

Baseline AI: Explainable Machine Learning for Hourly Building Energy Forecasting, Waste Detection and Savings Verification

Abdelouahab KRIM

COMPILIX INNOVATIONS, research lab of SPASU BAC Students, Algeria

Abstract

Most buildings are monitored through a monthly bill, which hides when and why energy is wasted and makes the savings of efficiency projects hard to prove. We present **Baseline AI**, an open-source system that predicts the hourly electricity use of buildings from weather, calendar and meter history, and turns these predictions into two services: detection of abnormal consumption and weather-adjusted verification of retrofit savings. Baseline AI is developed and evaluated on 591 real buildings from four university campuses, 9.25 million hourly readings over 2016–2017. Two gradient-boosted tree models are trained with a strict time split and evaluated on six months never used for training or tuning. A *day-ahead* model, which also sees consumption up to the previous day, reaches a median coefficient of variation of the root mean squared error (CV-RMSE) of 13.8% across buildings, 40% below a seasonal “same hour last week” baseline, and 88% of buildings meet the usual hourly accuracy limits for building energy models (CV-RMSE $\leq 30\%$, |NMBE| $\leq 10\%$). A *weather model*, which sees only weather, calendar and the building’s historical profile, reaches 21.5% and serves as the benchmark of expected use. In controlled experiments on real meter data, the system flags 74% of days carrying a +50% fault and 55% of days carrying a +30% fault while flagging 3.6% of normal days, and measures a known 15% saving to within 10 percentage points in 65% of buildings over three months. SHAP explanations show that building-specific weekly profiles and recent load dominate the forecasts, with a clear cooling signature at high temperature. We report the method, all results and the limits of the approach, and release the code, an interactive dashboard and a web platform.

Keywords: building energy; load forecasting; gradient boosting; LightGBM; explainable AI; SHAP; anomaly detection; measurement and verification; energy efficiency.

1 Introduction

Buildings are among the largest consumers of final energy and a major source of energy-related emissions, and much of their consumption is driven by heating, cooling, ventilation, lighting and plug loads that respond to weather and occupancy [1]. A large share of this energy is lost to avoidable causes: equipment that runs outside occupied hours, mistuned schedules, stuck valves or failed sensors. These faults are invisible in monthly bills. The underlying problem is the absence of a benchmark: without knowing how much a building *should* have used in a given hour, an excess cannot be detected, and the effect of an efficiency project cannot be separated from changes in weather and use.

Physics-based simulation tools can provide such a benchmark [4], but each building needs a detailed model of its geometry, envelope and systems, which costs weeks of expert work and does not scale to large portfolios. Data-driven models learn the relationship between consumption, weather and time directly from meter data [1]. With the spread of smart meters, they can cover thousands of buildings at almost no marginal cost, which is what large building portfolios and public efficiency programmes require.

This paper presents Baseline AI, a complete, reproducible system for hourly building energy prediction developed at COMPILIX INNOVATIONS. Its contributions are:

1. a data pipeline that cleans real meter data and builds calendar, weather, building-profile and lagged-load features for hundreds of buildings at once;
2. two complementary models, a *weather model* for expected consumption and a *day-ahead model* for operational forecasting, compared with two baselines, ridge regression and random forests on six unseen months under a leakage-free validation protocol;
3. two applications built on the weather model, anomaly detection with rolling calibration and savings verification with a retrained baseline, each validated with controlled experiments on real data;
4. explanations of every prediction with SHAP values, and an open implementation with a dashboard and a web platform.

Section 2 reviews related work. Section 3 describes the data and Section 4 the method. Section 5 reports the results, Section 6 discusses impact and limitations, and Section 8 concludes.

2 Related work

Data-driven building energy prediction. Reviews of the field report a steady shift from engineering and statistical models towards machine learning, with artificial neural networks, support vector regression and tree ensembles as the most common choices, and hourly forecasting of electricity as the most studied task [1]. A large public benchmark was set by a forecasting competition on more than a thousand buildings, whose winning solutions relied mainly on gradient-boosted decision trees with careful data cleaning and feature engineering [15]. The meter data of that competition was later released, together with metadata and weather, as the Building Data Genome 2 dataset [14], which we use here. More recently, large-scale benchmarks have studied short-term load forecasting with deep models pre-trained on simulated building stocks and evaluated on real buildings [5].

Models. Gradient-boosted trees, in particular LightGBM [9], handle heterogeneous tabular features, categorical variables with many levels and millions of rows on a single machine. Random forests [3] are a robust alternative. Recurrent networks [8] and attention-based architectures such as the Temporal Fusion Transformer [11] model sequences directly and are natural extensions of this work.

Measurement and verification. Savings from efficiency measures are usually estimated by comparing post-retrofit consumption with an *adjusted baseline*, the consumption predicted by a model trained before the retrofit, as formalised in standard measurement and verification (M&V) protocols [2, 6]. Automated M&V with data-driven baselines has been shown to reach useful accuracy on commercial buildings [7]. The same protocols define the goodness-of-fit criteria used in this paper.

Explainability. SHAP values attribute each prediction to its input features with guarantees of local accuracy and consistency [12], and can be computed exactly and quickly for tree ensembles [13]. They make it possible for engineers to check whether a model relies on physically plausible drivers.

3 Data

3.1 Source

We use the cleaned hourly electricity meters of the Building Data Genome 2 dataset [14], in the distribution prepared for the BuildingsBench benchmark [5]. It covers four university campuses, identified in the dataset by the pseudonyms Bear, Fox, Panther and Rat, with hourly outdoor air

temperature and relative humidity for each site. The primary use of each building (education, office, lodging, ...) is encoded in its meter name. Table 1 summarises the data after cleaning.

Table 1. Data after cleaning, by site. Temperatures are hourly outdoor values over 2016–2017.

Site	Buildings	Hourly readings	Mean load (kWh/h)	Temperature, mean [min, max] (°C)
Bear	91	1 479 436	181.9	15.1 [0.6, 38.3]
Fox	133	2 288 123	153.0	25.1 [2.2, 48.3]
Panther	105	918 135	106.9	22.8 [2.2, 35.0]
Rat	262	4 560 887	114.0	15.7 [−10.6, 37.8]
Total	591	9 246 581	133.8	

The portfolio is dominated by education buildings (243), followed by public (94), assembly (78), office (70) and lodging (52) buildings; parking, retail, health, warehouse and five smaller categories make up the rest. In total the meters record 1.24 TWh of electricity. Panther has data for 2017 only, which makes it a useful test of buildings with short histories.

3.2 Cleaning

Real meter data contains readings that do not reflect consumption. We apply three rules per building: (i) non-positive readings are removed; (ii) runs of 24 or more identical consecutive readings are removed as frozen meters; (iii) readings above five times the building’s 99th percentile are removed as spikes. These rules removed 41 587 readings. Weather series are aligned to the hourly grid and gaps of up to six hours are interpolated. Buildings are kept when they have at least 2 000 valid hours in the training period and 1 000 in the test period, which retains 591 of 611 buildings.

4 Method

Figure 1 gives an overview of the system.

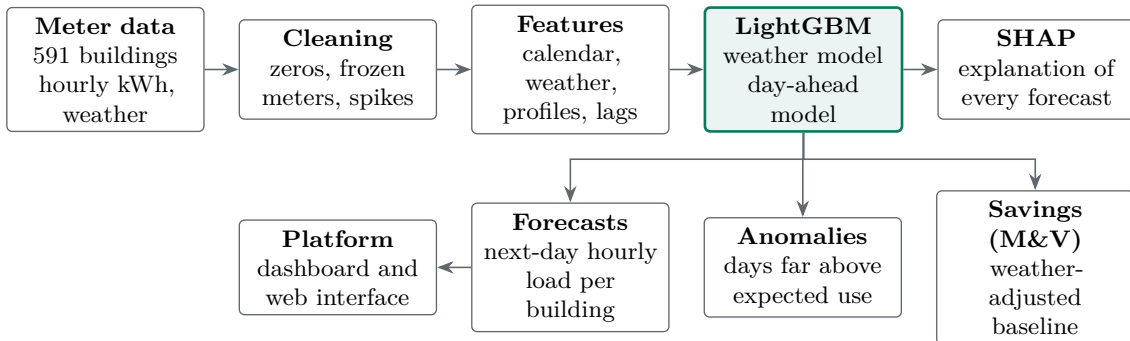


Figure 1. Overview of Baseline AI. Both models share the cleaning and feature pipeline; the weather model supports anomaly detection and savings verification, the day-ahead model supports operational forecasts. All outputs are explained with SHAP and exposed in a dashboard and a web platform.

4.1 Problem formulation

Let $y_{b,t}$ be the electricity use of building b in hour t . We learn a function f that maps a feature vector $\mathbf{x}_{b,t}$ to $\log(1 + y_{b,t})$ and predict $\hat{y}_{b,t} = \exp(f(\mathbf{x}_{b,t})) - 1$. Working on the logarithm makes errors relative, so that a hospital and a small office weigh comparably in training, and stabilises the strongly skewed distribution of loads. One global model is trained for all buildings.

We consider two tasks that differ only in the information available:

- **Weather model** (expected use). $\mathbf{x}_{b,t}$ contains the building identity and type, the site, calendar and weather variables and the building’s historical profiles, but no recent meter readings. Its

prediction is the consumption that weather and calendar explain, which is the quantity anomaly detection and M&V require: a fault or a retrofit must not leak into the benchmark.

- **Day-ahead model** (operational forecast). $\mathbf{x}_{b,t}$ also contains loads observed at least 24 hours before t , i.e. information available when a forecast for the next day is issued at midnight.

4.2 Features

Table 2 lists the features. Building profiles act as target encodings: they are averages of $\log(1 + y)$ computed on the training period only. The *recent* profiles cover the last eight weeks of training and capture slow drifts in a building’s level. Lagged features use only data at least 24 hours old.

Table 2. Features. The weather model uses the first four groups; the day-ahead model uses all five.

Group	Features
Identity	building identifier, site, building type (categorical)
Calendar	hour of day, day of week, month, hour of week, weekend flag, public-holiday flag
Weather	temperature, relative humidity, dew point (Magnus formula), cooling and heating degree hours (base 18 °C), 24 h and 72 h mean temperature, temperature 3 h earlier, daily maximum and minimum
Building profile	mean log-load; mean by hour of week; mean by month; recent mean and recent mean by hour of week (last 8 weeks of training)
Lagged load	load 24 h, 48 h, 168 h and 336 h earlier; mean of the previous day; mean of the previous week

4.3 Models

We compare five approaches. **Weekly-profile baseline:** the building’s training-period mean for the same hour of the week. **Last-week baseline** (day-ahead task): the load of the same hour one week earlier. **Ridge regression** on standardised numerical features and one-hot building types ($\alpha = 1$). **Random forest** [3] with 60 trees, maximum depth 24, minimum leaf size 5 and 60% of features per split, trained on a random sample of 600 000 training rows for tractability. **LightGBM** [9] with squared loss on $\log(1 + y)$, learning rate 0.05, 63 leaves, minimum leaf size 500, feature fraction 0.85, bagging fraction 0.8 and ℓ_2 regularisation 1, with building, site and type as native categorical features. We also report an ensemble of the LightGBM and random-forest weather models (weights 0.7 and 0.3).

4.4 Validation protocol

All models are trained on 1 January 2016 to 30 June 2017 (6.64 million rows) and tested on 1 July to 31 December 2017 (2.60 million rows). No test data is used for training, feature construction or any modelling decision. Feature sets, hyper-parameters and the number of boosting rounds are chosen on a *validation* period, May–June 2017, using models trained on data before May 2017 *and building profiles recomputed from that earlier data only*. This last point matters: in a first version, profiles that included the validation months made the validation error optimistic and stopped boosting after about 45 rounds. The number of rounds selected on the leak-free validation period (175 for the weather model, 214 for the day-ahead model) is increased by 10% for the final fit on the full training period.

4.5 Metrics

For each building, with n test hours, observed loads y_i , predictions $\hat{y}_i$ and mean $\bar{y}$:

$$\text{CV-RMSE} = \frac{100}{\bar{y}} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad \text{NMBE} = \frac{100}{n \bar{y}} \sum_{i=1}^n (y_i - \hat{y}_i). \quad (1)$$

These are the goodness-of-fit measures of M&V practice, where hourly models are commonly accepted when $\text{CV-RMSE} \leq 30\%$ and $|\text{NMBE}| \leq 10\%$ [2]. We report the median over buildings, the share of buildings meeting both limits, and, over all test hours, the root mean squared logarithmic error (RMSLE) and the mean absolute error (MAE).

4.6 Explainability

For each LightGBM model we compute exact TreeSHAP values [13] on 20 000 random test hours, in units of $\log(1 + y)$, and summarise feature importance by the mean absolute SHAP value.

4.7 Anomaly detection

Detection works on daily totals. For building b and day d , let A_d be the metered energy and E_d the sum of the weather model’s hourly predictions. Because buildings drift over months, the expectation is re-calibrated with the ratio of metered to expected energy over 28 days ending 7 days before d ,

$$c_d = \frac{\sum_{k=d-34}^{d-7} A_k}{\sum_{k=d-34}^{d-7} E_k}, \quad r_d = \frac{A_d - c_d E_d}{c_d E_d}. \quad (2)$$

The seven-day gap prevents an ongoing fault from raising its own expectation. A day is flagged when

$$r_d > \max\left(0.15, \text{med}_b(r) + 2.5 \times 1.4826 \text{ MAD}_b(r)\right), \quad (3)$$

that is, when consumption is at least 15% above expectation and an outlier for that building according to the robust median absolute deviation [10]. Consecutive flagged days form an event, ranked by excess energy in kWh.

To measure detection performance we inject synthetic faults into real test data: in 100 randomly chosen buildings, consumption is multiplied by 1.3 (or 1.5) during five consecutive days at a random date, and we record the share of fault days flagged and the share of other days flagged in the same buildings.

4.8 Savings verification

Following the adjusted-baseline approach of M&V protocols [6, 7], a retrofit at date t_r is evaluated by training the weather model on all data before t_r (with profiles recomputed up to t_r) and predicting the post-retrofit baseline $\hat{B}_t$. The estimated saving over the reporting period is

$$\hat{S} = 1 - \frac{\sum_{t \geq t_r} A_t}{\sum_{t \geq t_r} \hat{B}_t}. \quad (4)$$

We validate this with $t_r = 1$ October 2017 and a three-month reporting period. We apply a known saving of 15% to the metered data after t_r and measure the error $\hat{S} - 0.15$ in each building. We also run a placebo test without any saving, where the ideal estimate is zero.

5 Results

5.1 Model selection on the validation period

Table 3 shows the variants compared on May–June 2017. Smaller, more regularised trees with the recent-profile features were best for both tasks. Removing the day-of-year feature alone did not help, but it was kept out of the final models because, with only 18 months of history, it allows a model to memorise specific dates. The test period played no role in these choices.

5.2 Accuracy on unseen months

Table 4 and Figure 2 report test-period accuracy. For the day-ahead task, LightGBM reduces the median CV-RMSE from 22.95% for the last-week baseline to 13.76%, a 40% reduction, and raises

Table 3. Variants compared on the validation period (May–June 2017). “Original” trees: 255 leaves, minimum leaf size 100, learning rate 0.08; “smaller” trees: 63 leaves, minimum leaf size 500, learning rate 0.05.

Task	Variant	Rounds	Median CV-RMSE (%)	Median NMBE (%)
Weather	original features and trees	45	22.12	8.80
	+ recent profiles	49	21.84	8.47
	+ recent profiles, no day of year	40	22.54	9.29
	+ recent, no day of year, smaller trees	175	21.49	7.95
Day-ahead	original features and trees	46	15.46	3.12
	+ recent profiles, no day of year	45	15.65	3.39
	+ recent, no day of year, smaller trees	214	14.95	2.18

the share of buildings within the hourly accuracy limits from 73% to 88%. For the weather task, LightGBM improves on the weekly-profile baseline from 28.4% to 21.5% and on ridge regression (24.9%); it is comparable to the random forest (21.8%), and the ensemble of the two brings a marginal further gain (21.3%). The bias of all weather models (median |NMBE| around 6.5%) is larger than that of the day-ahead model (1.3%), which shows that much of the remaining error comes from slow changes in buildings that weather and calendar cannot explain.

Table 4. Test-period accuracy (1 July – 31 December 2017, 591 buildings). CV-RMSE, NMBE and the share within limits are computed per building; RMSLE and MAE over all test hours. Best value per task in bold.

Task	Model	Median CV-RMSE (%)	Median NMBE (%)	RMSLE	MAE (kWh)	Within limits (%)
Weather	Weekly-profile baseline	28.38	8.63	0.421	24.45	44.3
	Ridge regression	24.87	7.63	0.447	25.66	51.6
	Random forest	21.77	6.67	0.432	23.34	57.9
	LightGBM	21.49	6.43	0.425	22.86	58.9
	Ensemble LightGBM + RF	21.30	6.51	0.426	22.79	58.5
Day-ahead	Same hour last week	22.95	1.19	0.324	15.22	72.8
	LightGBM	13.76	1.29	0.206	10.60	88.2

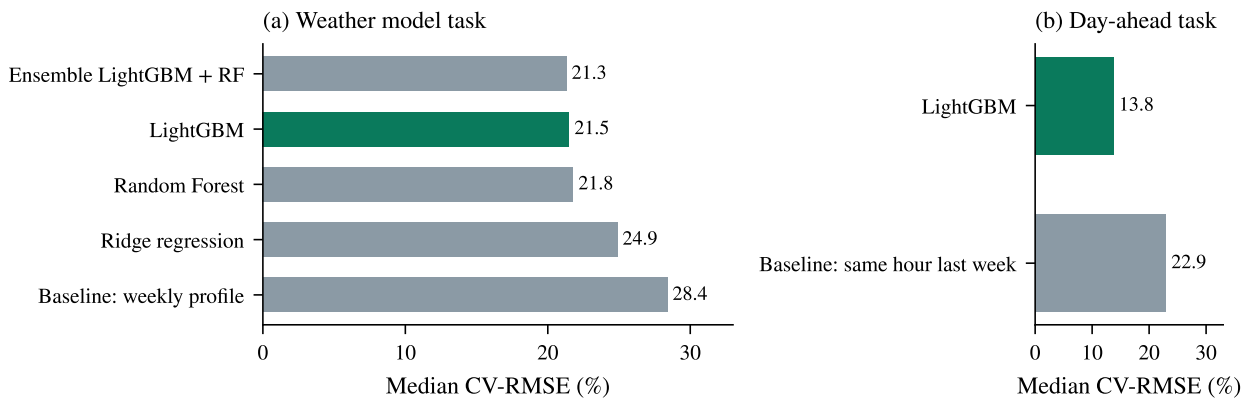


Figure 2. Median CV-RMSE on the test period. LightGBM models in green.

Figure 3 shows the distribution over buildings. For the day-ahead model, the CV-RMSE has quartiles of 9.9%, 13.8% and 19.6%, and 90% of buildings stay below 29.6%. The weather model has quartiles of 14.7%, 21.5% and 33.1%, with a long tail of buildings whose consumption changed abruptly after the training period. Figure 4 shows two weeks for the median building of three common types: both models follow the daily and weekly cycles, and the day-ahead model also tracks changes of level.

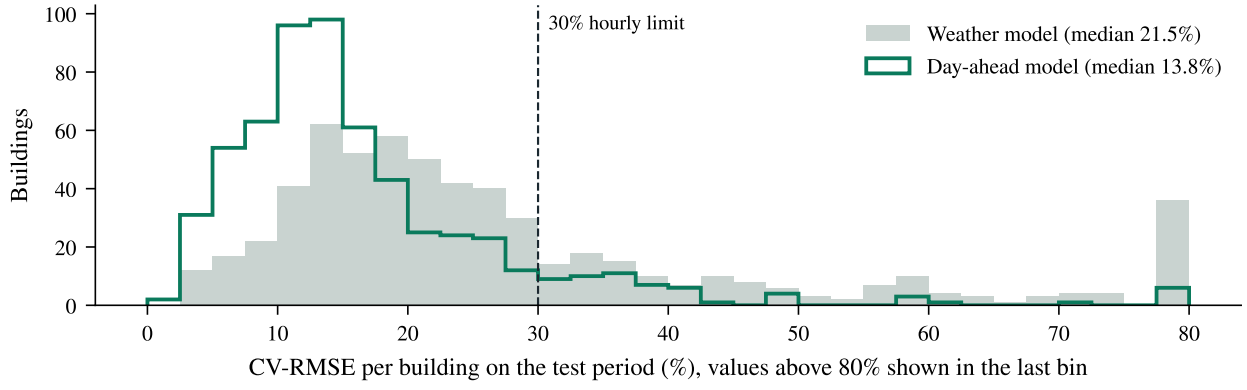


Figure 3. Distribution of per-building CV-RMSE on the test period. Dashed line: 30% hourly limit.

Table 5 breaks the error down by building type. Lodging, parking and education buildings, whose occupancy follows regular schedules, are the easiest to predict. Assembly, retail and warehouse buildings, whose use depends on events and operations not visible in the features, are the hardest.

Table 5. Median CV-RMSE (%) by building type, test period. Types with at least seven buildings.

Building type	Buildings	Weather model	Day-ahead model
Education	243	20.1	13.1
Public	94	19.4	14.1
Assembly	78	33.2	17.7
Office	70	21.5	13.8
Lodging	52	19.5	11.4
Parking	15	14.4	7.8
Other	9	19.7	13.2
Retail	8	29.9	20.4
Health	7	21.4	14.1
Warehouse	7	57.7	20.8

5.3 Explanations

Figure 5 ranks features by mean absolute SHAP value. The weather model relies first on the building’s weekly and monthly profiles, then on its identity, mean level and outdoor temperature. The day-ahead model relies first on the load 24 hours earlier, then on the weekly profile and the loads one week and 48 hours earlier. Figure 6 isolates the contribution of temperature and dew point in the weather model: it is lowest around 15–20 °C and rises steeply above roughly 25 °C, the cooling signature expected on these campuses; it also rises, more mildly, at low temperatures, consistent with some electric heating. The models therefore use physically plausible drivers.

5.4 Anomaly detection

On the test period, the detector flags 4038 building-days (3.7% of days), grouped into 1600 events, most of them lasting one or two days. Figure 7 shows a multi-day event in one building. In the fault-injection experiment (Table 6), 74.2% of days with a +50% fault and 54.6% of days with a +30% fault are detected, while 3.4–3.6% of normal days are flagged, about one day per month per building. The rolling calibration matters: without it, the same detector caught 50.6% of +30% fault days and flagged 3.2% of normal days.

5.5 Savings verification

With the baseline retrained up to the retrofit date, the estimated saving has a median absolute error of 6.7 percentage points against a true 15% saving, 42.5% of buildings are within 5 points and

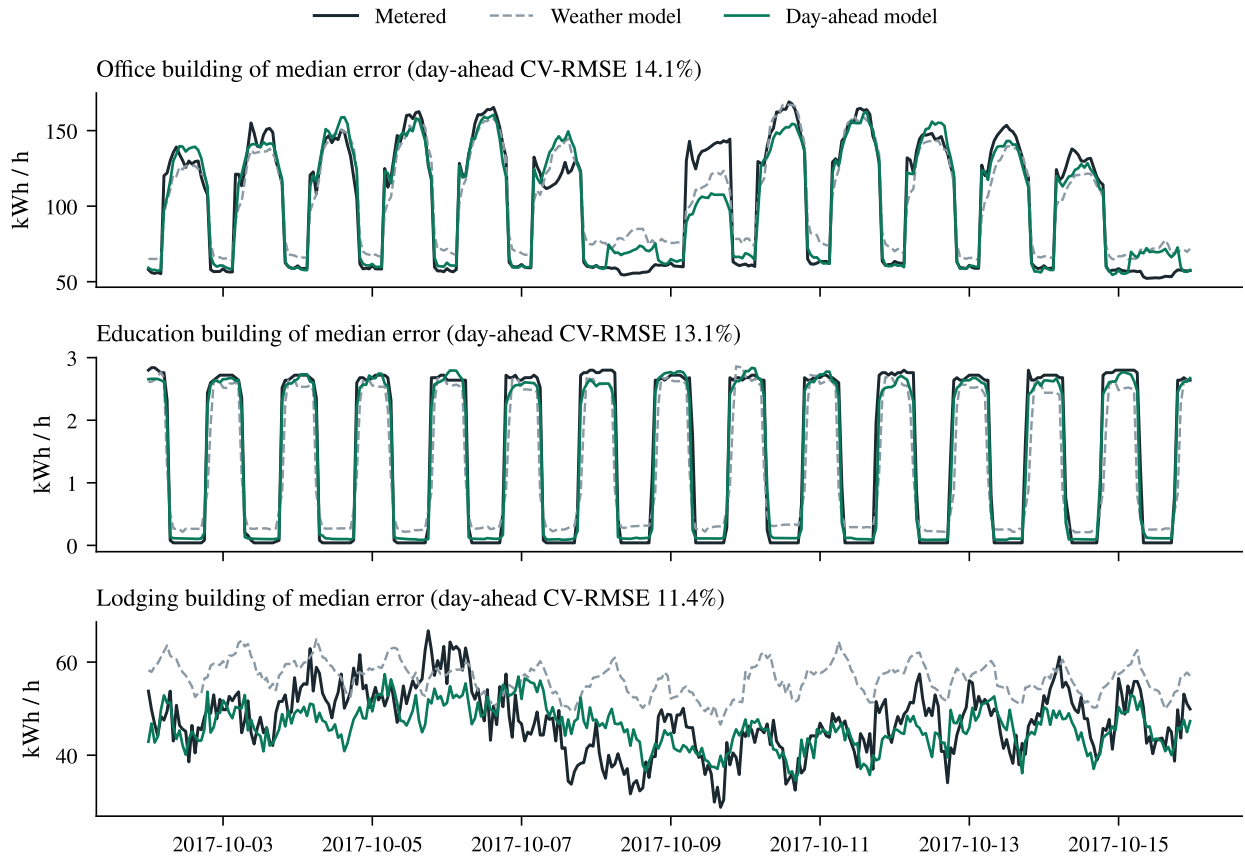


Figure 4. Metered load and predictions for 2–15 October 2017, for the building of median day-ahead error among offices, education and lodging buildings.

64.8% within 10 points (Figure 8). The placebo estimates are centred close to zero (median -1.6%), so the method is nearly unbiased across the portfolio. A simpler variant that only rescaled the original weather model with one calibration factor gave a median error of 7.6 points and a placebo median of -3.1% . At the level of single buildings and three months of data, the uncertainty is therefore of the order of the savings of light retrofits; the method is reliable for large savings, for portfolios, or with longer reporting periods.

5.6 Computational cost

The complete training script (features for the validation and test protocols, two baselines, ridge, random forest and both LightGBM models, predictions for 9.2 million rows and evaluation) runs in 9.3 minutes on a 4-core CPU with 15 GB of memory. The full pipeline, including data download, explanations and validation experiments, runs in about 30 minutes. No GPU is needed.

6 Discussion

Impact. The results support three uses. For facility teams, a daily comparison with the calibrated weather model surfaces large waste events within days, with about one false alarm per building per month. For utilities and operators, day-ahead forecasts within the usual accuracy limits for 88% of buildings can be aggregated into campus or feeder load and support demand response. For efficiency programmes and their funders, the retrained baseline gives a nearly unbiased, weather-adjusted estimate of savings across a portfolio. Because one global model covers all buildings and needs only meter and weather data, the cost of adding a building is negligible, in contrast with per-building engineering simulation.

What limits accuracy. The gap between the weather model (21.5%) and the day-ahead

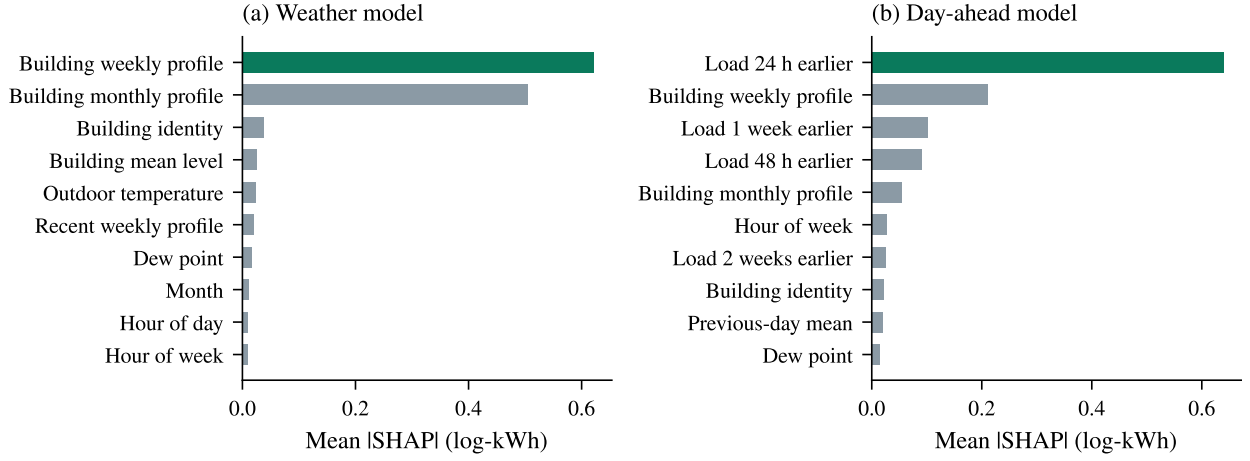


Figure 5. Top ten features by mean absolute SHAP value (20 000 test hours, log-kWh units).

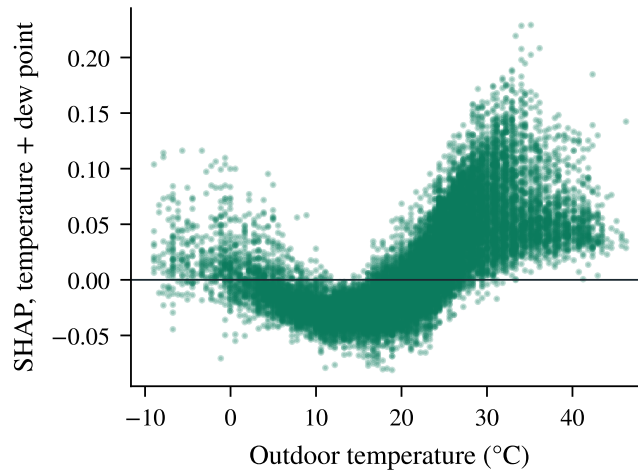


Figure 6. SHAP contribution of temperature and dew point in the weather model as a function of outdoor temperature.

model (13.8%) is mostly bias: buildings change over months in ways that weather and calendar do not explain (new equipment, occupancy, operating schedules). Recent-profile features and rolling calibration reduce, but do not remove, this effect.

Limitations. (i) Weather is limited to temperature and humidity; solar radiation and cloud cover are not available in this distribution of the data. (ii) Floor area and occupancy are not used. (iii) Faults in the detection experiment are synthetic step changes; real faults, such as gradual degradation or schedule errors, may be easier or harder to detect, and labelled fault data from facility teams would give a stronger test. (iv) The rolling calibration adapts to a fault that lasts several weeks: after the seven-day gap, a long fault raises its own expectation and is no longer flagged, so long-lasting excess must be tracked with a fixed reference period. (v) The buildings are university campuses in four locations; results may differ for other climates and building stocks. (vi) Deep sequence models [8, 11] were not evaluated in this study.

7 Software and platform

Baseline AI is implemented in Python with pandas, scikit-learn [16], LightGBM and SHAP. One command downloads the data and runs cleaning, training, explanation and validation; a separate importer prepares any hourly meter file with the columns `timestamp`, `building_id`, `kwh`, `temperature` and `humidity`. A Streamlit dashboard shows forecasts, anomalies, SHAP explanations and a savings calculator for every building, and a web platform presents the results

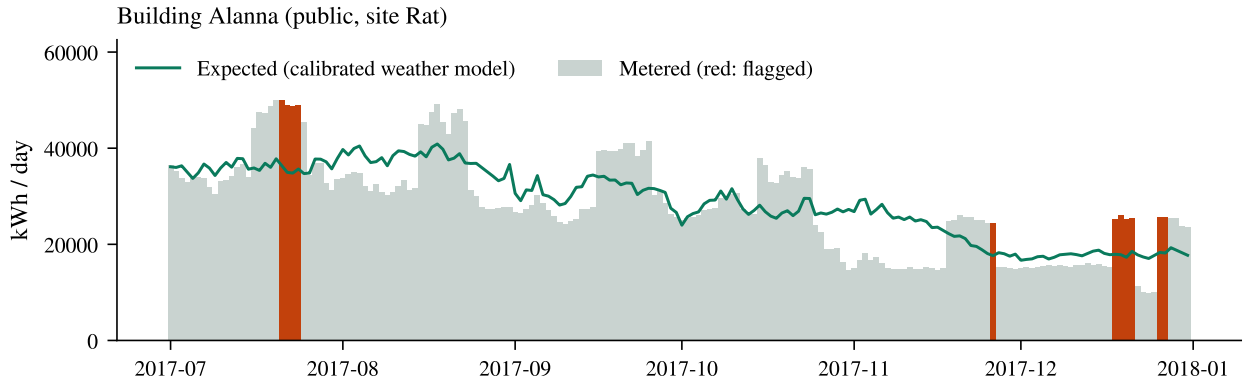


Figure 7. Daily metered and expected energy for one building over the test period; flagged days in red.

Table 6. Validation of the applications on real test data.

Experiment	Setting	Result
Fault detection	+50% for 5 days, 100 buildings	74.2% of fault days flagged
Fault detection	+30% for 5 days, 100 buildings	54.6% of fault days flagged
False alarms	normal days in the same buildings	3.4–3.6% flagged
Savings (M&V)	15% saving after 1 Oct. 2017, 3 months, 591 buildings	median error 6.7 points 42.5% within 5 points 64.8% within 10 points
Placebo	no saving applied	median estimate −1.6%

for representative buildings in the browser at <https://baselineai.bacai.space>. The code is available at <https://github.com/ouahab4/baseline-ai>.

8 Conclusion

Baseline AI shows that a single, explainable gradient-boosting system trained on meter and weather data can forecast the hourly electricity use of hundreds of real buildings within standard accuracy limits, and can turn these forecasts into validated tools for waste detection and savings verification. On six unseen months, the day-ahead model reduces the error of a seasonal baseline by 40%, and the weather model provides an unbiased benchmark of expected use at the portfolio level. Future work at COMPILIX INNOVATIONS will add solar radiation and occupancy proxies, probabilistic (quantile) forecasts, transfer learning to buildings with short histories, comparison with deep sequence models, and control strategies that use the forecasts for demand flexibility.

Data availability

The meter and weather data are from the open Building Data Genome 2 dataset [14], released under the CC BY 4.0 licence, accessed through the BuildingsBench distribution [5]. All results in this paper are produced by the released code.

Acknowledgements

The author thanks the creators and maintainers of the open datasets used in this work.

References

- [1] K. Amasyali and N. M. El-Gohary. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81:1192–1205, 2018.
- [2] ASHRAE. *Guideline 14-2014: Measurement of Energy, Demand, and Water Savings*. Atlanta, GA,

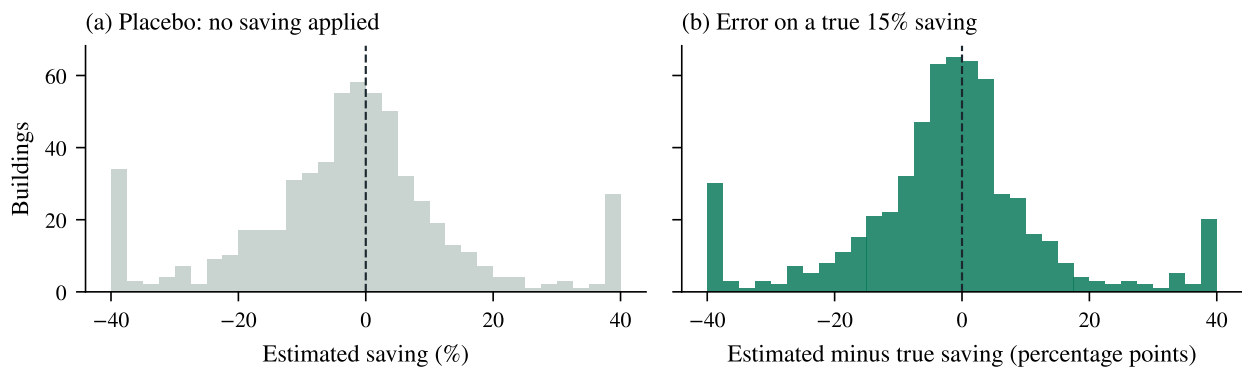


Figure 8. Savings verification over three months for 591 buildings. (a) Estimated saving when nothing changed. (b) Error of the estimate when a true 15% saving is applied. Values beyond ± 40 are shown at the edges.

2014.

- [3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [4] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, W. F. Buhl, Y. J. Huang, C. O. Pedersen, R. K. Strand, R. J. Liesen, D. E. Fisher, M. J. Witte, and J. Glazer. EnergyPlus: creating a new-generation building energy simulation program. *Energy and Buildings*, 33(4):319–331, 2001.
- [5] P. Emami, A. Sahu, and P. Graf. BuildingsBench: A large-scale dataset of 900K buildings and benchmark for short-term load forecasting. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023.
- [6] Efficiency Valuation Organization. *International Performance Measurement and Verification Protocol: Core Concepts*. 2022.
- [7] J. Granderson, S. Touzani, C. Custodio, M. D. Sohn, D. Jump, and S. Fernandes. Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings. *Applied Energy*, 173:296–308, 2016.
- [8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [9] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*, pages 3146–3154, 2017.
- [10] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.
- [11] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.
- [12] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 2017.
- [13] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.
- [14] C. Miller, A. Kathirgamanathan, B. Picchetti, P. Arjunan, J. Y. Park, Z. Nagy, P. Raftery, B. W. Hobson, Z. Shi, and F. Meggers. The Building Data Genome Project 2, energy meter data from the ASHRAE Great Energy Predictor III competition. *Scientific Data*, 7:368, 2020.
- [15] C. Miller, P. Arjunan, A. Kathirgamanathan, C. Fu, J. Roth, J. Y. Park, C. Balbach, K. Gowri, Z. Nagy, A. D. Fontanini, and J. Haberl. The ASHRAE Great Energy Predictor III competition: Overview and results. *Science and Technology for the Built Environment*, 26(10):1427–1447, 2020.
- [16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.