building's level between the two years, and the imprecision of the second leaves the size of any seasonal contribution open. What the data do show is that much of the gap between the designs closes when real-only Prophet is fitted on data close to the test period, which is also what a supplement that extends the training window to the origin provides for Prophet's trend.

The second mechanism is the trend. With a matched origin, real-only Prophet is accurate over the first week and degrades over the following months because the piecewise-linear trend fitted to a few weeks of data is extrapolated over a seven-month horizon. Any long and level supplement, synthetic or random, constrains that degree of freedom: Prophet sees more than a year of data with no systematic drift and estimates an almost flat trend. Three checks support this reading. A flat-trend Prophet on the real window alone is more accurate than every supplemented model; allowing the trend of the supplemented models to change inside the real segment makes them less accurate and removes the detectable advantage of the supplement at 10 and 20%; and a smaller changepoint prior scale chosen without test data, a weaker form of regularisation, recovers little of the real-only model's error. They do not prove that the trend is the only channel: the complement also anchors the level at the median-based scale, pulls the seasonal terms towards the calibration buildings' generic profile and changes the data from which the temperature coefficient is estimated, and the cost of the synthetic complement once the trend is flat ($+5.0$ to $+5.7$ pp) suggests that, once the trend is fixed, these remaining effects are harmful on balance. The most economical interpretation is that the supplement acts mainly as a regulariser of Prophet's trend, and that the same or better regularisation is available without it.

The third mechanism, the generator's temporal structure, made no measurable contribution to forecasting. No statistically supported advantage was detected for synthetic data over white noise, dispersion-matched noise or a full permutation of the synthetic complement, in any design or at any level, and the same held against white noise at four further origins, although there the estimates consistently leaned slightly towards the synthetic supplement (1–3 pp) without reaching significance; equivalence within ± 3 pp could be established for some of these comparisons but not all (Section 4.6). Removing the AR(1) autocorrelation and the load–temperature alignment from the synthetic complement, while keeping its calendar profile, made it, if anything, slightly better. Changing the AR(1) persistence did not alter that. This is consistent with Table 4: once the thermal term is small, the generator reduces to a profile shared by most education buildings plus noise, and a model that already has daily and weekly Fourier terms gains little from seeing that profile again.

5.2. Why Simple Baselines Win

The time-of-week profile computed from 25 days of recent data was more accurate than default Prophet trained on 17 months. This is less surprising than it looks. Hourly consumption in education buildings is dominated by the weekly occupancy cycle; the TOW profile estimates that cycle directly with one parameter per hour of the week and no trend, which is exactly the structure that TOWT models exploit in M&V practice [4,21].

Prophet with default settings spends its flexibility on a trend that is a liability at long horizons, and its error with the full history (median 24.0%) is already above that of the TOW profile with the same data (20.5%). The finding echoes the forecasting literature, where simple and combined statistical methods often match or beat more complex ones [19], and the machine learning literature, where untuned or missing baselines have repeatedly overstated progress [37,38]. For the present question it has a practical consequence: an augmentation pipeline whose result with the synthetic supplement is 26.2% NRMSE at 5% scarcity has not solved a problem that a profile average solves at 20.6%, and the ranking held at every additional origin. As an indicative reference only, since our out-of-sample NRMSE is not the in-sample CV(RMSE) of ASHRAE Guideline 14 (Section 3.9), the TOW profile computed from 5% of the history stays below an NRMSE of 30% in 71% of cases, more often than default Prophet with the full history (63%).

The transferred profile gives a second reading of the supplemented Prophet models. No statistically supported difference was detected between its error and that of Prophet with the synthetic supplement, although it uses no consumption data from the target building except its scale. Augmenting a scarce window with synthetic data from other buildings and fitting a flexible model therefore produced roughly the forecast that the other buildings would have given directly.

5.3. Anomaly Detection Depends on the Detector, Its Threshold and Its Scale Estimates

The cross-task contrast seen under the leading-window design, synthetic data helping forecasting and harming detection, needs three qualifications. On the forecasting side, the help is mainly a trend effect that disappears with a flat-trend model. On the detection side, the harm is specific to Isolation Forest and is not caused by the generator's temporal structure: dispersion-matched noise and the permuted complement did as much damage as synthetic data, and white noise, which is less dispersed, did less. The third qualification concerns where the harm arises. The threshold-free average precision of Isolation Forest is not lowered by the synthetic complement, and is slightly higher with it than with any unstructured control; it is the contamination-based threshold, set on a training sample that consists mostly of complement values more dispersed than the building's own data, that moves too far out and lets recall collapse. A supplement that is not specific to the building and is at least as spread as the real data enlarges the region that the threshold treats as normal and pushes the injected spikes inside it.

The profile-residual detector behaves differently because it models normality per hour of the week. When the real window covers only a few weeks, each hour-of-week cell contains three or four observations, some cells have a MAD of zero, the robust spread is poorly estimated and the detector over-flags. A 511-day window holds about 73 observations per cell; at the scarcity levels studied the complement adds between about 58 and 69 observations to each cell (before missing hours are removed). Even though they come from other buildings or from random draws, they bring the spread estimate to a realistic level, which raises precision more than it lowers recall. Here the calendar structure of the complement does matter, but only relative to equally dispersed unstructured complements: DM-noise and the permuted complement place the full between-hour spread in every cell and help much less, while white noise, whose low and constant spread suits a per-cell scale, helps about as much as the synthetic complement, with no statistically supported difference and no demonstrated equivalence. More importantly, in the main anomaly setting of the matched-origin design a real-only detector with a shrunk scale and a cross-fitted threshold reaches a mean F1 at least as high as that of every supplemented detector (the supplemented detectors are better for the largest anomalies and, marginally, at the lowest prevalence with a 5% window), and a real-only detector with a pooled scale has the highest

average precision. The benefit of a long complement for this detector is therefore a remedy for unstable small-sample scale and threshold estimates, and regularising those estimates achieves it without added data. With 20% of the history under the matched origin the real cells are better populated and the synthetic complement no longer has a detectable effect.

For both detectors, then, what matters is how the added data change the detector's scale and threshold, not the information that the generator carries about the building. Two practical lessons follow. The sign of an augmentation effect on detection is a property of the detector, its thresholding rule and the data, so a result obtained with one detector should not be generalised to anomaly detection as a task, and threshold-free metrics should be reported alongside thresholded ones. And the choice of detector mattered far more than the supplement: replacing Isolation Forest by the profile rule raised F1 from 0.091 to 0.430 with the full history, an improvement that no supplement approached. These conclusions concern injected point anomalies; persistent faults, drifts and collective anomalies were not studied (Section 5.6).

5.4. A Protocol for Evaluating Synthetic Data in Building Analytics

Table 18 condenses the lessons into a checklist. None of the items is new on its own; the contribution of this study is to show, on a single public dataset, how much each one changes the conclusion. Leaving out the matched origin overstated the forecasting gain severalfold (from a median paired reduction of 205.1 to 51.6 pp at 5% and from 65.4 to 3.1 pp at 20%), a change that reflects where the real window lies relative to the test period, not only the origin; leaving out the controls would have credited the generator with an effect it has not been shown to have; leaving out simple baselines would have hidden that the whole Prophet pipeline is dominated by a profile average; leaving out mechanism checks would have hidden that a flat trend or a regularised detector achieves what the supplement achieves; leaving out threshold-free metrics would have attributed to the supplement a loss of ranking ability that is in fact a threshold effect; and leaving out a second detector would have turned a model-specific result into a general statement about anomaly detection.

Table 18. Checklist for evaluating synthetic data supplements in building energy analytics, with the effect each item had in this study.

Item	Why	Effect in this Study
Matched forecast origin and several origins	Compared models must forecast the same period from the same origin, ideally with no record dated after it; one origin is one season	Median paired gain of the supplement fell from 205.1 to 51.6 pp at 5% and from 65.4 to 3.1 pp at 20%; its size varied across four further origins
Contrasts of season and recency	Moving a window changes more than the origin	Real-only Prophet improved markedly when the same weeks were moved one year closer to the test period; season contrasts too imprecise to size; trend-free models changed by 1–2 pp
Horizon-resolved errors	Trend and level effects appear at different horizons	Supplement worse than real only over the first week at every origin
Structure-free controls (white noise, dispersion-matched noise, permuted complements)	Separates window length, marginal distribution and dispersion from temporal structure	No statistically supported advantage of structure in forecasting; equivalence shown in only 4 of 18 comparisons. For the profile detector, structure helped relative to equally dispersed controls but not relative to white noise
Simple baselines (seasonal naive, time-of-week profile, transferred profile)	Shows whether the pipeline is needed at all	TOW profile best at every level; ahead of all scarce-window Prophet variants at every origin; at the primary origin also better than full-history Prophet

Table 18. *Cont.*

Item	Why	Effect in this Study
Mechanism checks (flat trend, changepoint placement, validated regularisation)	Identifies what the supplement actually changes	Results consistent with a trend regulariser; harmful once the trend is fixed; changepoint placement explains much of the gain at 10 and 20%
Threshold-free metrics and regularised real-only detectors	Separates ranking quality from threshold and scale estimation	Isolation Forest's loss was a threshold effect; a real-only profile detector with a shrunk scale and cross-fitted threshold was not detectably worse than the supplemented ones
Separate evaluation per task and per model	Effects need not transfer	Supplements harmed Isolation Forest F1 but helped the in-sample profile detector at 5%
Cluster-aware statistics, explicit estimands, equivalence tests	Buildings recur across splits and share sites; non-significance is not equivalence	The median and the Wilcoxon location disagreed at 20%, revealing gains concentrated in a minority of buildings
Fidelity metrics as description only	Similarity is neither necessary nor sufficient for utility	Synthetic series fail KS tests yet show no supported advantage over noise in utility

5.5. Relation to Prior Work

Our results do not show that data generation cannot help under scarcity, and they do not contradict the conclusion of Li et al. [7], which concerns different models, generators and protocols. They show that, with a leading-window design, a large apparent gain can arise without any contribution from the generator, and they suggest that comparisons of scarcity strategies should report a structure-free control and simple baselines alongside the generator. The same applies to more expressive generators [13,14]: a conditional diffusion model may well produce series that are more useful than our three-component generator, but that has to be shown against controls that restore the same window length, marginal distribution and dispersion. Our findings also support the emphasis that recent surveys place on utility over fidelity [11,16], with the addition that utility itself depends on how the evaluation is set up.

5.6. Limitations

The study has several limitations. All buildings belong to the education typology and to BDG2's North American and European sites; other typologies with weaker weekly cycles, such as offices with variable occupancy or lodging, may favour different methods. Only one family of generator was examined, and its thermal term was small in almost every split, so the results say little about temperature-driven synthetic consumption. The limited benefit of synthetic data found here should therefore not be generalised to other building types or generation methods. The quantitative results, namely the size of the gains, the ranking of the methods and the advantage of the time-of-week profile, rest on the strong weekly cycle of education buildings and on a generator that in practice reduces to a shared profile with noise. In building types whose consumption depends more on weather or on irregular occupancy, a time-of-week profile is a weaker baseline, and a generator with a substantial thermal term, or a conditional generative model that uses building metadata, could carry information that a short window lacks. What transfers is the evaluation logic: the confounding of a supplement with the forecast origin, the need for structure-free controls and simple baselines, and the dependence of detection results on a detector's threshold and scale estimates are properties of how the comparison is designed, not of the building type or the generator, and the checklist of Table 18 applies to other typologies and generators unchanged. The forecasting models were used with default settings apart from the changepoint checks; tuned or deep models might use synthetic data differently. The primary analysis uses one origin per building and an open-loop forecast; the four additional origins cover other seasons with 90-day horizons but are not a full

rolling-origin evaluation, and forecasts updated as new readings arrive were not studied. Temperature is taken at site level, and the main Prophet results assume that the test-period temperature is known; the temperature-free analysis shows that the conclusions about the supplement do not depend on this, but forecast weather was not evaluated. The anomalies are artificial point spikes of fixed nominal magnitude, some of them truncated at zero; collective anomalies, drifts, persistent faults and real faults may respond differently to supplements, and the profile-residual detector is naturally suited to point spikes defined relative to the series. This limits what the detection results say about the faults that matter most in operation. A persistent fault, such as equipment left running outside its schedule, or a gradual loss of performance shifts the level or shape of consumption over days or months rather than at single hours. A detector that scores each hour against a profile or a forest learned from a short window can absorb a fault already present in that window into what it treats as normal, and the lag features of Isolation Forest carry less information once a shift persists. Such faults are better sought with window-level or cumulative statistics, for example the cumulative residual against a baseline projection, and evaluated with event-level rather than point-wise metrics. Two of our findings concern how a detector is calibrated rather than the shape of the anomaly and are therefore likely to carry over: a dispersed complement moves a contamination-based threshold outwards, and unstable small-sample scale estimates can be regularised without added data. Whether a supplement helps or harms the detection of persistent faults and drifts, and how detectors rank on them, cannot be inferred from point-anomaly results and requires labelled real faults or injected faults of that kind. The cleaned BDG2 data had outliers and zero readings removed upstream over the whole series. Finally, the 82 evaluation buildings come from only 11 sites; the site-clustered intervals account for this dependence only approximately, condition on the calibration fits of the twenty splits and do not establish robustness to the removal of any single site, and the results remain conditional on the BDG2 education subset rather than on a defined population of buildings.

6. Conclusions

We re-evaluated a transparent statistical–temporal generator for supplementing short metering histories of education buildings, with a matched forecast origin, a seasonally matched lagged window, four additional origins, structure-free controls, simple baselines, mechanism checks and two anomaly detectors, over twenty building splits of BDG2.

RQ1. Under a matched forecast origin, the median paired reduction in Prophet's error obtained by adding synthetic data fell from 65.4–205.1 pp to 3.1–51.6 pp, had a site-clustered interval including zero at 20% real history, and was negative over the first week. Real-only Prophet improved markedly when the same calendar weeks were moved one year closer to the test period, a contrast that combines recency, extrapolation distance and changes in level between years; the season contrasts were too imprecise to size any seasonal contribution. The size of the remaining benefit varied across origins.

RQ2. No statistically supported contribution of the generator's temporal structure to forecasting was detected. The synthetic supplement did not differ detectably from white noise, dispersion-matched noise or a fully permuted complement in either design or at any level, nor from white noise at the additional origins. This is not evidence of equivalence: within ± 3 pp, equivalence was established for only 4 of the 18 comparisons at the primary origin and 1 of the 8 at the additional origins, and in the others a difference of that size could not be excluded. Against the calendar-preserving, hour-of-week-permuted complement the synthetic supplement was equivalent within ± 3 pp at every level, with a slightly higher error at 5% ($+0.4$ pp, Holm-adjusted $p = 0.046$). The remaining gain over real-only Prophet is consistent with a regularisation of Prophet's trend, which a flat trend achieves better.

RQ3. Under the matched origin, a time-of-week profile and a seasonal naive forecast computed from the scarce real window were more accurate than every supplemented model, and the profile was more accurate than default Prophet trained on the full history; the profile's advantage over the supplemented models held at every additional origin and without temperature information.

RQ4. All supplements lowered the F1-score of Isolation Forest, but not its average precision: the harm lies in the contamination-based threshold, which a dispersed complement moves too far out. For a profile-residual detector that was several times more accurate, every supplement raised F1 when the real window was shortest; the synthetic complement did better than equally dispersed unstructured complements but not better than white noise, and a real-only detector with a shrunk scale and a cross-fitted threshold reached a mean F1 at least as high without any added data. These results concern injected point anomalies and do not transfer directly to persistent faults or gradual degradation.

Synthetic data are therefore not a remedy that can be judged once and reused across tasks or models. Their usefulness must be shown on the downstream task, with a matched forecast origin at several origins, against structure-free controls, simple baselines and regularised versions of the models they are meant to help, and separately for each model. The findings are specific to education buildings in BDG2 and to a simple statistical–temporal generator. They do not show that synthetic data are of limited value for other building types or for more expressive generators, which remains to be tested, but the evaluation protocol applies to them unchanged. Future work should apply this protocol to more expressive generators conditioned on building metadata, to other building typologies, to rolling-origin and updated forecasts, to forecast weather and to real and persistent fault data, and should study partial supplements that add only as much synthetic data as a model needs.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/forecast8060113/s1, File S1: Archive containing the analysis code, an environment lock file, the SHA-256 checksums of the input files, the building and site memberships of every split, the timestamps that delimit every training window, all model settings and the per-building paired results of every experiment.

Author Contributions: Conceptualization, P.M. and M.A.; methodology, P.M. and T.E.; software, T.E. and P.M.; validation, P.M.; formal analysis, P.M. and T.E.; investigation, T.E.; data curation, T.E.; writing—original draft preparation, P.M. and T.E.; writing—review and editing, P.V., J.S., F.M. and M.A.; visualization, P.M.; supervision, P.M. and M.A. 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: The Building Data Genome Project 2 data analysed in this study are publicly available at https://github.com/buds-lab/building-data-genome-project-2 (accessed on 18 November 2026); the two files used are named in Section 3.1. The analysis code, an environment lock file, the checksums of the input files, the building and site memberships of every split, the timestamps that delimit every training window, all model settings and the per-building results of every experiment are provided in the Supplementary Materials. Further information is available on reasonable request to the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

ACF	Autocorrelation function
AP	Average precision
AR(1)	First-order autoregressive process
BDG2	Building Data Genome Project 2
CI	Confidence interval
CV(RMSE)	Coefficient of variation of the root mean squared error
DM-noise	Dispersion-matched noise
HL	Hodges–Lehmann estimate
HoW	Hour of week
IF	Isolation Forest
IQR	Interquartile range
KS	Kolmogorov–Smirnov
LW	Leading-window design
M&V	Measurement and verification
MO	Matched-origin design
NRMSE	Normalised root mean squared error
pp	Percentage points
RMSE	Root mean squared error
SL	Seasonally matched lagged-window diagnostic
TOW(T)	Time-of-week (and temperature) model

Appendix A. Reproduction of the Earlier Evaluation and Optimiser Diagnostics

The full pipeline was re-implemented and executed again from the public BDG2 files with Prophet 1.5.0. For the five original splits, building selection, the generator parameters (Table 4), all generator validation metrics and all Isolation Forest results (computed with the row-shifted lags used at the time) were identical to the outputs of the earlier evaluation. Prophet results for the supplemented scenarios and the full-history reference agreed closely (median absolute difference 0.04 pp; maximum 1.11 pp; Table A1), as expected from small numerical differences between optimiser builds. The scarce real-only scenarios of the leading-window design differed more (median absolute difference 2.7 pp; maximum 109 pp). These fits extrapolate a trend over more than a year, and small differences in the fitted trend are amplified over that distance: the earlier evaluation's own archived per-split files for two splits, produced by a separate execution, differ from its consolidated results for the same scenarios by up to 25 pp for a single building. Table A2 reports optimiser diagnostics for the five original splits. Prophet's default L-BFGS optimiser never terminated abnormally, so no fit fell back to Newton's method and the discrepancies are not caused by a fallback. Refitting every model with Newton's method changed the supplemented forecasts very little (median absolute difference 0.18 pp in the leading-window design and 0.15 pp in the matched-origin design), but changed the real-only forecasts considerably (10.4 and 2.8 pp; at most 434 pp). With a few weeks of data the trend of a real-only Prophet model is weakly identified, and different optimisers stop at different points of an almost flat objective. The real-only cases that differ by more than 5 pp from the earlier evaluation are also the most sensitive to the optimiser (median difference between optimisers 20.9 pp, against 4.5 pp for the other cases). The direction of the scenario comparisons does not depend on the optimiser: with Newton's method the synthetic complement lowers NRMSE in at least as many cases as with L-BFGS at every level and design, and the median paired change at matched-origin 5% is −58.0 pp, against −59.9 pp with L-BFGS. The conclusions of the earlier evaluation under its own design are unchanged.

Table A1. Mean NRMSE (%) over the 25 cases of the five original splits, leading-window design: earlier evaluation (consolidated results) and re-execution.

Scenario	Earlier	Re-Execution	Difference
5% real only	284.40	289.62	+5.22
5% real + synthetic	34.97	34.98	+0.01
5% real + noise	35.91	35.91	+0.01
10% real only	344.45	347.74	+3.29
10% real + synthetic	35.32	35.23	-0.09
10% real + noise	35.38	35.37	-0.02
20% real only	154.49	152.39	-2.11
20% real + synthetic	34.70	34.69	-0.01
20% real + noise	34.25	34.27	+0.02
100% real	31.23	31.24	+0.01

Table A2. Optimiser diagnostics for the five original splits (both designs, all scarcity levels): share of fits in which Prophet's default L-BFGS optimiser terminated abnormally and fell back to Newton's method, and absolute difference in NRMSE between the default optimiser and Newton's method used directly.

Design, Training Data	Fits	Fallback (%)	Median $\|\Delta\|$ (pp)	Max. $\|\Delta\|$ (pp)
Leading window, real only	75	0	10.37	434.1
Leading window, real + synthetic	75	0	0.18	1.4
Matched origin, real only	75	0	2.80	130.4
Matched origin, real + synthetic	75	0	0.15	1.2

Appendix B. Per-Split Results

Table A3 lists the mean NRMSE of each split for the matched-origin design at 5% and 20% real history. The ranking of the TOW profile below the supplemented Prophet models holds in 39 of the 40 split–level combinations.

Table A3. Mean NRMSE (%) per split in the matched-origin design.

	5% Real History				20% Real History			
Seed	Real	+Syn.	+Noise	TOW	Real	+Syn.	+Noise	TOW
42	126.4	29.6	26.7	19.1	47.4	26.6	23.9	17.3
100	244.1	40.4	39.3	35.6	108.0	40.7	39.7	35.1
200	133.7	30.1	32.2	22.0	75.7	31.0	32.7	23.7
400	124.5	38.7	43.5	30.9	197.5	43.3	47.7	30.4
500	109.6	25.1	31.6	18.0	69.4	24.8	27.9	18.0
600	146.9	34.5	33.5	28.7	63.8	33.0	32.1	33.8
800	96.3	31.8	35.5	21.5	48.6	31.8	33.4	20.5
900	142.8	31.7	30.3	24.5	112.7	29.7	29.2	26.7
1000	70.1	31.3	30.5	16.5	18.4	32.6	29.3	17.2
1200	137.3	26.6	25.1	19.5	66.3	28.5	26.9	17.7
1300	132.0	25.1	23.2	20.5	53.5	24.2	23.4	21.6
1400	77.4	31.7	34.4	25.6	84.5	55.9	56.2	23.6
1600	205.6	51.4	48.8	44.6	176.4	50.6	49.4	46.3
1700	90.2	39.1	44.8	34.2	158.9	45.5	49.8	30.9
1800	103.0	25.8	29.1	18.5	24.1	24.2	27.2	18.5
1900	135.6	39.3	42.3	32.7	228.2	48.3	48.5	34.7
2000	84.9	36.2	37.4	27.4	32.4	38.7	38.4	27.6
2100	52.9	25.0	17.6	14.8	35.3	24.1	19.5	17.0
2200	177.0	24.7	23.9	21.9	99.5	57.6	54.5	27.1
2300	43.9	31.8	27.5	21.4	31.0	32.8	27.3	21.4

References

1. United Nations Environment Programme; Global Alliance for Buildings and Construction. *Global Status Report for Buildings and Construction 2024/2025: Not Just Another Brick in the Wall*; UNEP: Nairobi, Kenya, 2025.
2. Amasyali, K.; El-Gohary, N.M. A review of data-driven building energy consumption prediction studies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095.
3. Deb, C.; Zhang, F.; Yang, J.; Lee, S.E.; Shah, K.W. A review on time series forecasting techniques for building energy consumption. *Renew. Sustain. Energy Rev.* **2017**, *74*, 902–924. https://doi.org/10.1016/j.rser.2017.02.085.
4. Granderson, J.; Touzani, S.; Custodio, C.; Sohn, M.D.; Jump, D.; Fernandes, S. Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings. *Appl. Energy* **2016**, *173*, 296–308. https://doi.org/10.1016/j.apenergy.2016.04.049.
5. Himeur, Y.; Ghanem, K.; Alsalemi, A.; Bensaali, F.; Amira, A. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Appl. Energy* **2021**, *287*, 116601. https://doi.org/10.1016/j.apenergy.2021.116601.
6. Miller, C.; Nagy, Z.; Schlueter, A. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1365–1377. https://doi.org/10.1016/j.rser.2017.05.124.
7. Li, G.; Zhan, L.; Fang, X.; Gao, J.; Xu, C.; He, X.; Deng, J.; Xiong, C. Performance comparison on improved data-driven building energy prediction under data shortage scenarios in four perspectives: Data generation, incremental learning, transfer learning, and physics-informed. *Energy* **2024**, *312*, 133640. https://doi.org/10.1016/j.energy.2024.133640.
8. Pinto, G.; Wang, Z.; Roy, A.; Hong, T.; Capozzoli, A. Transfer learning for smart buildings: A critical review of algorithms, applications, and future perspectives. *Adv. Appl. Energy* **2022**, *5*, 100084. https://doi.org/10.1016/j.adapen.2022.100084.
9. Iglesias, G.; Talavera, E.; González-Prieto, Á.; Mozo, A.; Gómez-Canaval, S. Data augmentation techniques in time series domain: A survey and taxonomy. *Neural Comput. Appl.* **2023**, *35*, 10123–10145. https://doi.org/10.1007/s00521-023-08459-3.
10. Wen, Q.; Sun, L.; Yang, F.; Song, X.; Gao, J.; Wang, X.; Xu, H. Time series data augmentation for deep learning: A survey. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), Montreal, QC, Canada, 19–27 August 2021; pp. 4653–4660. https://doi.org/10.24963/ijcai.2021/631.
11. Turowski, M.; Heidrich, B.; Weingärtner, L.; Springer, L.; Phipps, K.; Schäfer, B.; Mikut, R.; Hagenmeyer, V. Generating synthetic energy time series: A review. *Renew. Sustain. Energy Rev.* **2024**, *206*, 114842. https://doi.org/10.1016/j.rser.2024.114842.
12. Fang, H.; Tan, H.; Kosonen, R.; Yuan, X.; Jiang, K.; Ding, R. Study of the data augmentation approach for building energy prediction beyond historical scenarios. *Buildings* **2023**, *13*, 326. https://doi.org/10.3390/buildings13020326.
13. Wang, Z.; Hong, T. Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN). *Energy Build.* **2020**, *224*, 110299. https://doi.org/10.1016/j.enbuild.2020.110299.
14. Fu, C.; Kazmi, H.; Quintana, M.; Miller, C. Creating synthetic energy meter data using conditional diffusion and building metadata. *Energy Build.* **2024**, *312*, 114216. https://doi.org/10.1016/j.enbuild.2024.114216.
15. Kang, T.; Song, K. Data augmentation using multivariate time series decomposition for predicting daily energy consumption of new buildings. *Buildings* **2025**, *15*, 3467. https://doi.org/10.3390/buildings15193467.
16. Stenger, M.; Leppich, R.; Foster, I.; Kounev, S.; Bauer, A. Evaluation is key: A survey on evaluation measures for synthetic time series. *J. Big Data* **2024**, *11*, 66. https://doi.org/10.1186/s40537-024-00924-7.
17. Tashman, L.J. Out-of-sample tests of forecasting accuracy: An analysis and review. *Int. J. Forecast.* **2000**, *16*, 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0.
18. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028.
19. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE* **2018**, *13*, e0194889. https://doi.org/10.1371/journal.pone.0194889.
20. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 3rd ed.; OTexts: Melbourne, Australia, 2021. Available online: https://otexts.com/fpp3/ (accessed on 7 October 2026).
21. Mathieu, J.L.; Price, P.N.; Kiliccote, S.; Piette, M.A. Quantifying changes in building electricity use, with application to demand response. *IEEE Trans. Smart Grid* **2011**, *2*, 507–518. https://doi.org/10.1109/TSG.2011.2145010.
22. Zhang, L.; Wen, J.; Li, Y.; Chen, J.; Ye, Y.; Fu, Y.; Livingood, W. A review of machine learning in building load prediction. *Appl. Energy* **2021**, *285*, 116452. https://doi.org/10.1016/j.apenergy.2021.116452.
23. Huang, J.; Kaewunruen, S. Forecasting energy consumption of a public building using Transformer and support vector regression. *Energies* **2023**, *16*, 966. https://doi.org/10.3390/en16020966.
24. Taylor, S.J.; Letham, B. Forecasting at scale. *Am. Stat.* **2018**, *72*, 37–45. https://doi.org/10.1080/00031305.2017.1380080.
25. Xing, Z.; Pan, Y.; Yang, Y.; Yuan, X.; Liang, Y.; Huang, Z. Transfer learning integrating similarity analysis for short-term and long-term building energy consumption prediction. *Appl. Energy* **2024**, *365*, 123276. https://doi.org/10.1016/j.apenergy.2024.123276.

26. Ribeiro, M.; Grolinger, K.; ElYamany, H.F.; Higashino, W.A.; Capretz, M.A.M. Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. *Energy Build.* **2018**, *165*, 352–363. https://doi.org/10.1016/j.enbuild.2018.01.034.
27. Kang, Y.; Hyndman, R.J.; Li, F. GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. *Stat. Anal. Data Min.* **2020**, *13*, 354–376. https://doi.org/10.1002/sam.11461.
28. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 15. https://doi.org/10.1145/1541880.1541882.
29. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422. https://doi.org/10.1109/ICDM.2008.17.
30. Fan, C.; Xiao, F.; Zhao, Y.; Wang, J. Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data. *Appl. Energy* **2018**, *211*, 1123–1135. https://doi.org/10.1016/j.apenergy.2017.12.005.
31. Schmidl, S.; Wenig, P.; Papenbrock, T. Anomaly detection in time series: A comprehensive evaluation. *Proc. VLDB Endow.* **2022**, *15*, 1779–1797. https://doi.org/10.14778/3538598.3538602.
32. Wu, R.; Keogh, E.J. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 2421–2429. https://doi.org/10.1109/TKDE.2021.3112126.
33. Kim, S.; Choi, K.; Choi, H.S.; Lee, B.; Yoon, S. Towards a rigorous evaluation of time-series anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 7194–7201. https://doi.org/10.1609/aaai.v36i7.20680.
34. Han, J.; Chong, A. Hybrid-CGAN: A hybrid approach combining simulation and generative adversarial networks for fault detection and diagnosis in buildings. *Energy Build.* **2025**, *335*, 115531. https://doi.org/10.1016/j.enbuild.2025.115531.
35. Tian, C.; Li, M.; Yan, A.; Zhang, H.; Liu, Q. Research on anomaly detection of building energy consumption integrating density-aware oversampling and improved positional encoding. *Energy Build.* **2026**, *353*, 116922. https://doi.org/10.1016/j.enbuild.2025.116922.
36. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.
37. Ferrari Dacrema, M.; Cremonesi, P.; Jannach, D. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In Proceedings of the 13th ACM Conference on Recommender Systems, Copenhagen, Denmark, 16–20 September 2019; pp. 101–109. https://doi.org/10.1145/3298689.3347058.
38. Lipton, Z.C.; Steinhardt, J. Troubling trends in machine learning scholarship. *ACM Queue* **2019**, *17*, 45–77. https://doi.org/10.1145/3317287.3328534.
39. 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. https://doi.org/10.1038/s41597-020-00712-x.
40. Meta Open Source. Prophet: Trend Changepoints. Available online: https://facebook.github.io/prophet/docs/trend_changepoints.html (accessed on 4 November 2026).
41. ASHRAE. *ASHRAE Guideline 14-2014: Measurement of Energy, Demand, and Water Savings*; ASHRAE: Atlanta, GA, USA, 2014.
42. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall/CRC: New York, NY, USA, 1993.
43. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. https://doi.org/10.2307/3001968.
44. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
45. Fisher, R.A. *Statistical Methods for Research Workers*, 4th ed.; Oliver and Boyd: Edinburgh, UK, 1932.
46. Lakens, D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Personal. Sci.* **2017**, *8*, 355–362. https://doi.org/10.1177/1948550617697177.
47. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.