A Temporal Features-Enhanced Mixture-of-Experts Approach for Indoor Temperature Prediction

Kanxuan He [1] **Hongshan Guo** [1]

Abstract

This study presents a flexible modeling pipeline for indoor temperature prediction that leverages a Mixture-of-Experts (MoE) framework built upon Light Gradient Boosting Machine (LightGBM) models. The approach incorporates a set of temporal feature-enhanced experts using methods such as Moving Average (MA) and Exponentially Weighted Moving Average (EWMA) to embed temporal trends. A model selector is trained to assign dynamic soft weights to each expert at every time step based on contextual features, enabling the final prediction to be a weighted combination of all experts' outputs. The Soft-MoE framework achieved a mean absolute error (MAE) of 1.1839 °F across all rooms over the entire validation period. Notably, during periods with pronounced diurnal temperature fluctuations, the EWMA-enhanced expert reduced MAE by 45.4% compared to the base mode. The proposed MoE framework demonstrates strong adaptability to diverse temporal dynamics and is readily applicable in real-world building environment control systems. The complete Jupyter Notebook is available at: `https://drive.google.com/file/d/1Os6GDuHBo0CpUwvMVGwr5qiuXFoEawaB/view?usp=sharing`.

1. Introduction

Indoor temperature predicting is essential for building environment control, serving as a prerequisite for downstream applications such as energy consumption simulation, Heating, Ventilation, and Air Conditioning (HVAC) control and thermal comfort modeling (Palaić et al., 2023; Sasaki et al., 2024; Xu et al., 2021). However, accurate prediction remains challenging due to the highly coupling interactions between numerous exogenous factors driven by weather, occupants behaviors and HVAC regulations (Jiang et al., 2024; Mtibaa et al., 2020). *The Smart Buildings Competition* at the *ICML 2025 CO-BUILD Workshop* offers an opportunity to deep dive into the field. The task involves leveraging the building sensor data from Google's open-source *Smart Building Simulator framework* (Goldfeder & Sipple, 2023) (hereafter referred to as SBS dataset), which is part of *the Smart Buildings Control Suite* (Goldfeder et al., 2025) — an interactive benchmark framework for building control. The developed models, trained on the provided dataset, are required to predict indoor temperatures over a validation dataset of six month period. In addition to the above mentioned inherent difficulties, the competition poses further unique challenges as follows.

- **Substantial Missing in Datasets:** Total missing data ratio in training dataset is 35.18%, the time period of which is almost overlapped across all sensors, including a over-one-month period (from 2022-04-08 to 2022-05-17). For the validation set, 19.51% of the sensor data is missing with the similar overlapping pattern but more sparsely distributed. The missing pattern across timestamps and devices is shown in App. Figure 6. The data missing in the training set, especially the long term ones, hinders the model to learn and track the temporal pattern. During inference, the data gap causes drift problem and thus more severe cumulative error.
- **Disproportion of Training and Validation Dataset Length**: Both datasets are recorded from a 6 month period (Jan-June 2022 and July-Dec 2022 respectively). This equal split is unusual in Machine Learning (ML), especially in time series forecasting, where the training datasets usually span over a year or at least cover the involved months in the validation set to allow the model to capture the whole temporal pattern. The limited training horizon calls for higher robustness and generality of the model.
- **Room-wise Variability in Sensor Modalities**: The task involves over a hundred sensors with various ob-

[1] Department of Architecture, University of Hong Kong, Hong Kong SAR, China. Correspondence to: Kanxuan He <kanxuan.he@connect.hku.hk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

served variables across a two-floor indoor space. Although not a fully multivariate forecasting, the heterogeneity in input space introduces further challenge in capturing the shared pattern yet with nuanced, room-specific differences for robust prediction.

To address these challenges, this study proposes a multi-stage imputation and temporal context aware Soft Mixture-of-Experts (Soft-MoE) framework, training a set of Gradient Boosting Machine-based models with various methods of additional temporal smoothing features. The proposed MoE approach achieves a mean absolute error (MAE) of 1.1839 °F over the whole span of validation set, with stable performance across rooms (MAE standard deviation within 0.4 °F). The additional temporal smoothing features enhance the model's robustness in response to fluctuation and enable flexible model ensemble based on the forecasting context. The pipeline is well-suited for real-world applications, effectively leveraging weather data and enabling efficient online prediction with lightweight deployment.

2. Methodology

2.1. Modeling Framework Overview

Our approach follows a modular, multi-stage design tailored for the challenges in indoor temperature forecasting with sparse and heterogeneous sensor data. As shown in Figure 1, the high-resolution outdoor temperature signal reconstructed with an external coarse-grained meteorological data serves as an anchor to impute missing data in the training dataset. Based on the completed sequences, a set of Light Gradient Boosting Machine (LightGBM) models (Ke et al., 2017), incorporating temporal smoothing method such as Moving Averages (MA) and Exponentially Weighted Moving Average (EWMA) to enhance the model's ability to capture temporal dynamics. During inference, the same feature processing is applied to the validation data, and a Soft-MoE selector dynamically calculate the weighted predictions from each expert mode for each timestep. This design enables the framework to adapt to varying temporal patterns and improves overall prediction accuracy, which is validated through further evaluation and spatiotemporal analysis.

2.2. Data Imputation and Consolidation

2.2.1. Data Source

The datasets used through the pipeline are from two sources, the SBS dataset provided by the competition organizer and the local weather data collected from National Oceanic and Atmospheric Administration (NOAA) API, which is a part of Global Historical Climatology Network daily (GHCNd) dataset (National Centers for Environmental Information, 2023). The SBS dataset is originally recorded from a commercial office building located in Mountain View, California, with a total area of 68,000 square feet across two stories (Goldfeder & Sipple, 2023). Various observation and control action data are collected from 127 HVAC devices with a 5-minutes interval, 123 of which have target indoor temperature records. The GHCNd dataset is an integrated collection of daily climate summaries from land-based meteorological stations worldwide, curated and maintained by NOAA (Menne et al., 2012). The geographically closest to the aforementioned office building and relatively complete weather data comes from a station located in San Jose, California, United States (latitude 37.35938, longitude -121.92444). The collected weather contains the daily maximum temperature (TMAX) and daily minimum temperature (TMIN) with full coverage of year 2022.

2.2.2. Data Consolidation

Before imputation, the training and validation datasets are each reorganized into graph structures by grouping the observations data by rooms and sensors. Each room is represented as a node containing metadata (e.g. room area, device information) and room-specific observation data. Two additional global nodes are introduced: `exterior_space`, which holds outdoor sensor data including outdoor temperature (T_{out}), and `global_weather`, which stores TMAX and TMIN from external weather file. This consolidated graph serves as the sole input for subsequent processing.

2.2.3. Data Imputation

To address the substantial data missing problem in SBS dataset, a multi-step imputation method is applied. First, T_{out} is imputed by leveraging external weather file and Machine Learning modeling. A proxy outdoor temperature series is reconstructed by scaling the normalized daily temperature template to match the TMAX and TMIN from the weather file as shown in Equation (1) and (2). Then a LightGBM model is trained to map the relationship of the T_{out} from the sensor and the constructed series, combining with time-related features (hour, weekday, etc.) to impute the missing data in T_{out}. After imputation, a weighted linear smoothing is applied to avoid jump at the joint points.

$$\tilde{x} = \frac{1}{|D|} \sum_{d \in D} \frac{x^{(d)} - \min(x^{(d)})}{\max(x^{(d)}) - \min(x^{(d)})} \tag{1}$$

$$\hat{x}_t^{(i)} = \mathrm{TMIN}_t + \tilde{x}^{(i)} \cdot (\mathrm{TMAX}_t - \mathrm{TMIN}_t) \tag{2}$$

where D denotes the set of days with complete data, $x^{(d)}$ is the observed temperature series for day d, $\tilde{x}$ is the normalized average daily temperature pattern, and $\hat{x}_t^{(i)}$ is the scaled temperature at time index i on day t.

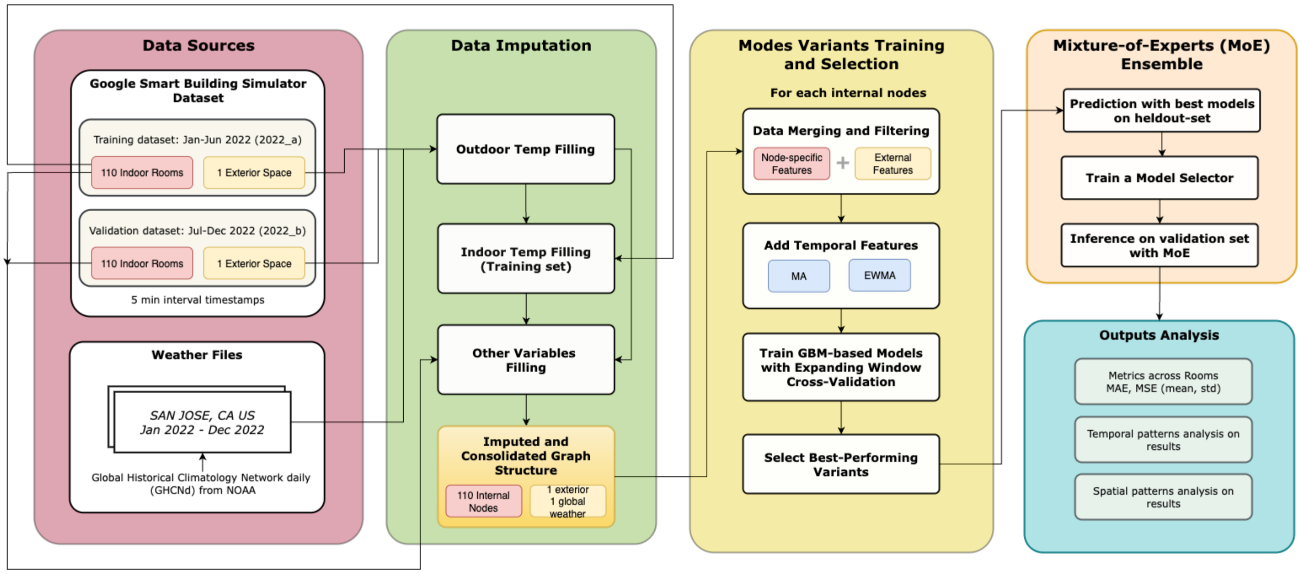
Data Sources
Google Smart Building Simulator Dataset
Training dataset: Jan-Jun 2022 (2022_a)
110 Indoor Rooms
1 Exterior Space
Validation dataset: Jul-Dec 2022 (2022_b)
110 Indoor Rooms
1 Exterior Space
5 min interval timestamps
Weather Files
SAN JOSE, CA US
Jan 2022 - Dec 2022
Global Historical Climatology Network daily (GHCNd) from NOAA
Data Imputation
Outdoor Temp Filling
Indoor Temp Filling (Training set)
Other Variables Filling
Imputed and Consolidated Graph Structure
110 Internal Nodes
1 exterior
1 global
weather
Modes Variants Training and Selection
For each internal nodes
Data Merging and Filtering
Node-specific Features
+
External Features
Add Temporal Features
MA
EWMA
Train GBM-based Models with Expanding Window Cross-Validation
Select Best-Performing Variants
Mixture-of-Experts (MoE) Ensemble
Prediction with best models on heldout-set
Train a Model Selector
Inference on validation set with MoE
Outputs Analysis
Metrics across Rooms MAE, MSE (mean, std)
Temporal patterns analysis on results
Spatial patterns analysis on results

Figure 1. Data processing and modeling framework

Then, the indoor temperatures (T_{in}) across different rooms in training dataset are imputed with the full T_{out} series by similar modeling and smoothing method, which is further applied to the other exogenous variables with the combination of T_{in} and T_{out} as main training features (exclude T_{in} when completing validation dataset). All the imputed rows are marked in an artificial column for future training, enabling the model to learn the potential pattern distinction. With the above iterative imputation steps, 14,931 missing rows out of 51,852 in training set and 13,589 out of 52,716 in validation set are restored, substantially improving the data completeness for subsequent modeling.

2.3. Prediction Model Training and Inference

Based on the completed datasets, the entire training and inference pipeline consists of following stages: (1) temporal context feature engineering, (2) best-performing expert model selection through cross-validation, (3) expert models and model selector training, and (4) inference on the validation set.

2.3.1. Temporal Context Feature Engineering

To enhance temporal consistency and capture short-and-medium-term trends in the sensor data, we apply two common temporal smoothing techniques—Moving Average (MA) and Exponentially Weighted Moving Average (EWMA)—as part of our temporal context feature engineering process. These techniques transform the original time series into smoothed versions that better reflect underlying temporal structures, which are added as additional training features.

We define a general smoothing transformation $\mathcal{S}$, which maps a raw temporal feature sequence $x^{(i)} = \{x_t^{(i)}\}_{t=1}^{T}$ to its smoothed version $\tilde{x}^{(i)} = \{\tilde{x}_t^{(i)}\}_{t=1}^{T}$. Two specific smoothing methods are applied:

- **Moving Average (MA):** A symmetric rolling mean with window size w, defined as:

$$\tilde{x}_t^{(i)} = \frac{1}{w} \sum_{j=0}^{w-1} x_{t-j}^{(i)} \tag{3}$$

- **Exponentially Weighted Moving Average (EWMA):** A recursive smoothing function that emphasizes recent values:

$$\tilde{x}_t^{(i)} = \alpha \cdot x_t^{(i)} + (1-\alpha) \cdot \tilde{x}_{t-1}^{(i)} \tag{4}$$

  where the decay factor $\alpha \in (0,1)$ is computed from the smoothing span s as $\alpha = \frac{2}{s+1}$.

As the span increases, the smoothing parameter α decreases, thereby reducing the weight assigned to the most recent data point and increasing the influence of past observations, which leads to better preservation of medium-to-long-term trends.

In practice, these smoothed values are computed for all numeric columns as additional features in the dataset and configure the following modes.

- `base`: uses external and node-specific features only, without any temporal smoothing features.
- `MA_winW`: adds MA features of each time series from the `base` mode, where W represents the rolling window size. (1 interval equals 5 minutes)
- `EWMA_spanS`: adds EWMA features, where S is the span parameter.

2.3.2. Cross-Validation and Modes Evaluation

To facilitate the training of both expert models and the mode selector in our Soft-MoE framework, we first split the original training set $\mathcal{D}_{\text{train}}$ into two parts: a cross-validation set $\mathcal{D}_{\text{cv}}$, which is used to train and evaluate individual expert modes, and a held-out set $\mathcal{D}_{\text{held}}$, which is formed by selecting a fixed ratio (e.g., the last 10%) of the training sequence. The internal held-out set is reserved for training the MoE model selector and ensuring that the selector does not overfit to the training process of expert modes.

Expanding Window Cross-Validation (EWCV). To robustly evaluate each temporal mode under realistic forecasting conditions, *Expanding Window Cross-Validation* (EWCV) is employed on $\mathcal{D}_{\text{cv}}$. This strategy mimics real-world forecasting where models are updated with more data rolling over time.

We begin by dividing $\mathcal{D}_{\text{cv}}$ into P equal-length, contiguous blocks, where each block $\mathcal{B}_p$ represents a fixed time span:

$$\mathcal{D}_{\text{cv}} = \bigcup_{p=1}^{P} \mathcal{B}_p \tag{5}$$

K folds of validation are defined based on a growing training window. Let p_0 be the number of blocks used in the initial training set (e.g., $p_0 = 3$). For each fold $k \in \{1, \ldots, K\}$, the training and validation subsets are:

$$\begin{aligned} \mathcal{D}_{\text{train}}^{(k)} &= \bigcup_{p=1}^{p_0+(k-1)} \mathcal{B}_p \\ \mathcal{D}_{\text{val}}^{(k)} &= \mathcal{B}_{p_0+k} \end{aligned} \tag{6}$$

This results in an expanding training set and a fixed-width validation set that slides forward in time.

Evaluation Procedure. For each mode (e.g., base, MA with various window sizes, EWMA with various spans), we perform EWCV independently on each room. At each fold, a LightGBM model is trained on $\mathcal{D}_{\text{train}}^{(k)}$ and evaluated on $\mathcal{D}_{\text{val}}^{(k)}$ using MAE and MSE and are aggregated across folds and rooms. The top-performing modes from each group (i.e. `base`, `MA_` and `EWMA_`) are selected as candidate experts for the Soft-MoE ensemble.

2.3.3. Expert Models and Model Selector Training

Given the selected M expert modes (e.g., Base, MA-w^*, EWMA-s^*), we retrain each corresponding expert model using the full training dataset excluding held-out portion. Each expert model is trained per room using LightGBM regressors. The input features include: global features X^{ext}, room-specific features X^{node}, and additional smoothed features depending on the mode, i.e. X^{ma} or X^{ewma}.

After training, each expert model performs inference on the held-out set to generate per-room predictions. These predictions, alongside the input features, serve as supervision signals for training the Soft-MoE model selector to learn which mode performs best under different temporal contexts. Specifically, the selector takes the current input features as predictors and is trained to estimate a soft probability distribution over all expert modes (i.e., mode-wise weights). These probabilities are interpreted as soft confidence scores that guide how much each expert should contribute to the final prediction at each time step.

2.3.4. Inference on the Validation Set

Finally, dynamic inference is performed over the full time span of validation set. At each time step, the model selector generates soft weights for all experts, and the final prediction is computed as the weighted average of the predictions from all experts. Other individual modes are also evaluated on the validation set for further analysis. The full inference process is described in Algorithm 1. Due to the incompleteness of the validation set, the error metrics are only calculated where the true values exists.

3. Results

3.1. Data Imputation Results

In the data imputation process, T_{out} from the outdoor sensor is the the key variable, bridging the external weather file and the other sensor data including T_{in}. The imputation accuracy is evaluated with MAE between the model-predicted values and corresponding ground truth values where available, with 2.0505 and 2.0111 °F on training and validation dataset respectively. Despite the close metric, how the imputed T_{out} series is enveloped within the TMIN and TMAX range is noticeably different between the two datasets, as illustrated in App. Figure 7. While the model generally captures temperature spikes well, during a long missing period in July, it consistently underestimates peak temperatures—potentially due to more substantial deviations of daily patterns from seasonal averages in summer.

Algorithm 1 Inference with Soft-MoE

```
Input: Validation graph $G_{\text{val}}$, trained room-level expert models $\{M_n^{\text{base}}, M_n^{\text{ma}}, M_n^{\text{ewma}}\}$, soft model selector $S$
for each node $n$ in $G_{\text{val}}$ do
    if any model $M_n^{\cdot}$ is missing then
        continue
    end if
    Extract external features $X^{\text{ext}}$ and room-level features $X^{\text{node}}$ for node $n$
    Construct base features $X^{\text{base}}$ = $X^{\text{ext}}$ + $X^{\text{node}}$
    Apply smoothing on $X^{\text{base}}$ to obtain $X^{\text{ma}}$ and $X^{\text{ewma}}$
    Concatenate to form selector input $X^{\text{select}}$ = $X^{\text{base}}$ + $X^{\text{ma}}$ + $X^{\text{ewma}}$
    Use selector $S$ to predict soft weights $\mathbf{w}_t \in \mathbb{R}^3$ for each time step $t$   {Each $\mathbf{w}_t = (w_t^{\text{base}}, w_t^{\text{ma}}, w_t^{\text{ewma}})$}
    Predict expert outputs:
        $\hat{y}_t^{\text{base}} = M_n^{\text{base}}(X_t^{\text{base}})$
        $\hat{y}_t^{\text{ma}} = M_n^{\text{ma}}(X_t^{\text{base}} + X_t^{\text{ma}})$
        $\hat{y}_t^{\text{ewma}} = M_n^{\text{ewma}}(X_t^{\text{base}} + X_t^{\text{ewma}})$
    Compute final prediction:
        $\hat{y}_t = w_t^{\text{base}} \cdot \hat{y}_t^{\text{base}} + w_t^{\text{ma}} \cdot \hat{y}_t^{\text{ma}} + w_t^{\text{ewma}} \cdot \hat{y}_t^{\text{ewma}}$
    Record prediction $\hat{y}_t$, expert weights $\mathbf{w}_t$, and ground truth $y_t$
end for
Output: Room-level predictions, ground truth and soft expert weights at each timestamp.
```

3.2. Prediction Results Overview

To better assess the impact of different mode configurations, all tested modes—including the MoE approaches-alongside the AutoRegressive Integrated Moving Average (ARIMA) model (Box & Jenkins, 1970) as the baseline (with hyperparameters selected via grid search) are deployed to generate predictions over the entire 6-month validation period. The prediction results are summarized in Table 1, reporting the mean and standard deviation (std) of MAE and MSE across all rooms with target variable T_{in}.

Based on cross-validation results, the best-performing variants from each group—`base`, `MA_win108`, and `EWMA_span108`—are selected as expert candidates for the MoE framework. In addition to the Soft-MoE described in Section 2.3.3, a Hard-MoE variant is also evaluated for comparison, where the selector chooses the single best expert for each instance instead of computing a weighted ensemble of all expert predictions. The Soft-MoE mode achieved the best overall performance on validation set with a mean MAE of 1.1839 °F. Despite the marginal improvement over the second-best mode in terms of mean MAE (i.e., `MA_win72`), it ensures both accuracy and stability with 14.47% lower standard deviation across rooms. Among the individual modes, `MA_win72` performed the best, reducing the average MAE by 62.29% and 9.66% compared to the baseline and `base` mode respectively. While EWMA modes show slightly higher errors, they offer more stable predictions; notably, `EWMA_span72` achieves an 20.64% reduction in the std of MSE compared to `base` mode. The optimal rolling window or span is 72-108 intervals, suggesting that a 6-to-9-hour historical context strikes a good balance between capturing long-term trends and short-term variations.

In terms of runtime, on a MacBook Air equipped with an Apple M3 chip, all MA and EWMA variants require 2.20–2.62 seconds per room for training and inference, while the base mode is notably faster at 1.82 seconds. For the MoE modes, assuming the models have been pre-trained, per-room inference takes 1.66 seconds (Soft-MoE) and 1.35 seconds (Hard-MoE). These LightGBM-based methods show strong potential for lightweight, real-time forecasting with minimal computational overhead.

3.3. Temporal Pattern Investigation

General error metrics alone are not sufficient to fully capture the nuance differences between modes. In this section, we further compare the performance of various modes with respect to temporal characteristics, such as seasonal pattern and their responsiveness to abrupt or periodic changes in the series. As shown in Figure 2, it is interesting to notice that all modes have best performance during hottest months. The Soft-MoE and `MA_win72` modes demonstrate more consistent performance throughout the year. In contrast, EWMA-based modes lag behind notably during the colder months.

Nevertheless, EWMA-based modes demonstrates more robust stability in response to recurrent fluctuations in temperature as illustrated in Figure 3 and Figure 4. Between mid-September and mid-October, the EWMA-based modes demonstrated a stronger ability to adapt to cyclical fluctuations in daily temperature, achieving higher prediction accuracy, followed by the MA modes with long windows. Interestingly, the MA modes with short windows performed even worse than the base mode in this period. However, with the intense volatility of daily temperature during winter, EWMA modes lose advantage while all modes underperformance. This observation confirms that a smaller α (bigger *span*) allows EWMA to more effectively capture long-term smoothing patterns, at the tradeoff of reduced sensitivity to short-term spike, which further highlights the effectiveness of a dynamic model selection approach and suggests that in practical applications, selecting an appropriate model or weights should be guided by the expected temporal characteristics of the data.

3.4. Spatial Pattern Investigation

The prediction error is relative stable across all rooms as shown in the room-level MAE heatmap (Figure 5), ex-

Table 1. Prediction results across all rooms with different settings.

Mode	MAE mean (°F)	MAE std (°F)	MSE mean ($^\circ F^2$)	MSE std ($^\circ F^2$)
ARIMA (baseline)	3.1532	0.9477	18.2929	9.1192
Base	1.3161	0.3426	3.3299	1.6778
MA_win12	1.2715	0.3613	3.1546	1.7568
MA_win36	1.2258	0.3761	2.9565	1.7961
MA_win72	1.1889	0.3871	2.7386	1.7948
MA_win108	1.1938	0.3303	**2.6652**	1.5324
EWMA_span12	1.3292	0.2734	3.3741	1.3090
EWMA_span36	1.3810	0.2865	3.5664	1.3253
EWMA_span72	1.3463	**0.2719**	3.3252	**1.2267**
EWMA_span108	1.3058	0.2882	3.0974	1.2913
Soft-MoE	**1.1839**	0.3311	2.6841	1.4609
Hard-MoE	1.2547	0.3227	2.9874	1.4886

MAE and MSE values are first computed individually for each room and then aggregated across all rooms to calculate the mean and standard deviation value.

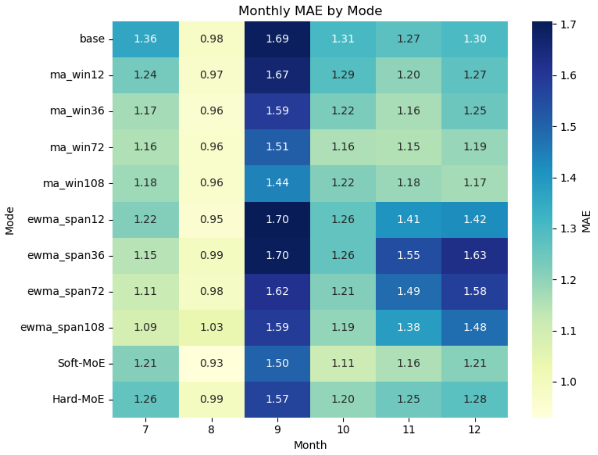

Figure 2. Average MAEs by different modes and months across all rooms.

cept for several rooms on the second floor with notably high MAEs. By calculating the Pearson correlation between all trained features and MAEs, it is found that the features related to supply air control—such as the std of flowrate and damper commands—show the highest correlation with prediction error. For example, room `2-16` with the highest prediction error has a supply air flowrate setpoint std (`supply_air_flowrate_setpoint_std`) 13.84% higher than the average across all rooms. This finding suggesting that rooms with more dynamic HVAC regulation are harder for the model to predict accurately.

4. Conclusion

This study was conducted in response to the Smart Buildings Competition and proposes an automated data imputation, modeling and prediction pipeline that leverages external weather information and Gradient Boost Regression models. The pipeline incorporates temporal smoothing methods such as MA and EWMA to enrich the model with historical context. Building upon these individual models, a Soft Mixture-of-Experts (Soft-MoE) framework is developed, which assigns learned weights to each expert at every time step, further enhancing predictive accuracy and robustness across varying temporal patterns.

Among the various modes, the Soft-MoE mode achieved the overall best performance on the whole validation set, reaching a MAE of 1.1839 °F. Both the MA and EWMA-based features demonstrated improved compatibility with temperature fluctuations at daily scale, validating the effectiveness of integrating traditional time series modeling techniques into ML workflows, which shows greater flexibility and ease of deployment while maintaining competitive predictive accuracy.

Several directions remain open for future work. First, al-

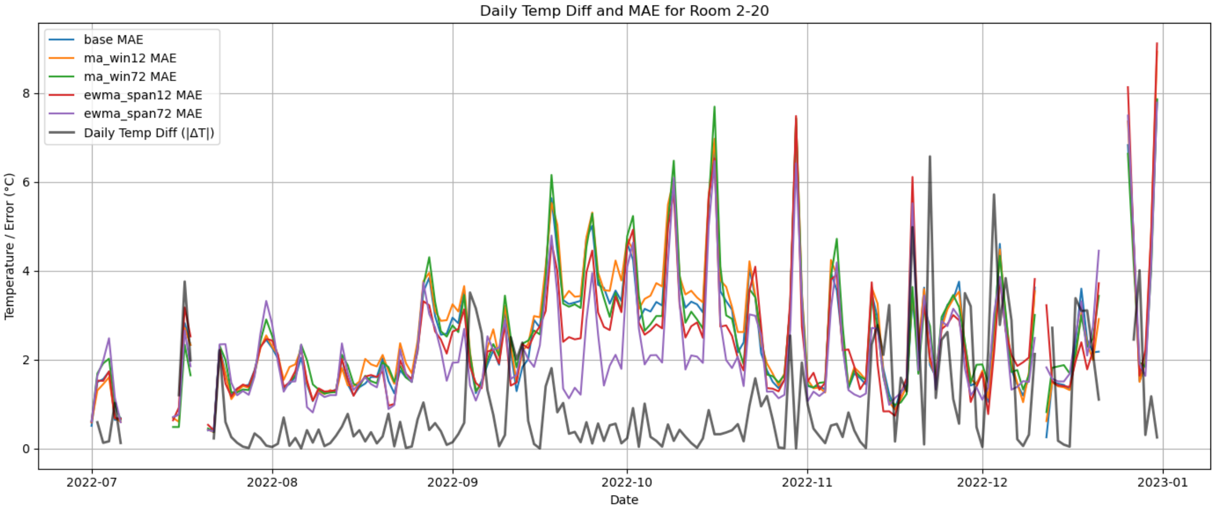

Figure 3. Prediction accuracy comparison across different modes in respect to daily temperature variation (daily temperature difference is calculated as the first-order difference between today's and yesterday's average temperatures).

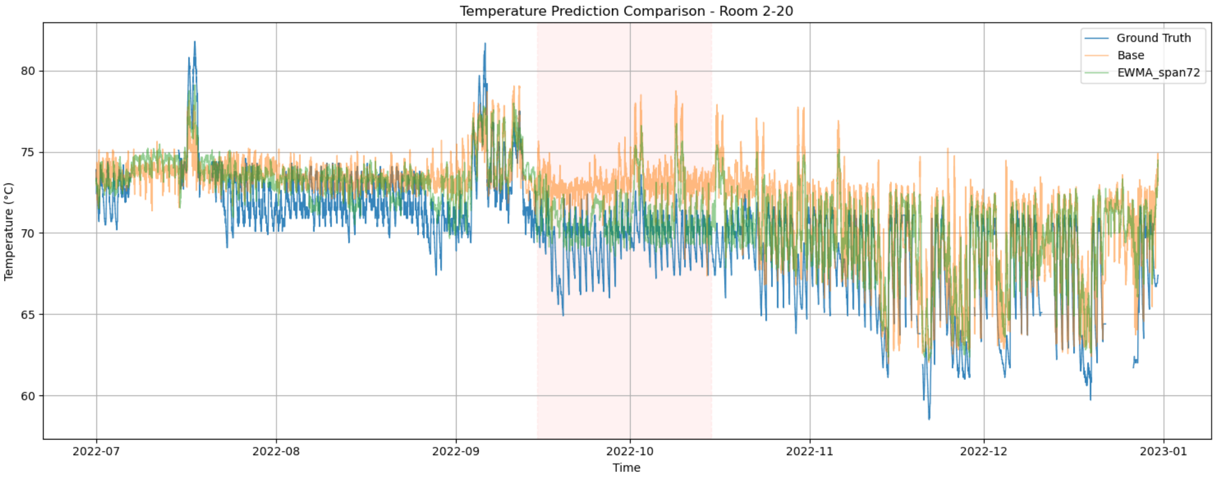

Figure 4. The EWMA mode demonstrates improved performance during the seasonal regime shift from September to October, characterized by a cooling trend and pronounced daily temperature fluctuations.

though the dataset is organized as a graph, its structure has not been fully utilized for feature enhancement—such as incorporating physical attributes of individual rooms or modeling inter-room thermal dynamics. Second, the adaptability of EWMA to different temporal fluctuations highlights its potential for further exploration, for instance, introducing a dynamic self-calibration mechanism for the smoothing parameter α in an online prediction setting, to enable the model to better accommodate various temporal pattern and improve real-time forecasting robustness.

References

Box, G. E. P. and Jenkins, G. M. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970. ISBN 978-0-8162-1094-7.

Goldfeder, J., Dean, V., Jiang, Z., Wang, X., dong, B., Lipson, H., and Sipple, J. The smart buildings control suite: A diverse open source benchmark to evaluate and scale hvac control policies for sustainability, 2025. URL https://arxiv.org/abs/2410.03756.

Goldfeder, J. A. and Sipple, J. A. A Lightweight Calibrated

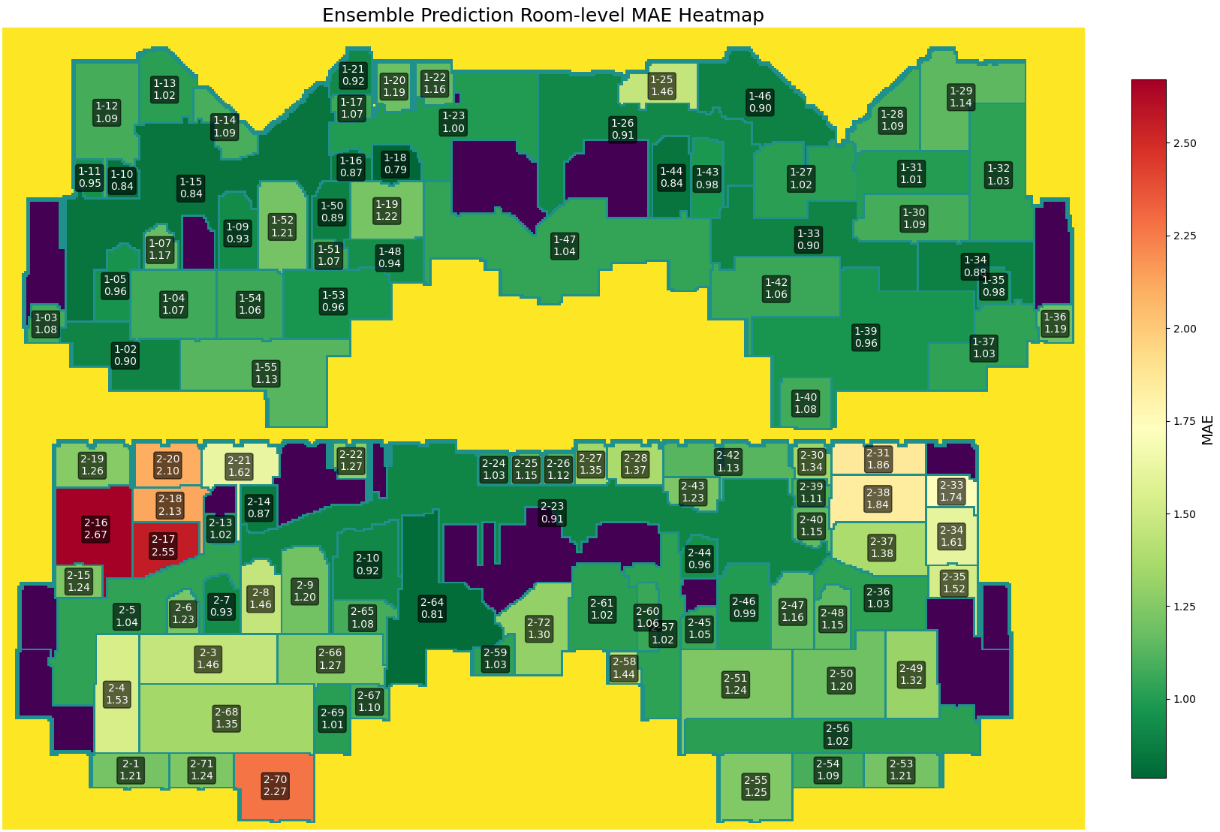

Figure 5. MAE heatmap across all involved rooms under Soft-MoE mode.

Simulation Enabling Efficient Offline Learning for Optimal Control of Real Buildings. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pp. 352–356, Istanbul Turkey, November 2023. ACM. ISBN 979-8-4007-0230-3. doi: 10.1145/3600100.3625682.

Jiang, K., Shi, T., Yu, H., Mahyuddin, N., and Lu, S. A systematic review of multi-output prediction model for indoor environment and heating, ventilation, and air conditioning energy consumption in buildings. *Indoor and Built Environment*, 33(9):1574–1604, November 2024. ISSN 1420-326X. doi: 10.1177/1420326X241258678.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pp. 3146–3154, 2017.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, 29(7):897–910, July 2012. ISSN 0739-0572, 1520-0426. doi: 10.1175/jtech-d-11-00103.1.

Mtibaa, F., Nguyen, K.-K., Azam, M., Papachristou, A., Venne, J.-S., and Cheriet, M. LSTM-based indoor air temperature prediction framework for HVAC systems in smart buildings. *Neural Computing and Applications*, 32 (23):17569–17585, December 2020. ISSN 1433-3058. doi: 10.1007/s00521-020-04926-3.

National Centers for Environmental Information. Global historical climatology network - daily (ghcn-d). https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily, 2023. Accessed: 2025-07-08.

Palaić, D., Štajduhar, I., Ljubic, S., and Wolf, I. Development, Calibration, and Validation of a Simulation Model for Indoor Temperature Prediction and HVAC System

Fault Detection. *Buildings*, 13(6):1388, June 2023. ISSN 2075-5309. doi: 10.3390/buildings13061388.

Sasaki, R., Kato, K., Zhao, D., Nishikawa, H., Taniguchi, I., and Onoye, T. Indoor temperature prediction for hvac energy management using smart remote controller. In *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '24, pp. 225–226, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707063. doi: 10.1145/3671127.3698702. URL https://doi.org/10.1145/3671127.3698702.

Xu, X., Fu, B., Wu, Z., and Sun, G. Predictive control for indoor environment based on thermal adaptation. *Science Progress*, 104(2), April 2021. ISSN 0036-8504, 2047-7163. doi: 10.1177/00368504211006971.

A. Appendix

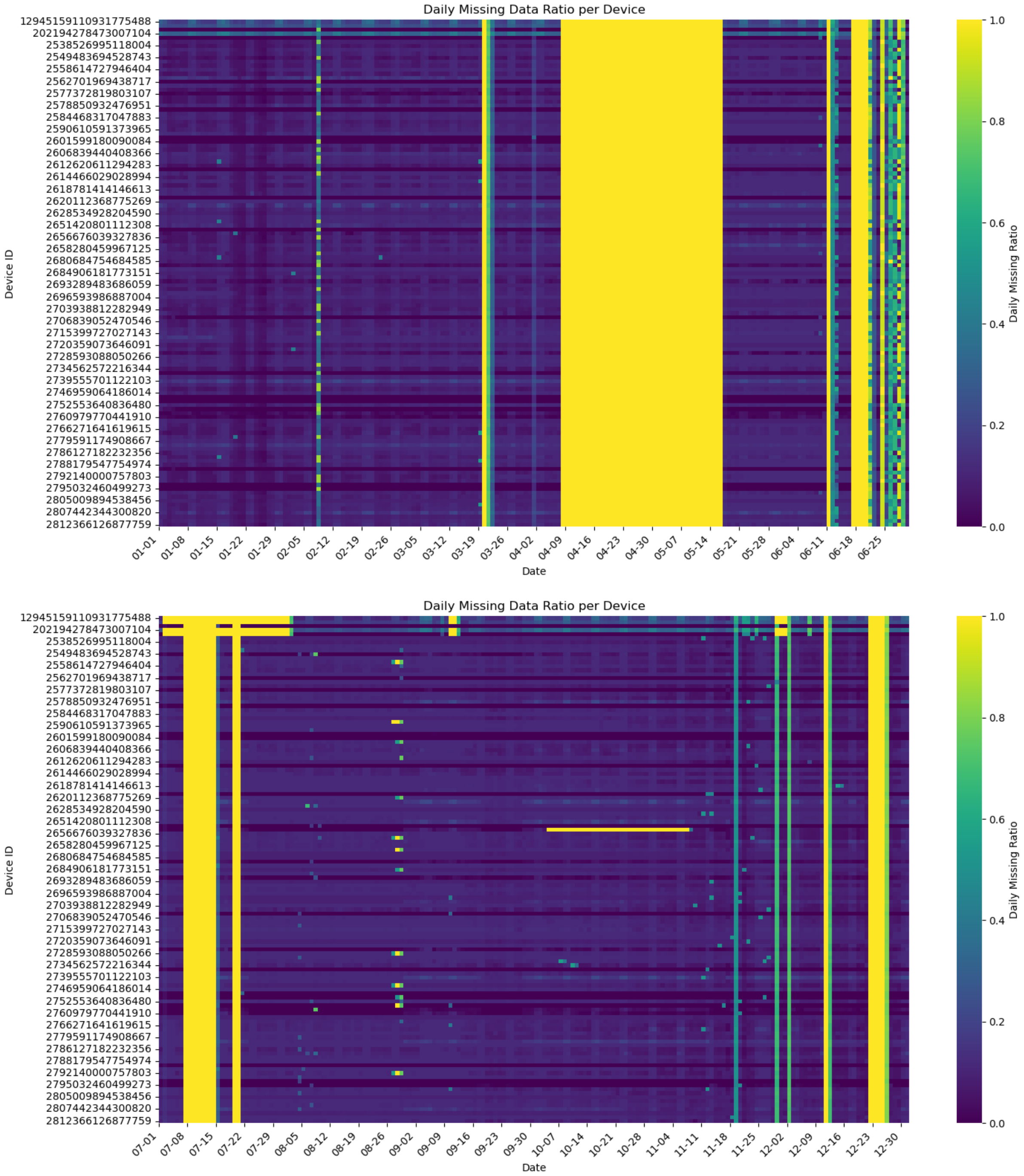

Figure 6. Missing data points across timestamps and sensors in training dataset (above) and validation dataset (lower).

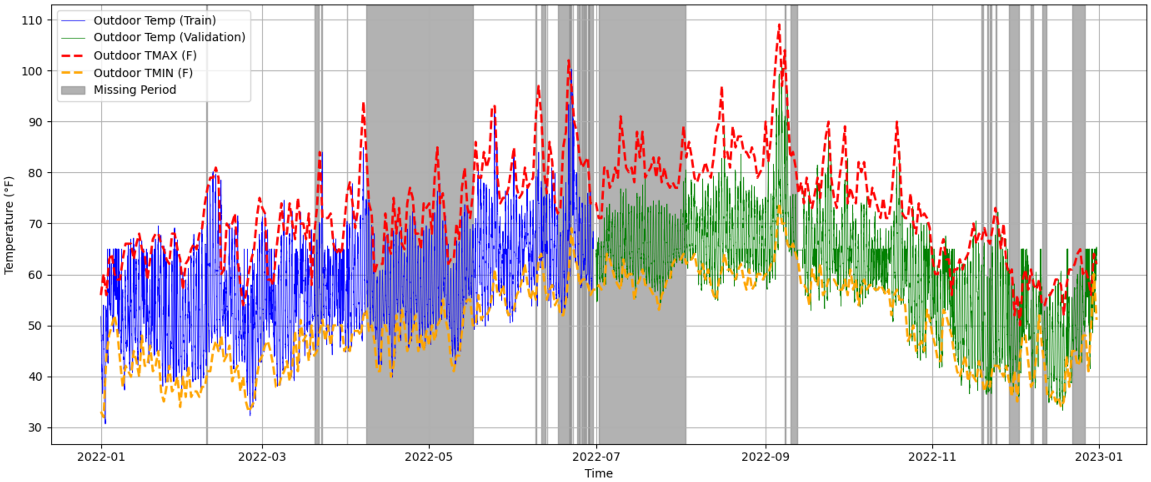

Figure 7. Filled outdoor temperature series with external weather data