Hybrid XGBoost-PINN Multi-Stage Framework for Scalable Building Temperature Prediction

Tushar Shinde, Rohan Saha
MIDAS (Multimedia Intelligence, Data Analysis and compreSsion) Lab
Indian Institute of Technology Madras, Zanzibar, Tanzania
shinde@iitmz.ac.in

Abstract

Building temperature prediction is crucial for energy optimization and control in smart cities. We present a hybrid framework combining XGBoost with physics-informed neural networks (PINN) in a multi-stage sequential scaling approach. Starting from single-zone, single-day predictions, we progressively scale to multi-zone, multi-year forecasts using real-world data from Google's Smart Building Simulator. Our method incorporates physics-enhanced features, temporal encodings, and inter-zone interactions, achieving mean absolute errors (MAE) as low as 0.169°F for weekly multi-zone predictions. For longer horizons, we employ ensemble strategies, demonstrating robust performance up to 2.5 years. This work advances urban AI by enabling accurate long-term building dynamics modeling for downstream control tasks. Link to the code: https://colab.research.google.com/drive/1ul_Qzwq-awYVQYI7yLk__jW93oZE5aIt?usp=sharing

1 Introduction and Related Work

Urban buildings consume over 70% of city energy, making accurate temperature prediction crucial for energy-efficient control and planning Wang and Ma [2008]. Forecasting building thermal dynamics is challenging due to complex inter-zone interactions and temporal variations. Our work leverages real-world datasets to bridge short-term forecasting and long-term planning, addressing spatial dependencies, seasonal effects, and temporal drifts Saha and Shinde. Building energy modeling has evolved from detailed physics-based simulations, such as EnergyPlus Crawley et al. [2001], to machine learning approaches that capture temporal patterns from sensor data using LSTMs and Transformers Reza et al. [2022]. While deep learning excels in short-term predictions, purely data-driven models can struggle with long-term extrapolation.

Physics-informed neural networks (PINNs) present a promising alternative by embedding governing physical laws, such as energy conservation or Newton's law of cooling, directly into the learning objective Goldfeder et al. [2024]. By constraining predictions to adhere to known physical principles, PINNs improve generalization and robustness, particularly in scenarios with limited or noisy data. Complementing these, hybrid approaches that combine tree-based ensemble methods like XGBoost with physics-informed features have shown robustness and interpretability in time-series prediction tasks Sahin [2020]. XGBoost effectively captures non-linear dependencies while physics-derived features constrain the solution space to physically plausible regions.

Building on these advancements, we introduce a multi-stage scaling framework that progressively extends prediction horizons from short-term, single-zone scenarios to ultra-long-term, multi-zone forecasts. Unlike prior studies that focus on either isolated short-term predictions or purely physics-driven models, our approach systematically mitigates cumulative errors and seasonal drifts through sequential validation, physics-enhanced features, and horizon-specific ensembles. Specifically, our

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

framework integrates building metadata (e.g., floorplans, device layouts), physics-informed lag features, and adjacency-based spatial encodings to enhance predictive fidelity.

The proposed hybrid XGBoost-PINN framework combines the computational efficiency and interpretability of XGBoost with the physical consistency enforced by PINNs Goldfeder et al. [2024], while ensuring scalability via sequential stage-wise modeling. Our contributions are threefold:

- A sequential scaling methodology that extends prediction horizons from one day to 2.5 years, enabling comprehensive temporal coverage and rigorous validation at each stage.
- Development of physics-enhanced features, including adjacency matrices for inter-zone interactions and cyclical temporal encodings to capture diurnal and seasonal patterns.
- Horizon-specific ensemble strategies for ultra-long-term predictions, which combine specialized models to achieve state-of-the-art performance.

2 Method

We propose a framework of a multi-stage sequential scaling strategy that methodically increases both the temporal horizon and spatial complexity of predictions, ensuring that each stage builds upon the validated foundations of the previous ones to achieve robust and scalable performance. This progressive approach mitigates the risks of abrupt scaling, such as overfitting or instability, by allowing incremental incorporation of complexities like inter-zone dependencies and long-term patterns, while providing clear checkpoints for evaluation and refinement.

2.1 Data Preparation

We utilize the Smart Building dataset from Goldfeder et al. [2024], which comprises comprehensive time-series matrices including observations for sensor readings, actions for control inputs, and reward information for performance metrics. From this, temperature targets are precisely extracted from zone air temperature sensors, serving as the primary prediction variables, whereas exogenous features encompass a range of inputs such as weather conditions and operational setpoints, providing contextual signals that influence thermal dynamics. The training data is drawn from the 2022_a split, covering January to June with 51,852 timesteps, offering a diverse representation of seasonal transitions (see Figure 3 in Appendix). Validation is performed on the 2022_b split, spanning July to December with 53,292 timesteps, to assess generalization to unseen periods. Additionally, physics metadata, including device information dictionaries for sensor specifications, zone information for spatial attributes, floorplan arrays for geometric layouts, and device layout maps for connectivity, are integrated to enrich the feature space with domain-specific knowledge, enabling more physically grounded modeling.

2.2 Physics-Enhanced Feature Engineering

To incorporate physical realism into the predictive framework, we construct a set of features that explicitly capture both spatial and temporal dynamics of building thermodynamics Gokhale et al. [2022], Chen and Guestrin [2016]. Let $\mathbf{A} \in \mathbb{R}^{Z \times Z}$ denote the adjacency matrix derived from floorplan metadata, where Z is the number of zones. Each entry A_{ij} encodes a proximity-based weight between zones i and j, decaying with Euclidean distance, to represent heat conduction and convection pathways:

$$A_{ij} = \exp(-\alpha\, d_{ij}), \tag{1}$$

where d_{ij} is the inter-zone distance and α is a decay coefficient. This captures the principle that thermal energy transfers more readily between neighboring zones, enabling realistic temperature propagation across connected spaces. Temporal dynamics are represented using cyclical embeddings for hour-of-day, day-of-week, and day-of-year:

$$\sin_t = \sin\left(2\pi \frac{t}{T}\right), \quad \cos_t = \cos\left(2\pi \frac{t}{T}\right), \tag{2}$$

where T is the period (24 hours, 7 days, or 365 days), mapping periodic phenomena into a continuous space and enabling gradient-based learners to capture recurring patterns Elloumi et al. [2025].

Physics-informed lag features encode thermal inertia. For each zone z, lagged temperature values at intervals $\tau \in \{1, 3, 6\}$ hours are appended: $x_z(t-\tau), \quad \forall\tau$, while inter-zone differences are computed as: $\Delta T_{ij}(t) = T_i(t) - T_j(t), \quad \forall(i,j)$ s.t. $A_{ij} > 0$, capturing thermal gradients that drive diffusive heat flow Incropera et al. [1990].

All features are aligned across training and validation sets, with missing values imputed via the median and dimensional mismatches resolved by truncation, ensuring computationally tractable and physically meaningful inputs.

2.3 Multi-Stage Scaling

The multi-stage framework incrementally scales both spatial and temporal horizons. **Stage 1:** Single zone, one-day prediction using conservative XGBoost settings (shallow depth, strong regularization) establishes a reliable baseline:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |T_{\text{pred}}^{(i)} - T_{\text{true}}^{(i)}|. \tag{3}$$

Stage 2: All zones (123) for one day, trained with inter-zone features to capture heterogeneous behaviors and spatial interactions. **Stage 3:** One-week prediction for a single zone, including weekly temporal features (e.g., business hours) and extended lags to capture accumulated thermal effects. **Stage 4:** Multi-zone weekly prediction combining spatial and temporal components, demonstrating improved MAE (Table 1). **Stage 5:** Two-week predictions with bi-weekly features and extended lags to handle transitional patterns. **Stage 6:** One-year predictions via seasonal decomposition with heating/cooling indicators Crawley et al. [2001] to address non-stationary weather effects. **Stage 7:** Ultra-long-term (2.5 years) predictions incorporating multi-year aging features to model gradual system degradation (e.g., insulation wear) Wang et al. [2020].

Across all stages, XGBoost hyperparameters are tuned iteratively, with early stopping for generalization. PINNs augment longer stages, optimizing a combined loss:

$$\mathcal{L} = \|T_{\text{pred}} - T_{\text{true}}\|_2^2 + \lambda\|\mathcal{F}(T_{\text{pred}})\|_2^2, \tag{4}$$

where $\mathcal{F}$ represents the discretized heat transfer operator and λ balances data fidelity with physical consistency Gokhale et al. [2022].

3 Experiments and Results

The framework is implemented in Python, utilizing libraries such as XGBoost for core modeling, scikit-learn for preprocessing and metrics, and NumPy for efficient array operations, ensuring reproducibility through fixed random seeds and version-controlled dependencies. Experiments are conducted in a Kaggle/Colab environment with standard CPU/GPU resources, simulating accessible computational settings for broader applicability. Training times are meticulously recorded for each stage to assess scalability, while a simple mean prediction baseline, computed from training temperatures, is used throughout for relative performance benchmarking, highlighting the framework's added value over trivial approaches.

Evaluation Metrics. The primary metric employed is the Mean Absolute Error (MAE) for temperature predictions, selected for its direct interpretability in degrees Fahrenheit and sensitivity to prediction deviations that impact control decisions. Secondary metrics include Root Mean Squared Error (RMSE) to emphasize larger errors that could signify model instability, and R^2 to quantify explained variance, providing insight into how well the framework captures underlying dynamics relative to a naive mean baseline. These metrics are computed per stage, aggregated across zones for multi-zone evaluations, and reported with distributions to highlight consistency and outliers.

Results. We evaluate the proposed multi-stage XGBoost framework across increasing temporal horizons and spatial complexities. Table 1 summarizes the mean absolute error (MAE) across all stages, illustrating a clear progression: Stage 1 (single-zone, one-day) yields an MAE of 0.424°F, which decreases to 0.101°F at Stage 5 (two-week, multi-zone) before increasing for longer horizons due to accumulating uncertainties in extended forecasts, highlighting the effectiveness of the sequential hybrid ML-physics approach in capturing weekly dynamics.

Table 1: Sequential scaling performance of the proposed framework, reported as mean absolute error (MAE) in °F across different stages.

Stage	Description	Zones	MAE (°F)
1	Single Zone, 1 Day	1	0.424
2	All Zones, 1 Day	123	0.325
3	Single Zone, 1 Week	1	0.173
4	All Zones, 1 Week	123	0.169
5	All Zones, 2 Weeks	123	0.101
6	All Zones, 1 Year	123	2.080
7	All Zones, 2.5 Years	123	2.826

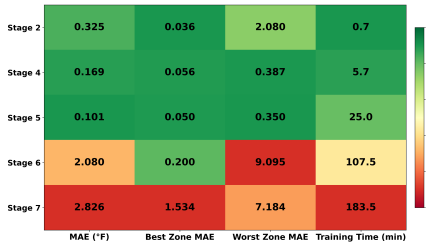
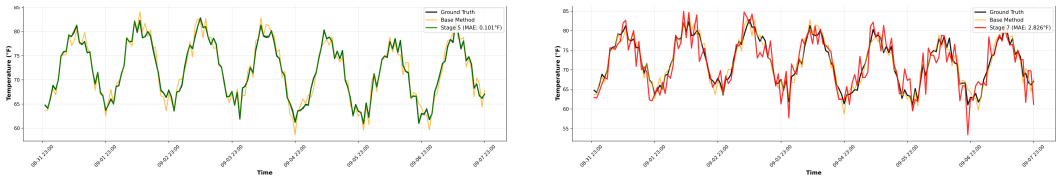


Figure 1: Sequential scaling analysis of the multi-stage framework.



(a) (b)

Figure 2: Prediction performance across stages: (a) Stage 5 shows high accuracy and minimal deviation from ground truth, representing optimal performance of the framework. (b) Stage 7 exhibits significant prediction errors and instability over extended temporal horizons, demonstrating limitations in long-term forecasting.

Figure 1 provides a performance overview across stages, demonstrating the framework's ability to maintain excellent scores for early and mid-term horizons while gradually declining for ultra-long-term predictions. The MAE trends over prediction horizons for all zones are shown Figure **??** (in Appendix). Short-term stages (Stages 2–4) show consistent low errors, with 78.9% of zones in Stage 4 below 0.2°F, while Stages 6–7 exhibit larger deviations, peaking at 2.826°F in Stage 7. These results indicate that the ensemble strategies mitigate error accumulation to some extent, but the inherent limitations of the dataset and long-term dependencies remain. **Computational Time Analysis.** Training complexity scales with horizon length, from 0.16 seconds in Stage 1 to 10,857 seconds in Stage 7 (see Figure 1), reflecting increased data volume and model complexity. Nevertheless, the offline training remains practical for realistic deployment scenarios.

Discussion. The short-term prediction stages (1–4) provide sub-0.2°F MAE, suitable for real-time building control and proactive HVAC adjustments (Figure **??** in Appendix). Longer horizons (Figure **??** in Appendix) exhibit elevated errors due to compounding uncertainties, such as unmodeled occupant behavior, equipment drift, and external weather variations. Despite these limitations, predictions remain informative for strategic planning and maintenance scheduling. Incorporating physics-informed features, including inter-zone interactions through adjacency matrices and temperature gradients, significantly improves generalization and enforces physical plausibility, reducing dependence on data-driven learning alone. The sequential design enables systematic analysis and debugging of errors across stages, and the hybrid framework offers a pathway for integrating ML and physics-based modeling for urban AI applications.

4 Conclusion

In this work, we have developed a scalable hybrid XGBoost-PINN framework for building temperature prediction that demonstrates robust performance across a wide range of temporal and spatial scales, from single-day single-zone forecasts to multi-year multi-zone modeling. By integrating physics-informed features, ensemble strategies, and a multi-stage scaling approach, our method advances the state of urban AI, providing accurate and physically consistent predictions that enable improved building control, energy efficiency, and sustainability in smart cities. Future work should address dataset limitations, particularly the 2.5-year coverage, to better validate multi-year performance. Extensions could include full PINN loss functions to embed heat dynamics equations and additional datasets to capture long-term drift and aging effects.

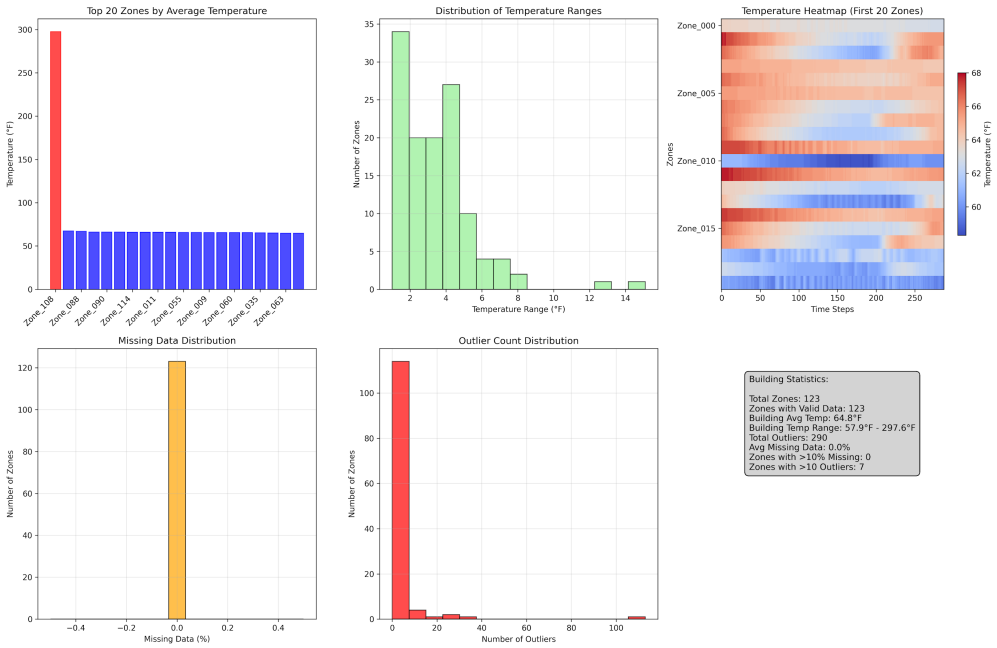


Figure 3: Building summary for January 1, 2022, showing the top 20 zones with average temperature, distribution, outliers, and a heatmap for each timestamp. This visualization provides insights into spatial and temporal temperature patterns, highlighting areas of potential concern or irregular behavior across the building.

A Zone-Level Sensor Analysis

To evaluate the reliability and interpretability of building sensor data, we conducted a detailed analysis of two representative zones: Zone 102 and Zone 108 from the `2022_a` dataset. Each zone was examined at a daily resolution (288 timesteps for January 1, 2022), and all available sensors were categorized, visualized, and summarized. Comprehensive sensor-level plots and CSV summaries were generated for reproducibility.

Zone 102 Analysis. Zone 102 contained a diverse set of 9 sensors spanning multiple categories, including temperature setpoints, flow sensors, valve and damper commands, and one zone air temperature sensor. The sensor breakdown was: Other (1), Valve Command (1), Temperature Setpoint (3), Temperature (1), Flow Sensor (2), Damper Command (1). All sensors were successfully analyzed, producing 9 diagnostic plots and a consolidated summary file. Our findings reveal that more than half of the sensors exhibited binary or status-type behavior (e.g., valve and damper commands). Three sensors, namely *Valve Command*, *Temperature Setpoint*, and *Supply Air Flowrate*, demonstrated frequent state changes (3 events each), identifying them as the most active signals driving thermal regulation dynamics. Meanwhile, two sensors exhibited high variability across the day, suggesting sensitivity to operational or environmental changes. This diversity highlights Zone 102 as a control-intensive environment, where both continuous and discrete sensor modalities interact to regulate comfort and efficiency.

Zone 108 Analysis. Zone 108 was comparatively simpler, with 5 active sensors identified: Other (3), Temperature Setpoint (1), and Temperature (1). Unlike Zone 102, most sensors here showed limited variability, with only the *Temperature Setpoint* changing once during the observation period. Both the zone air temperature and supporting air temperature sensors remained stable, indicating relatively static conditions and lower operational complexity compared to Zone 102.

Comparative Analysis. The contrast between Zones 102 and 108 illustrates the heterogeneous nature of building subsystems. Zone 102 exhibits frequent actuation and strong coupling between control signals (valve, damper, flowrate) and thermal states, requiring careful modeling of interdependencies. In contrast, Zone 108 reflects a more stable thermal environment dominated by passive monitoring with minimal active interventions. These findings emphasize the importance of zone-specific modeling

strategies: control-heavy zones demand feature representations that account for actuator-driven variability, while monitoring-dominant zones can benefit from simpler, stability-focused forecasting approaches.

References

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Drury B Crawley, Linda K Lawrie, Frederick C Winkelmann, Walter F Buhl, Y Joe Huang, Curtis O Pedersen, Richard K Strand, Richard J Liesen, Daniel E Fisher, Michael J Witte, et al. Energyplus: creating a new-generation building energy simulation program. *Energy and buildings*, 33(4):319–331, 2001.

Yaroub Elloumi, Salim Khazem, Ibrahim Krayem, and Jeyakaran Mahesananthan. Cyclical temporal encoding for ensemble deep learning in multistep energy forecasting. 2025.

Gargya Gokhale, Bert Claessens, and Chris Develder. Physics informed neural networks for control oriented thermal modeling of buildings. *Applied Energy*, 314:118852, 2022.

Judah Goldfeder, Victoria Dean, Zixin Jiang, Xuezheng Wang, Hod Lipson, John Sipple, et al. The smart buildings control suite: A diverse open source benchmark to evaluate and scale hvac control policies for sustainability. *arXiv preprint arXiv:2410.03756*, 2024.

Frank P Incropera, David P DeWitt, Theodore L Bergman, Adrienne S Lavine, et al. *Fundamentals of heat and mass transfer*, volume 1072. New York John Wiley & Sons, Inc., 1990.

Selim Reza, Marta Campos Ferreira, José JM Machado, and João Manuel RS Tavares. A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. *Expert Systems with Applications*, 202:117275, 2022.

Rohan Saha and Tushar Shinde. Scalable building temperature prediction for smart hvac control: A multi-stage learning framework. In *ICML 2025 CO-BUILD Workshop on Computational Optimization of Buildings*.

Emrehan Kutlug Sahin. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using xgboost, gradient boosting machine, and random forest. *SN Applied Sciences*, 2(7):1308, 2020.

Jian Qi Wang, Yu Du, and Jing Wang. Lstm based long-term energy consumption prediction with periodicity. *energy*, 197:117197, 2020.

Shengwei Wang and Zhenjun Ma. Supervisory and optimal control of building hvac systems: A review. *Hvac&R Research*, 14(1):3–32, 2008.