2026 European Conference on Computing in Construction
Corfu, Greece
July 12-15, 2026

A MONITORING-ORIENTED DIGITAL TWIN PIPELINE FOR DATA-DRIVEN PREDICTION OF RESIDENTIAL PROSUMER ENERGY PROFILES

Amin Rahmani[1], Mahsa Moradian[1], Sonia Lupica Spagnolo[1], and J.J. McArthur[2]
[1] Politecnico di Milano, Milan, Italy
[2] Toronto Metropolitan University, Toronto, Canada

Abstract

Residential prosumer buildings with PV and high-load subsystems exhibit irregular demand, yet current residential digital-twin and baselining studies rarely address missing-data realism and causal deployment constraints. This paper presents a monitoring-oriented operation-phase digital twin analytics pipeline for an instrumented villa in Segrate, Italy, linking whole-building demand, PV output, pool-circuit power, and outdoor air temperature to a one-hour-ahead baseline task. A fixed Random Forest is evaluated across eight dataset configurations against persistence and Multiple Linear Regression. Under deployment-realistic causal conditions, the best configuration achieves RMSE 576.5 W, demonstrating the value of a traceable leakage-safe baselining workflow for residential prosumer settings.

Introduction

Digital twins, which were initially developed as models for the design phase, are now being used as effective tools to monitor and manage building operations (Oulefki et al., 2025; Mengual Torres et al., 2022). In this setting, data-driven energy baselining has emerged as a method to generate short-horizon reference predictions of building demand, supporting fault detection, performance tracking, and measurement and verification (Xie et al., 2023).

In the residential sector, smart home automation and IoT monitoring have been extensively studied, but residential digital twins are still less prevalent compared to their use in commercial facility management (Mengual Torres et al., 2022). Many existing residential studies concentrate on basic monitoring dashboards or the optimisation of specific components like batteries or heat pumps, lacking a robust and traceable connection between sub-metered devices and operational building-level representations (Xie et al., 2023). Residential prosumer buildings are also more difficult to model than conventional offices due to more stochastic occupancy patterns and the presence of onsite generation and high-load amenities, such as rooftop photovoltaics and swimming pools, which lead to non-linear and intermittent demand profiles necessitating comparison against both persistence and linear benchmark models to isolate the value of complex learners (Ghaemi et al., 2024). In the absence of a dependable short-horizon baseline that reflects these factors, building owners and facility managers do not have a clear reference point for detecting anomalies or assessing system upgrades (Ghaemi et al., 2024; Xie et al., 2023).

This gap is addressed in this paper using an instrumented residential villa in Segrate (Milan, Italy) as a documented operational-phase digital-twin testbed, linking heterogeneous IoT telemetry (whole-building electrical demand, PV production, swimming-pool circulation pump load, and outdoor air temperature) into a traceable data-processing pipeline that enforces domain constraints and prevents time leakage. This paper investigates a monitoring-oriented operation-phase digital twin analytics pipeline and documented residential testbed for causally consistent short-horizon baselining, addressing a gap narrower than a general digital-twin one: few residential prosumer studies jointly examine one-hour-ahead operational baselining, sub-metered telemetry, explicit missing-data policy, causal versus non-causal imputation, and comparison against persistence and linear regression within a leakage-safe design (Krishnan et al., 2024). Here, digital twin denotes a traceable operational representation linking telemetry, subsystem context, and analytics; it does not imply bidirectional control or formal BIM integration (Xie et al., 2023).

The one-hour-ahead horizon is chosen because the baseline serves as an operational reference for residual-based anomaly screening rather than long-range forecasting. At this horizon the baseline can support near-real-time fault detection in subsystems such as the pool circulation pump or the PV array, and it can provide building owners or facility managers with a timely reference for detecting deviations before they propagate into larger energy waste or equipment damage. The research question is whether adding sub-metered telemetry improves one-hour-ahead residential demand baselines over demand-only models and how that benefit varies across dataset configurations that differ in imputation strategy and missing-data policy.

Related work

Operation-phase monitoring pipelines and metadata traceability in residential digital twins

Operation-phase digital twin applications rely on robust data integration and metadata management to connect time-series data to building context (Xie et al., 2023). Recurring challenges in metadata consistency for smart-

building analytics motivate digital twin architectures that incorporate a traceable data integration layer linking building entities to operational data streams through curated metadata configurations (Balaji et al., 2018; Oulefki et al., 2025).

Data-driven energy prediction for occupied buildings

Building energy prediction has been extensively examined through three primary modelling paradigms: physics-based ("white-box"), data-driven ("black-box"), and hybrid ("grey-box") approaches. Reviews commonly observe that white-box models offer interpretability, but demand detailed and precise building parameters along with significant calibration efforts. In contrast, data-driven approaches are generally easier to implement when historical operational data are accessible and can deliver strong forecasting performance (Krishnan et al., 2024).

For operation-phase baselining and short-horizon forecasting, data-driven methods are frequently adopted as practical predictors, provided that issues such as data quality, missing values, and non-stationary operation (e.g., discontinuities from in-use residential operation) are appropriately addressed during preprocessing and evaluation (Krishnan et al., 2024). Within this framework, the methodological significance of baseline construction lies not only in final accuracy but also in how the inclusion of explanatory operational features, such as weather data, subsystem telemetry, and temporal context, contributes to the reliability and interpretability of predictive references relative to naïve benchmarks (Zhou, Yang & Wang, 2024). However, many studies still treat data handling and deployment constraints as secondary implementation aspects rather than as central methodological considerations (Krishnan et al., 2024; Xie et al., 2023).

The prosumer challenge and subsystem variability

Operation-phase analytics is more complex in residential prosumer contexts, where onsite generation and flexible or high-load subsystems lead to intermittent and non-linear demand patterns (Ghaemi et al., 2024). Recent peer-reviewed surveys on AI-enabled digital twins and data-driven energy management in buildings indicate that scalable applications such as forecasting and operational analytics require reliable data pipelines and consistent contextual information. Despite this, many residential studies continue to focus on component-level optimization, such as PV self-consumption, often without addressing the construction of a causally consistent whole-building operational baseline (compared against persistence and linear regression reference models) supported by sub-metered data (Xie et al., 2023).

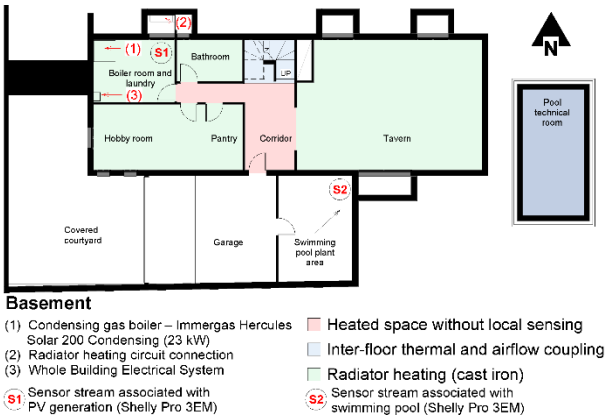
Recent residential short-term forecasting studies provide the closest benchmark context for this deployment-oriented case study. Kychkin & Chasparis (2025) conduct a broad comparative analysis of hourly and day-ahead load forecasting models for a small residential energy community, demonstrating that computationally efficient persistence-based and autoregressive models can match or outperform more complex deep learning approaches on aggregate community load. Popławski, Dudzik & Szeląg (2023) apply LASSO regression and Random Forest to forecast the net electricity balance of a single prosumer household with rooftop PV, using short training windows and both endogenous and exogenous scenarios. Neither study jointly addresses sub-metered high-load subsystem telemetry, explicit causal versus non-causal imputation policy, or a leakage-safe multi-configuration evaluation design under deployment-realistic missing-data conditions. The gap addressed here is therefore narrower than a general residential forecasting gap: this study examines one-hour-ahead operational baselining with sub-metered pool-circuit and PV telemetry under systematically varied imputation strategies and a chronologically enforced train/test split designed to reflect realistic deployment constraints.

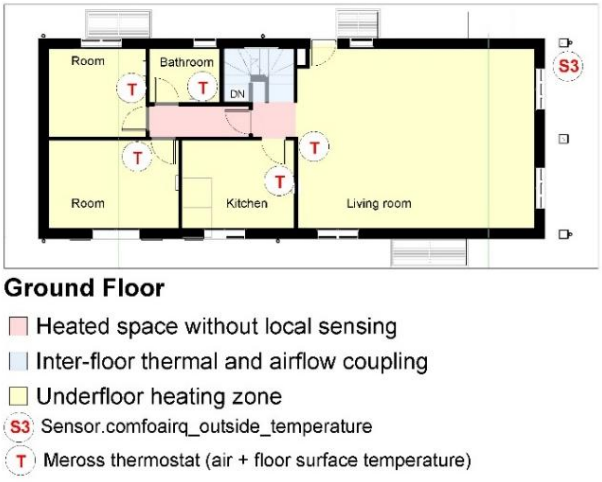
Methodology

This study implements a monitoring-oriented operation-phase digital twin analytics pipeline for operational energy baselining in a residential prosumer context following a three-stage empirical approach: characterising the case study, deploying the data integration and analytics architecture, and constructing and evaluating the operational baseline (Krishnan et al., 2024).

Case study apparatus: the monitored villa

The testbed is a detached residential villa situated in Segrate (Milan), Italy. It consists of three floors and has numerous sensors installed throughout (Fig. 1). Built in 1993, it falls under climatic zone E, with a heating season from 15 October to 15 April and 2404 degree days. The building envelope has been upgraded to comply with contemporary efficiency standards, with a calculated design heating load $Q_{H,des}$ of ~12.9 kW (approx. 44 W/m^2) according to its last Energy Performance Certificate (EPC). In the renovated sections, the transmittance of opaque envelope elements reaches values as low as 0.18 $W/(m^2K)$, reflecting a high level of thermal insulation. The net useful floor area is 291.5 m^2, spread across three levels (underground 105.8 m², ground floor 95.9 m^2, first floor 89.8 m^2). The heated gross volume is 1087 m^3, with a heat losing area/heated volume $S/V = 0.61\ m^{-1}$. In contrast to typical residential datasets that consider the building as a black box, this facility operates as a documented micro energy system with detailed information on both the building envelope and technical systems. A swimming pool and PV array complete the site (Fig. 2).





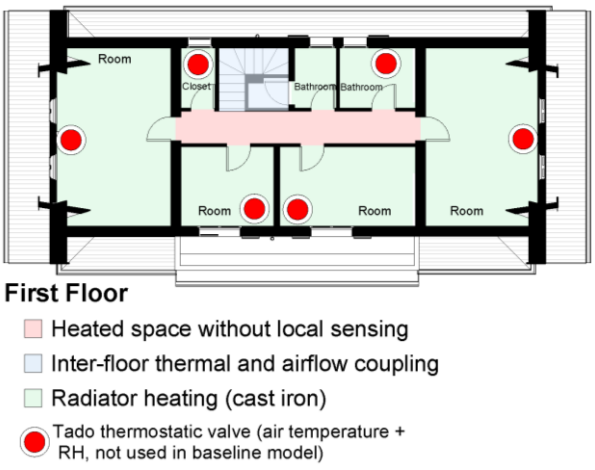


Figure 1. Villa floor plans

The outdoor swimming pool (approximately 50 m³) has a significant electrical load due to circulation and filtration. Ground-floor heating is provided by a 7-kW air-to-water heat pump with low-temperature hydronic underfloor heating, while the underground and first floors use a 23-kW condensing gas boiler with cast-iron radiators. The building also includes smart thermostatic control and a Zehnder ComfoAir Q heat-recovery ventilation unit.

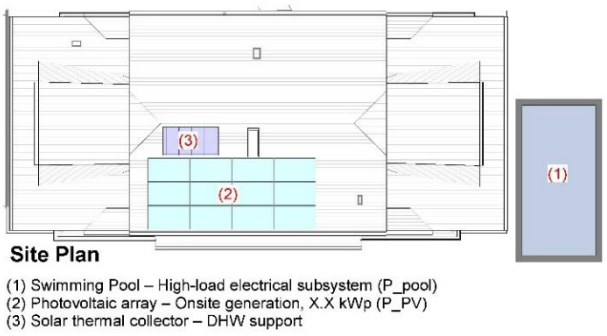


Figure 2. Villa site plan

Instrumentation consists of a heterogeneous IoT network including Shelly Pro 3EM energy meters, smart thermostats, and MVHR telemetry. The operational boundary is defined by four telemetry streams accessed through the *Home Assistant* platform: whole-villa electrical power *P_villa*, PV electrical power *P_PV*, pool circuit electrical power *P_pool*, and outdoor air temperature *T_outdoor*. Streams were retrieved as historical exports and processed in Python over a one-year period from 18 January 2025 to 18 January 2026, resampled to $\Delta t = 60$ minutes. Telemetry includes discontinuities and missing periods consistent with in-use residential operation, motivating explicit missing-data handling in the analytics pipeline.

Operation-phase digital twin architecture

Figure 3 shows the implemented architecture, in which Shelly and Zehnder sensors are integrated via the Home Assistant gateway and processed through a Python analytics pipeline. Sensor-to-subsystem association is maintained through a curated metadata configuration table that maps sensor identifiers to physical subsystems. The implemented architecture links the villa's IoT telemetry streams to a curated subsystem mapping within the analytics layer, enabling traceable association between sensor streams and operational building subsystems. This architecture supports traceable monitoring and baseline construction only; it does not implement bidirectional control, online actuation, or formal BIM-linked semantic integration.

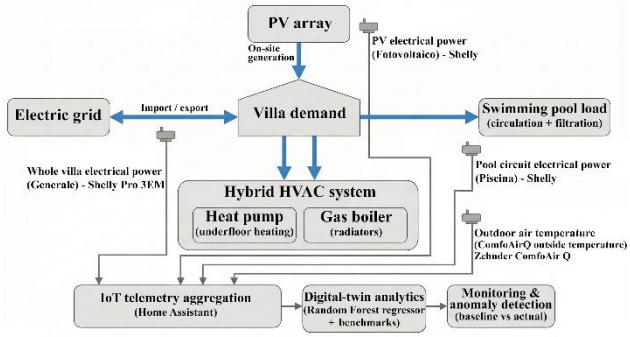


Figure 3. Energy system architecture and data pipeline integrating onsite generation, high-load subsystems, and hybrid HVAC operation with an IoT aggregation layer (Home Assistant) and a baseline-analytics module.

The data processing pipeline aligns asynchronous PV and whole-villa measurements to the nearest hourly timestamp, constructs datasets in multiple configurations based on feature availability (Table 1), imputes data using the noted missing-data policy, and generates engineered features to capture daily and seasonal patterns related to occupancy and solar exposure. For demand-only configurations, the feature set comprises the current demand reading, 1h, 24h, and 168h lags of demand, cyclic hour encodings (sine and cosine), day of week, month, weekend and public holiday indicators, a grid-export flag, and an imputation flag for the demand stream. For telemetry configurations, the feature set additionally includes the current readings and 1h, 24h, and 168h lags of P_PV, P_pool, and T_outdoor, together with per-stream imputation flags, yielding a feature vector of 28 variables for keep-imputed telemetry configurations and a reduced set for clean-only configurations where imputation flags are excluded.

The two missing data policies are as follows. For the first Ax series (causal imputation): missing values are imputed using only information available up to time t (bounded forward-fill plus signal-specific fallbacks):

$$\tilde{x}_t = x_t \quad (1)$$

if observed; otherwise

$$\tilde{x}_t = \tilde{x}_{t-1} \quad (2)$$

when the gap length $\leq H$, else

$$\tilde{x}_t = g(x,signal) \quad (3)$$

For example, PV approaches zero at night and the temperature was set based on an existing climate database. Signal-specific forward-fill bounds H were applied to each stream based on its physical rate of change. For P_villa, P_PV, and P_pool, H was set to 6 hours, reflecting the operational volatility of electrical signals where longer imputation would introduce implausible values. For T_outdoor, H was set to 24 hours, where temperature varies slowly enough that a longer fill window remains physically reasonable. Gaps exceeding these bounds fall back to the signal-specific default g(x, signal) defined in Eq. 3.

For the second Bx series (offline interpolation upper bound): for gaps up to 2 h, missing points are filled by linear interpolation between the nearest available samples x_a and x_b:

$$\tilde{x}_t = x_a + \frac{t-a}{b-a}(x_b - x_a) \quad (4)$$

with

$$b - a \leq 2 \text{ h} \quad (5)$$

Gap-filling and interpolation are applied separately within the training and test partitions to prevent test-period information from influencing the training data. In addition, both keep-imputed and clean-only data selection policies are considered to distinguish realistic deployment conditions from an upper performance bound.

Because interpolation uses both the past and the future endpoint within a gap, Bx represents an offline upper bound; Ax uses past-only information and therefore better approximates conditions where only information available up to time t is used (no future data).

Table 1: Data Treatments

Dataset	Imputation Pipeline	Missing Data Policy	Feature Set
A1	Pipeline A: causal-fill	Keep-imputed	Demand-only
A2	Pipeline A: causal-fill	Clean-only	Demand-only
A3	Pipeline A: causal-fill	Keep-imputed	Demand + telemetry
A4	Pipeline A: causal-fill	Clean-only	Demand + telemetry
B1	Pipeline B: time-based interpolation	Keep-imputed	Demand-only
B2	Pipeline B: time-based interpolation	Clean-only	Demand-only
B3	Pipeline B: time-based interpolation	Keep-imputed	Demand + telemetry
B4	Pipeline B: time-based interpolation	Clean-only	Demand + telemetry

The analytics layer accesses the processed feature vectors and constructs short-horizon operational energy baselines using a fixed Random Forest regressor (Krishnan et al., 2024).

Energy baseline prediction strategy

The analytics function seeks to develop a dynamic energy baseline model that estimates what the building's total consumption should be under current conditions based on historical behaviour. This is framed as a short-horizon baseline construction task (Krishnan et al., 2024). Let y_t denote the total electrical demand at time t. The baseline function f predicts the demand one hour ahead as:

$$\hat{y}_{t+1} = f(X_t) \quad (6)$$

Following the pipeline design, eight dataset configurations are constructed by combining feature availability (demand-only vs. demand augmented with PV, pool, and temperature telemetry), missing-data policy (keep-imputed vs. clean-only), and imputation strategy (Pipeline A: causal fill vs. Pipeline B: time-based interpolation). Imputation strategies are critical to ensure that missing data points or sensor errors do not introduce gaps that could adversely affect the predictive performance of the digital twin (Xie et al., 2023).

A data quality assessment of A1, the causal-fill, keep-imputed, demand-only dataset, indicates substantial operational missingness (e.g., demand missing/imputed 15.9%, PV missing 27.2%, and on-site temperature sensor missing 56.3%), leaving 38.4% fully clean rows; this motivates the explicit keep-imputed vs. clean-only policies and the imputation ablations.

These missingness rates should be understood as a property of an in-use residential monitoring environment rather than as a laboratory-grade sensing setup. In this case study, discontinuities are consistent with intermittent subsystem activity, telemetry or gateway interruptions, sensor outages, and periods in which parts of the monitoring pathway were unavailable. For the outdoor temperature stream specifically, the on-site sensor was

unreliable throughout the monitoring period, and the pipeline substituted values from a meteorological database where available; the reported 56.3% missingness rate reflects the combined proportion of rows where neither source provided a valid observation. For the PV stream, missingness appears as scattered individual hours consistent with intermittent gateway synchronisation rather than extended outage periods. This level of missingness is unusually high for a continuously monitored residential system, but it provides a challenging test of deployment-oriented baseline construction. From a deployment perspective, this means that a practically useful baseline cannot assume fully clean streams and must remain interpretable under missing and reconstructed observations. For this reason, missing-data handling is treated here as a first-order design variable rather than as a secondary preprocessing detail.

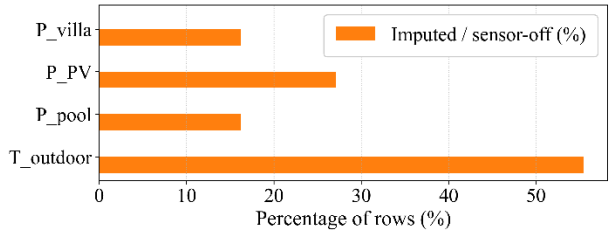


Figure 4. Missingness proxy by signal for dataset A3, expressed as the percentage of rows flagged as Imputed or sensor-off.

Pipeline A constitutes the primary deployable dataset family used for reporting results, while Pipeline B is included only as an offline upper-bound diagnostic that quantifies the performance cost of enforcing causal imputation under otherwise matched conditions; it is not presented as an independent contribution and its near-identical performance relative to Pipeline A is an expected and interpretable outcome rather than a redundancy. Data are split chronologically into 70% training and 30% testing to preserve temporal order and avoid look-ahead bias. To evaluate performance, we compare the Random Forest regressor against two benchmark models: (i) a naïve persistence model, which assumes that next-hour demand equals the current hour's demand (Eq. 7), and (ii) a Multiple Linear Regression (MLR) model trained on the same feature set to test whether simple linear relations are sufficient for short-horizon baselining.

$$\hat{y}_{t+1} = y_t \tag{7}$$

Together, these establish a minimum performance floor that must be exceeded to demonstrate predictive value (Zhou et al., 2024). A Random Forest regressor is used as a fixed learner to capture non-linear interactions between weather, prosumer subsystems, and temporal patterns (Krishnan et al., 2024). We keep the model class fixed (Random Forest) across all experiments to isolate dataset effects. Hyperparameters were tuned on training data only using *RandomizedSearchCV* with *30 iterations* and negative RMSE as the scoring metric, applied via *TimeSeriesSplit* with four folds to avoid look-ahead bias. The search space *covered n_estimators* in the range *300–1100*, *max_depth {None, 10, 20, 40, 60}*, *min_samples_split 2–30*, and *min_samples_leaf 1–20*. The resulting best hyperparameters were frozen per dataset configuration, and the test set was used only once for final evaluation.

Algorithm selection and validation

The Random Forest regressor is adopted as the primary learner because it can capture non-linear relationships in medium-sized tabular time-series data and remains robust with one year of operational observations. The learner is kept fixed across dataset variants to isolate the effect of dataset construction rather than algorithm choice. Feature importances are reported using impurity-based importance, defined as the normalized total reduction in split criterion across trees as implemented in scikit-learn (Pedregosa et al., 2011); these values should not be interpreted as equivalent to Pearson correlation. Performance is evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Tree-based ensembles are widely reported as strong baselines for tabular prediction under irregular target behaviour (Grinsztajn, Oyallon & Varoquaux, 2022).

Results

The Random Forest regressor was benchmarked against both persistence and MLR on the chronological test set, achieving an 8.4% RMSE reduction versus persistence (629.26 W → 576.50 W) and a 5.9% reduction versus MLR (612.0 W → 576.5 W) on A3. Feature importance analysis revealed that sub-metered PV and pool data were the most influential predictors. These findings demonstrate the practical utility of the monitored villa testbed and the proposed pipeline for short-horizon operational baselining in this residential prosumer case study.

Direct RMSE comparison across residential forecasting studies should be interpreted cautiously because buildings, climates, target ranges, horizons, and feature availability differ substantially; for this reason, the present study is positioned as a deployment-oriented case study rather than as a universal benchmark. The full one-year time series of all four input signals alongside whole-building demand is not shown here due to space constraints; signal quality, missingness patterns, and the operational range of each stream are characterised in the Methodology section and summarised in Fig. 4.

Fig. 5 presents the Pearson correlation heatmaps of the hourly processed datasets under causal fill, time-based interpolation, and clean-only selection, focusing on the relationships between subsystem telemetry and whole-villa electricity demand.

PV generation shows the strongest negative linear association with whole-building demand, which is expected because it reflects onsite generation rather than demand-side consumption. Outdoor temperature also shows a negative association consistent with seasonal heating behaviour. Pool-circuit power exhibits a weaker linear correlation, but this does not imply low predictive value because pool operation is intermittent and state-like rather than smoothly linear.

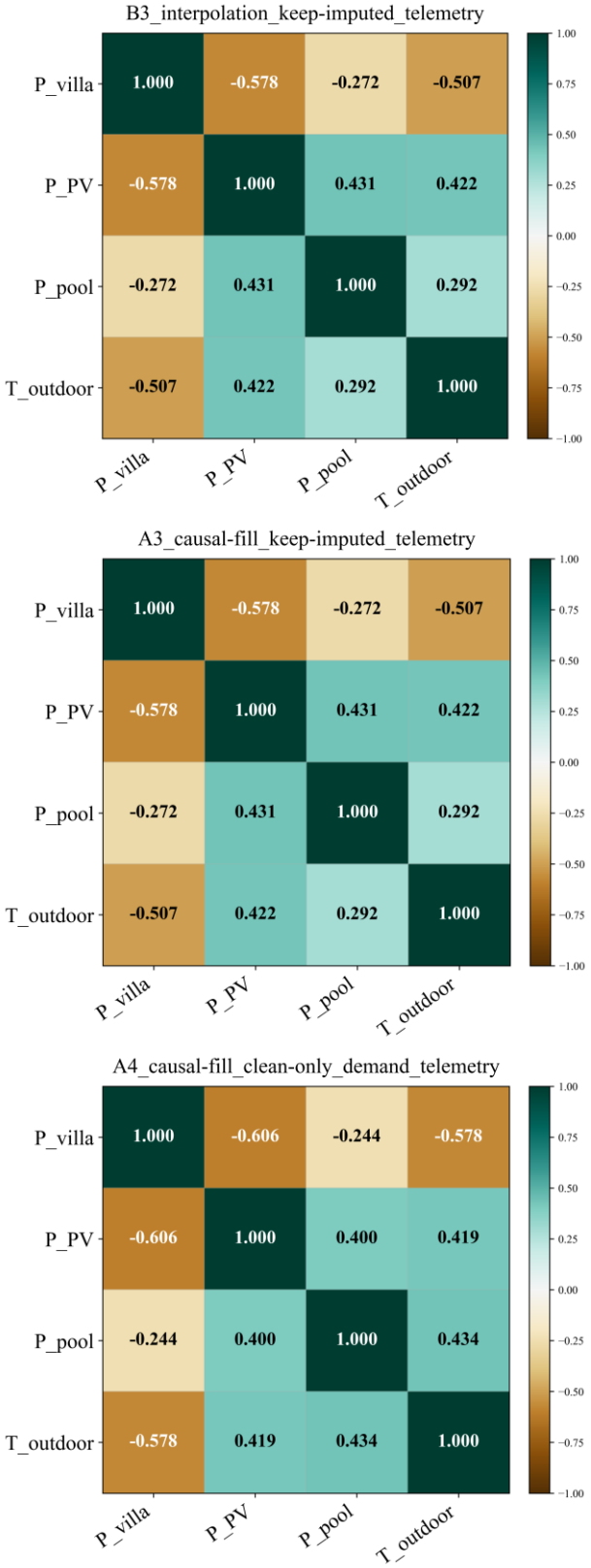


Figure 5. Pearson correlation heatmaps for selected processed datasets: B3 (top), A3 (middle), and A4 (bottom).

For this reason, feature-importance results and Pearson correlation should be interpreted as complementary rather than equivalent diagnostics. The near-identical correlation matrices for A3 and B3 are expected: in both configurations missing values are retained through imputation rather than removed, so the underlying observed data structure is identical and only the imputed segments differ; since these segments are small relative to the full dataset and consistent with the surrounding trend, the global linear association structure remains effectively unchanged. In contrast, the A4 and B4 configurations show noticeably different correlation values because clean-only filtering removes imputed rows and thereby alters the sample composition and temporal coverage. The maintained metadata mapping is used only to preserve subsystem traceability within the analytics layer. Correlation patterns are broadly similar across the compared datasets, while any differences introduced by clean-only filtering should be interpreted cautiously because dataset coverage and evaluated timestamps change across configurations.

Operational baseline model performance

Table 2 summarises model performance on the chronological test set (30% of the available data; 2529 samples for keep-imputed datasets) and reports Random Forest RMSE alongside those from the persistence and MLR benchmarks, showing that the datasets using the "Keep-imputed" data policy alongside "Demand + telemetry" features outperformed the others. Note that A4 and B4 have substantially lower usable hours than the other datasets, as reflected in the test-row counts reported in Tables 2 and 3, partially accounting for their decreased performance. This reduction in usable hours is not only a sample-size issue. Clean-only filtering also changes which operating regimes remain in the evaluated subset and therefore changes the task itself. In A4 and B4, the retained rows no longer represent the same full operational variability as the keep-imputed datasets. As a result, the predictive advantage of the richer feature set and the non-linear flexibility of Random Forest can be reduced, while a simpler model such as MLR may become comparatively more competitive on the filtered subset. This regime shift also accounts for the MAE results in Table 3, where RF underperforms MLR in A4 and B4; on a filtered subset with reduced operational variability, the non-linear flexibility of Random Forest provides no advantage over a linear model and the smaller sample size limits effective hyperparameter tuning. The eight configurations test three directional hypotheses embedded in the pipeline design. Keep-imputed configurations should outperform clean-only counterparts by preserving operational variability that filtering removes. Telemetry configurations should outperform demand-only configurations by providing subsystem-level predictive signal that demand history alone cannot capture. Pipeline A and B should converge under keep-imputed conditions because the imputation method has limited influence on the global data structure when imputed values are retained. Tables 2 and 3 confirm all three patterns.

Table 2: Summary of results – RMSE

Dataset	Test rows	RF RMSE (kW)	Persistence RMSE (kW)	MLR RMSE (kW)
A1	2529	0.597	0.629	0.634
A2	2118	0.652	0.681	0.655
A3	2529	**0.576**	0.629	0.612
A4	939	0.831	0.859	0.804
B1	2529	0.596	0.629	0.634
B2	2118	0.652	0.681	0.655
B3	2529	**0.573**	0.629	0.612
B4	939	0.831	0.859	0.804

We additionally report MAE (Table 3) to indicate typical behaviour error, while the RMSE values indicate the model sensitivity to extreme conditions.

Table 3: Summary of results – Mean Absolute Error

Dataset	Test rows	RF MAE (kW)	Persistence MAE (kW)	MLR MAE (kW)
A1	2529	0.366	0.378	0.426
A2	2118	0.431	0.440	0.473
A3	2529	**0.362**	0.378	0.409
A4	939	0.633	0.618	0.620
B1	2529	0.365	0.378	0.426
B2	2118	0.431	0.440	0.473
B3	2529	**0.361**	0.378	0.409
B4	939	0.633	0.618	0.620

The Random Forest regressor trained on A3 achieved MAE = 0.362 kW and RMSE = 0.576 kW, representing the best trade-off between accuracy and deployment realism among the tested configurations, since B3 achieves marginally lower error but relies on non-causal interpolation that is not available at deployment time. The building exhibits a skewed demand distribution with a median of 0.76 kW and a mean of 1.04 kW driven by occasional high-load and PV export events, making the RMSE of 0.576 kW a substantial fraction of typical instantaneous demand. This indicates that the baseline is suited to detecting large residual deviations consistent with equipment faults or sustained anomalous consumption but may not meet the accuracy requirements of fine-grained measurement and verification applications. In this paper, the baseline is intended as an operational reference for residual-based monitoring and measurement-oriented comparison, not as a full physical emulator of building behaviour. Its practical value lies in improving over persistence and MLR under leakage-safe and deployment-realistic conditions.

Adding telemetry (A3 vs A1) reduces RMSE by ~3.5%, while RF improves over persistence and MLR by 8.4% and 5.9%, respectively, quantifying the practical benefit of subsystem instrumentation under deployment-like constraints.

A3 and B3 show nearly identical results (RMSE $\leq$ 0.5%). The cost of enforcing deployable causality is therefore modest in this case study. Across all keep-imputed configurations, the Random Forest regressor outperforms the persistence benchmark, confirming predictive value beyond trivial extrapolation. Feature importance analysis ranks P_pool (32%), recent demand history (28%), P_PV (19%), and T_outdoor (12%) as the dominant explanatory variables, with temporal features making up the balance across all keep-imputed configurations. These rankings are consistent across all keep-imputed configurations; demand-only configurations assign full importance to demand history and temporal features by construction. The prominence of pool-circuit telemetry in impurity-based feature importance, despite its weaker Pearson correlation with demand, is plausible in this setting because the pool signal acts as a regime marker for intermittent switching behaviour and interacts non-linearly with other predictors. In other words, a feature can contribute strongly to tree-based prediction even when its standalone linear association with the target is modest. It should also be noted that impurity-based importance is known to inflate scores for intermittent or high-variance signals because such features create frequent and uneven splits across trees; the reported importance values should therefore be interpreted as indicative of predictive relevance within this specific Random Forest configuration rather than as absolute measures of physical influence.

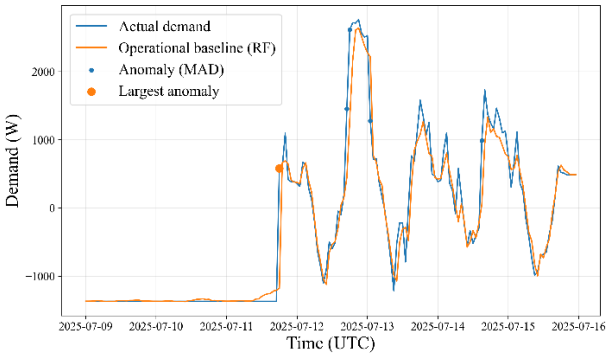


Figure 6. Sample week showing actual demand vs operational baseline (RF with A3 data configuration)

Fig. 6 illustrates a representative week of measured demand, and the one-hour-ahead operational baseline produced by the best-performing A3 configuration using Random Forest. The baseline reproduces daily structure and prosumer-specific effects such as PV self-consumption periods and pool-pump cycling. A prominent deviation is consistent with a documented circulation-pump fault due to voltage instability (resolved by installing a stabiliser in July 2025); such episodes are detected from the baseline residual using the MAD-based threshold defined in Eq. 9 and are flagged for separate post-hoc analysis. They are not removed from the test set used for RMSE and MAE

computation, and the current study does not implement an online retraining loop in which flagged anomalies are used to update the fitted baseline.

Anomalies are detected from the baseline residual rather than from raw demand. Let

$$r_{t+1} = y_{t+1} - \hat{y}_{t+1} \quad (8)$$

where y_{t+1} is the measured demand and $\hat{y}_{t+1}$ is the Random Forest baseline prediction. We compute the median residual $m = \text{median}(r)$ and $MAD = \text{median}(|\, r - m \,|)$, and flag time step $t + 1$ as anomalous if

$$|\, r_{t+1} - m \,| > K \cdot MAD \quad (9)$$

In this study we use $K = 9$, selected via a threshold sweep to reduce spurious detections during stable operation.

Conclusions

This study demonstrates that deployment-realistic missing-data handling and sub-metered subsystem telemetry are first-order determinants of operational baseline quality in residential prosumer buildings. The monitored villa testbed and the eight-configuration ablation show that imputation strategy and data policy shape predictive performance as much as feature selection, and that the cost of enforcing causal deployability is modest when imputed values are retained.

Key findings

Sub-metered subsystem telemetry, especially PV and pool-circuit power, improves predictive baselines when used within a controlled and leakage-safe dataset design. Baseline quality is strongly shaped by feature availability and missing-data policy rather than algorithm choice alone. Under keep-imputed causal conditions, A3 achieved the lowest deployable RMSE (0.576 kW). The A3 baseline residuals successfully flagged the documented circulation-pump fault in July 2025, confirming practical anomaly screening capability under in-use conditions.

Limitations and future work

Occupancy is not directly measured, so the current feature set relies on temporal proxies and may not distinguish between unoccupied and low-activity conditions. The present study is limited to monitoring and baseline construction on a single residential testbed. Accordingly, the current implementation remains an analytics-oriented operational representation rather than a closed-loop control system or a formally BIM-integrated workflow. Future work should validate the approach across multiple residential buildings, enrich behavioural and occupancy context, and extend the current offline pipeline toward continuous deployment, supervisory diagnostics, and optional richer semantic alignment were justified by the monitoring objective.

References

Balaji, B., Bhattacharya, A., Fierro, G., Gao, J., Gluck, J., Hong, D., Johansen, A., Koh, J., Ploennigs, J., Agarwal, Y., Bergés, M., Culler, D., Gupta, R.K., Kjærgaard, M.B., Srivastava, M., & Whitehouse, K. (2018) Brick : Metadata schema for portable smart building applications, Applied Energy, 226, pp. 1273–1292.

Ghaemi, A., Rezgui, Y., Petri, I., Beach, T., & Ghoroghi, A. (2024) AI and Digital Twin Applications in Building Energy Management: A State-of-the-Art Review, IEEE Access, 12, pp.1-11.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022) Why do tree-based models still outperform deep learning on typical tabular data? in Advances in Neural Information Processing Systems.

Krishnan, S., Divakaran, A., Rethnam, O.R., Markarian, E., Qiblawi, S., Thomas, A. & Azar, E. (2024) Evaluating the capabilities of surrogate modeling techniques in predicting hourly building energy consumption, in EC3 Conference 2024, 5, European Conference on Computing in Construction.

Kychkin, A., & Chasparis, G.C. (2025) Hourly Short Term Load Forecasting for Residential Buildings and Energy Communities. arXiv preprint arXiv:2501.19234.

Mengual Torres, S. G., Tzuc, O. M., Aguilar-Castro, K. M., Tellez, M. C., Sierra, J. O., Cruz, A. D. R. C. Y., & Barrera-Lao, F. J (2022) Analysis of Energy and Environmental Indicators for Sustainable Operation of Mexican Hotels in Tropical Climate Aided by Artificial Intelligence, Buildings, 12(8), p. 1155.

Oulefki, A., Kheddar, H., Amira, A., Kurugollu, F., Himeur, Y., & Bounceur, A., (2025) Innovative AI strategies for enhancing smart building operations through digital twins: A survey, Energy and Buildings, 335, pp. 115567–115567.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp. 2825–2830.

Popławski, T., Dudzik, S., & Szeląg, P. (2023) Forecasting of Energy Balance in Prosumer Micro-Installations Using Machine Learning Models. Energies, 16(18), 6726.

Xie, X., Merino, J., Moretti, N., Pauwels, P., Chang, J. Y., & Parlikad, A. (2023) Digital twin enabled fault detection and diagnosis process for building HVAC systems, Automation in Construction, 146, p. 104695.

Zhou, F., Yang, C. & Wang, Z. (2024) Prediction of building energy consumption for public structures utilizing BIM-DB and RF-LSTM, Energy Reports, 12, pp. 4631–4640.