CO-BUILD Smart Buildings Competition: An Empirical Comparison of HVAC Temperature Prediction Models

Sohei Arisaka [1] **Eikichi Ono** [1] **Hiroyasu Miura** [2] **Yutaka Shoji** [2] **Yangayang Li** [2] **Kuniaki Mihara** [2]

Abstract

This paper presents a comprehensive comparison of temperature prediction models for HVAC systems using the CO-BUILD Smart Buildings Competition dataset. We evaluate five modeling approaches—Naive Mean, Light Gradient Boosting Machine (LightGBM), Time-series Dense Encoder (TiDE), Time Series Foundation Model (TimesFM), and a Multimodal Large Language Model—across prediction horizons from 5 minutes to 2 weeks. Through exploratory data analysis, we identify key building characteristics, device relationships, and operational patterns that inform our preprocessing pipeline, which includes timezone conversion, missing data handling, and feature selection incorporating both direct VAV measurements and cross-device CO_2 influences. Our results demonstrate that LightGBM achieves superior short-term performance (up to 3 hours), while TiDE proves effective for longer horizons. TimesFM accurately predicts weekly temperature patterns in a zero-shot setting, and a multimodal LLM exhibits unique reasoning capabilities, successfully forecasting temperature shifts during operational transitions. This study provides practical insights for model selection in building energy management systems.

1. Introduction

Smart buildings leverage high-resolution temporal and spatial data collection to monitor occupant behavior, indoor environmental conditions, and system states. This rich data ecosystem enables the deployment of sophisticated applications for analysis, prediction, and control optimization. For HVAC (Heating, Ventilation, and Air Conditioning) systems, achieving optimal operation requires balancing occupant comfort with energy efficiency through dynamic optimization approaches that account for the inherent variability in building and system behavior. Such optimization strategies necessitate accurate predictive models for boundary conditions, including occupancy patterns and outdoor weather conditions.

Since most HVAC systems are operated to maintain indoor temperatures within a specific comfort range, accurate prediction of future indoor temperatures plays a critical role in implementing dynamic optimal control. While typical HVAC systems operate on a daily schedule (e.g., from 8:00 AM to 7:00 PM), making a prediction time horizon of several hours to one day generally sufficient, systems with greater thermal capacity, such as those incorporating thermal energy storage, may require longer horizons spanning several days. Additionally, for integrated energy demand-supply control at district or urban scale (e.g., demand response systems), even longer prediction horizons of several weeks may be required.

In this competition, we set a practical prediction time horizon ranging from five minutes up to two weeks. We investigate the prediction accuracy of various modeling approaches over short- to long-term forecasting horizons.

2. Data Overview

This section provides an overview of the provided dataset. The dataset consists of one year of HVAC system operation data with five-minute interval from a building (Goldfeder et al., 2025). First half of the data is used for training and the second half is used for testing. The prediction target is the indoor temperature obtained from temperature sensors in 123 VAV (Variable Air Volume) boxes. We refer to this temperature time series data as target and all other data as covariates, which can be used as features for prediction. The covariates are further classified into three categories: past covariates are covariates known only into the past, future covariates are covariates known into the future, and static covariates are covariates that remain constant over time. Table 1 shows the details of each covariate.

[1]Kajima Technical Research Institute Singapore, Kajima Corporation, Singapore [2]Kajima Technical Research Institute, Kajima Corporation, Tokyo, Japan. Correspondence to: Sohei Arisaka <s.arisaka@kajima.com.sg>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Table 1. Data overview.

Category	Variable	Description
Target	Zone air temperature	Zone air temperature of each VAV
Past covariates	VAV sensor values	All measurement data of each VAV
	AHU sensor values	All measurement data of 2 AHUs
	Outdoor conditions	All measurement data of weather station
	CO_2 setpoints	All CO_2 setpoints of all devices
Future covariates	TMY data	Typical Meteorological Year data
	Scheduled setpoints	Estimated setpoint schedules
	Time features	Month, day, hour, day of week, etc.
Static covariates	Device ID	Unique ID of each target device

2.1. Building Location and Weather Data

Since the only locational information provided was that the buildings are located in the United States, we estimated the geographical region based on the provided outdoor air temperature data. The temperature patterns suggest a region with four distinct seasons but relatively mild winters, indicating a likely location in the southern United States. By comparing monthly statistical values of the provided weather data with TMY (Typical Meteorological Year) datasets (OneBuilding, 2025) and actual weather data in 2022 (Time & Date, 2025) for several cities, we estimated that the building is located in or near San Francisco, California, as shown in Figure 1. Based on this assumption, we used the TMY data for San Francisco as future covariates in our modeling.

The provided timestamps appear to be in UTC format. Consistent with our location analysis placing the building in California, we converted all timestamps to Pacific Time (America/Los_Angeles, GMT-8/GMT-7) to facilitate accurate temporal analysis and account for daylight saving time transitions.

2.2. Connections between Zones and Devices

The target building is a two-story facility that regulates indoor temperature and CO_2 concentration across over 500 rooms using 123 VAV boxes equipped with thermostats. While the device registry lists six Air Handling Units (AHUs), analysis of the naming conventions and operational status indicates that duplicated and non-functional entries exist, suggesting that two primary AHUs are responsible for air conditioning operations.

Analysis of VAV-zone relationships revealed that the number of zones connected to a single VAV ranges from 1 to 41, and the number of VAVs connected to a single zone ranges from 1 to 8. Since mechanical drawings were not provided, it is difficult to determine the control logic or spatial relationships for devices connected to multiple zones. Including all connected VAVs through zones as features may overcomplicate the modeling with excessive inter-VAV interactions. Therefore, for single-VAV modeling, we decided to limit features to those directly relevant to the target VAV, including associated AHU data and outdoor air conditions.

To elucidate relationships between measurement variables, we applied Dynamic Mode Decomposition (DMD) (Schmid, 2010) and analyzed the resulting coefficient matrix shown in Figure 2. This analysis revealed that CO_2 setpoints from VAVs actively managing CO_2 levels significantly influence room temperatures of VAVs not directly involved in CO_2 control. We hypothesize that when a VAV adjusts its operation to satisfy CO_2 control requirements, it affects the operational state of the connected AHU, which subsequently influences room temperatures of other VAVs served by the same AHU. To capture this cross-device influence, we included CO_2 setpoints from all devices as past covariates in our modeling approach.

2.3. Cooling and Heating Temperature Setpoints

Analysis of cooling and heating zone air temperature setpoint data revealed a one-hour shift in HVAC operation timing due to daylight saving time and seasonal changes in the setpoint values themselves. Using the training period data and weekly average outdoor temperatures calculated from San Francisco's TMY data, we identified a threshold of 14 °C as a decision point for adjusting setpoints. This outdoor temperature-dependent schedule was used as a future covariate to predict targets.

2.4. Missing and Duplicated Data

Both training and test periods contain intervals where target zone air temperatures are continuously recorded as zero. Given the physical implausibility of sustained zero temperatures, we treated these periods as missing data. Since missing target values preclude meaningful error evaluation, these intervals were excluded from both training and evaluation procedures. Additionally, we identified complete data gaps (e.g., June 30) where timestamps themselves are absent.

The test period data also contains duplicate entries, specifically complete data duplication for July 29, 2022. These duplicates were removed to ensure valid model evaluation.

2.5. Temperature Unit Inconsistency

While all devices report temperature measurements in Fahrenheit, device ID `16286830034440683520` uses Kelvin units. To maintain evaluation consistency across all targets, we converted this device's temperature values to Fahrenheit for both training and evaluation phases.

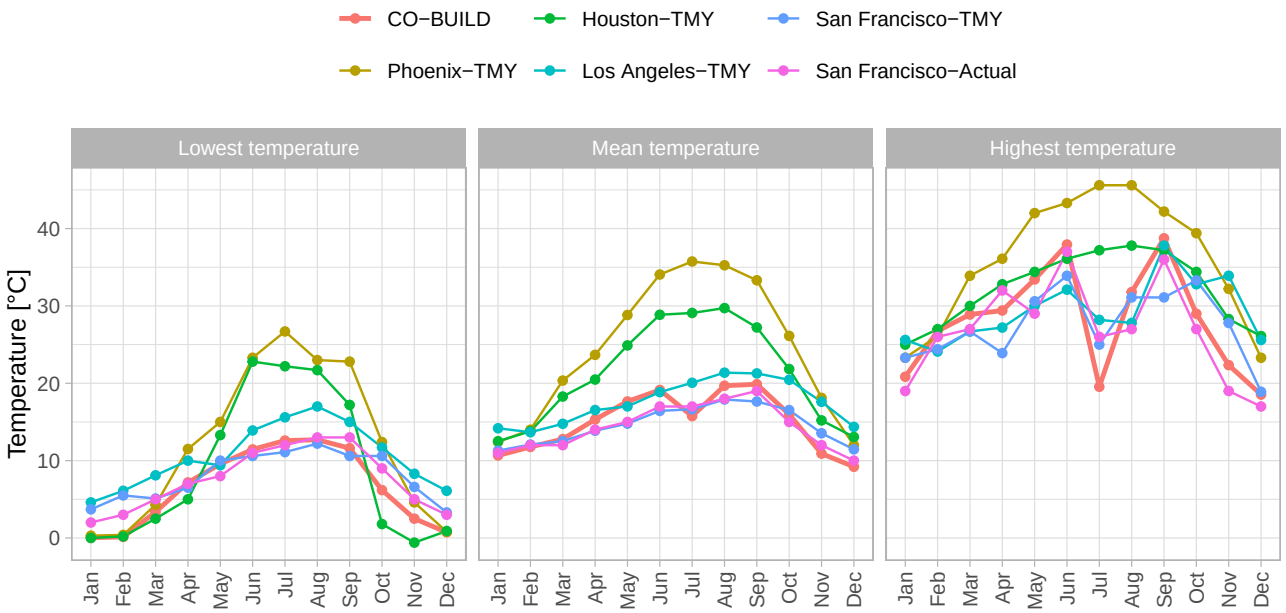

Figure 1. Monthly statistics of outdoor temperatures in US cities.

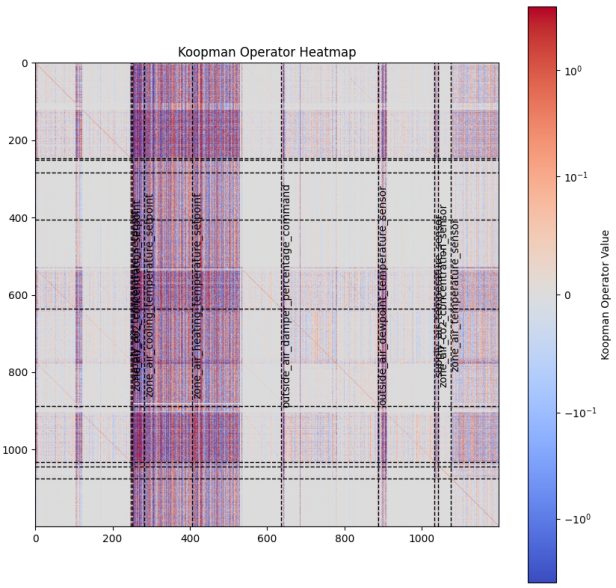

Figure 2. The DMD coefficient matrix.

3. Modeling Approach

This section details our modeling methodology for the competition. Our fundamental strategy involves training unified machine learning models capable of predicting target values across all VAVs. Given that observable fields vary slightly across devices, we standardized the feature space by taking the union of all observable fields and imputing missing values with -1 for devices lacking specific measurements.

We evaluated five distinct modeling approaches—Naive Mean, Light Gradient Boosting Machine, Time-series Dense Encoder, Time Series Foundation Model, and Multimodal Large Language Model—across ten forecasting horizons ranging from 5 minutes to 2 weeks. To ensure implementation consistency and evaluation reliability, we utilized the Darts library (Herzen et al., 2022). Detailed descriptions of each model follow.

Unless otherwise specified, models utilize historical input windows matching the prediction horizon length. Specifically, for n-step ahead forecasting, models incorporate: ordinally encoded static covariates, past n steps of past covariates and target values, and both past and future n steps of future covariates. This design enables models to capture historical patterns while leveraging known future information such as scheduled setpoints and temporal features.

To prevent overtuning to the test data (Schneider et al., 2025), we strictly excluded all evaluation period data (July 1st onwards) from the modeling pipeline, including hyperparameter tuning. Model optimization was conducted using validation splits of the training data (through June 30th).

3.1. Naive Mean

We implemented a naive mean baseline that predicts future target values as the average of past values within a window equal to the prediction horizon. For instance, 6-hour forecasts use the mean of the preceding 6 hours of target values. This simple baseline provides a performance reference for evaluating more sophisticated approaches.

3.2. Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework utilizing tree-based learning algorithms (Ke et al., 2017). We employed a multi-step forecasting strategy that trains separate models for each prediction step. For n-step forecasting, this approach involves training n specialized models, each targeting a specific time step ahead. This step-specific design enables the capture of distinct temporal patterns at different horizons, potentially improving accuracy over single models predicting all steps simultaneously.

All hyperparameters were maintained at Darts library defaults.

3.3. Time-series Dense Encoder

Time-series Dense Encoder (TiDE) (Das et al., 2023) is a multi-layer perceptron (MLP) architecture designed for long-term time series forecasting. Unlike traditional recurrent or convolutional approaches, TiDE employs a streamlined dense encoder-decoder framework that captures complex temporal patterns while maintaining computational efficiency.

The architecture features a dense encoder processing historical time series and covariates to extract temporal features, coupled with a dense decoder generating multi-step predictions. This design enables TiDE to handle both short-term dependencies and long-term trends while seamlessly integrating past and future covariates—capabilities particularly well-suited for HVAC systems where scheduled setpoints and temporal features provide valuable predictive information.

We configured TiDE with L1 loss, 512 hidden units, 3 encoder and decoder layers, 32-dimensional decoder output, and 0.1 dropout. The batch size was set to 1024, with learning rates determined via the Darts learning rate finder (Smith, 2017). Training proceeded for 20 epochs with learning rate reductions by factor 10 at epochs 10 and 15.

3.4. Time Series Foundation Model

Foundation models such as TimesFM (Das et al., 2024) and Chronos (Ansari et al., 2024) represent emerging paradigms for time series forecasting. These pre-trained transformer-based models leverage diverse temporal patterns learned from large-scale datasets to achieve competitive performance across domains with minimal task-specific adaptation.

We employed the pre-trained TimesFM model for one-week (168-hour) predictions using four-week (672-hour) input contexts. Each VAV zone was treated as an independent time series, with zone-specific predictions performed separately rather than multivariate forecasting. No model fine-tuning was conducted.

Foundation models offer several advantages for this application: reduced training data requirements, robustness to operational anomalies, and the ability to capture complex temporal patterns without domain-specific architecture design. However, they also present limitations including lack of physical constraints, limited interpretability, and significant computational overhead.

Additionally in this competition, it should be noted that the target dataset may have been included in the foundation model's pre-training data, which could potentially influence the evaluation results and should be considered when interpreting performance comparisons.

3.5. Multimodal Large Language Model

We present a novel methodology for time series forecasting using multimodal Large Language Models (LLMs). While conventional forecasting relies on sequence models, recent advances suggest that LLM reasoning capabilities can enhance predictive accuracy. Specifically, providing time series data across multiple modalities—text, image, and audio—has shown promise for improving model performance.

Daswani et al. (2024) compared the predictive accuracy of a multimodal LLM when time series data were provided as either text or image. It showed that image-based inputs led to a classification accuracy improvement of up to 120% and a tenfold reduction in token usage and cost. However, the study addressed a classification task, fall detection from wearable sensor data. In this paper, we extend the multimodal approach to a time series forecasting task. Additionally, we examine whether converting time series data into audio format and supplying it to the LLM yields further benefits.

We employed the latest Gemini 2.5 Pro (Comanici et al., 2025), which features an extended context window and enhanced visual reasoning capabilities. However, due to the constraint of LLM inference cost, we limited the scope of our experiments as follows:

- Forecast Horizons: Constrained by Gemini's maximum output token limit (65,535 tokens), three horizons were chosen: 6 hours, 24 hours, and 3 days.
- Devices: Four devices were randomly selected.
- Observable Fields: A common set of fields was used to ensure data consistency: zone air temperature, cooling temperature setpoint, heating temperature setpoint, supply air flowrate setpoint, supply air flowrate sensor, and outside air temperature sensor.
- Prediction Start Dates: To ensure data stability, three start dates were selected: Oct 15th 00:00, Oct 17th

00:00, and Oct 20th 00:00.

For the context for the LLM, time series data from the six observable fields listed above were visualized as graphs and provided to the model as images. Since this is a forecasting task, we also provided the context data for the target variable as raw text in addition to the image. The length of the historical data provided to the model was set to be identical to the forecast horizon (e.g., a 3-day context for a 3-day forecast). Finally, to investigate the potential impact of an additional modality, we converted the indoor temperature time series into an audio file and included it in the context.

We prompted the model to return structured output, containing 5-minute interval forecasts for the indoor temperature. The input prompt was structured to first present a series of time-series graphs, each accompanied by a brief text description (e.g., "This graph shows the outside air temperature."). Following the images, the full raw text data for the target indoor temperature was provided, along with the audio file. Finally, a set of instructions detailed the forecasting task, including the horizon, interval, and required structured output format. Each experiment was repeated five times with different random seeds, and the overall MAE for each forecast horizon was calculated by averaging the results across all seeds, start dates, and devices.

4. Results and Discussion

This section presents the evaluation results for each model across different prediction horizons. The evaluation uses Mean Absolute Error (MAE) as specified in the competition rule. The ground truth data consists of all data from the test period (July 1st onwards) where target values are not missing. Note that the number of evaluation points varies by device due to slight differences in missing data periods across devices.

For prediction evaluation, we performed forecasting by sliding the prediction window one step at a time with overlapping predictions over the entire test period. The absolute error between each prediction and the corresponding ground truth was calculated, and the final MAE was computed as the average across all predictions on all devices.

Table 2 summarizes the MAE results for each model across different prediction horizons. The results demonstrate distinct performance characteristics for each modeling approach across various temporal scales.

4.1. Naive Mean

The Naive Mean model serves as our baseline, predicting future temperatures as the mean of past values within the same time period as the prediction horizon. As shown in Table 2, the model achieves low MAE values for short-term predictions (0.0704 °F to 0.418 °F for 5 minutes to 1 hour), reflecting the temporal stability of indoor temperatures in HVAC-controlled environments.

Performance deteriorates significantly for medium-term predictions (0.993 °F to 1.83 °F for 3-12 hours) as the averaging assumption becomes less valid, particularly failing to capture diurnal temperature variations such as the transition from daytime to nighttime temperatures.

For long-term predictions (1 day to 2 weeks), the MAE gradually increases from 1.55 °F to 1.88 °F, indicating progressive deterioration as the prediction horizon extends. This degradation may occur because seasonal temperature changes become increasingly influential over longer periods, which the simple mean-based approach cannot capture.

4.2. Light Gradient Boosting Machine

The LightGBM model demonstrates strong performance for short-term predictions, achieving MAE values of 0.0834 °F, 0.163 °F, 0.230 °F, and 0.434 °F for 5-minute, 30-minute, 1-hour, and 3-hour horizons respectively. These results represent significant improvements over the Naive Mean baseline, particularly for the 30-minute to 3-hour prediction horizons where LightGBM achieves approximately 33-56% lower MAE.

Interestingly, for the 5-minute prediction horizon, LightGBM performs slightly worse than the Naive Mean baseline (0.0834 °F vs 0.0704 °F). A possible reason is that the inherent temporal stability of indoor temperatures in HVAC-controlled environments makes simple averaging highly effective for very short-term predictions, while the complexity of the LightGBM model may introduce unnecessary noise for such stable conditions.

However, our LightGBM implementation faces computational limitations for longer prediction horizons (6 hours and beyond), as evidenced by the missing results in Table 2. This limitation arises from the multi-step forecasting approach requiring separate models for each prediction step, combined with larger feature sets needed for extended temporal dependencies. The result is prohibitive computational requirements for horizons beyond 3 hours.

To enable LightGBM for longer-term predictions, strategies such as data subsampling, feature aggregation, and alternative forecasting approaches would be necessary, trading some predictive granularity for computational feasibility.

4.3. Time-series Dense Encoder

The TiDE model demonstrates strong performance across various prediction horizons, as shown in Table 2. For short-term predictions, TiDE achieves competitive performance with 0.0708°F, 0.205°F, and 0.264°F MAE for 5-minute,

30-minute, and 1-hour horizons respectively. The model particularly excels at medium-term predictions, achieving the best performance for 6-hour (0.869°F), 12-hour (1.01°F), and 1-day (1.05°F) horizons.

Figure 3 illustrates the training loss curves for different prediction horizons, comparing the 2-week and 30-minute forecasting tasks. The results reveal that longer prediction horizons, particularly the 2-week forecast, suffer from reduced training data availability due to missing target values in the dataset. This data scarcity leads to potentially underfitting, as evidenced by the learning curves. The comparison suggests that with increased data availability, there is considerable room for improvement in long-term prediction performance. This limitation highlights the importance of data completeness for training robust models capable of extended forecasting horizons.

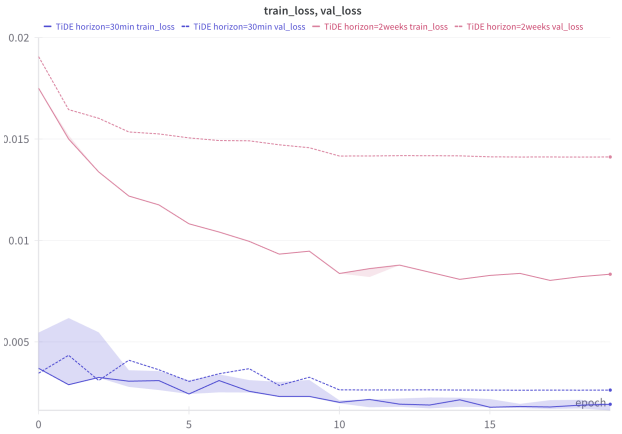

Figure 3. Loss curves of TiDE for different prediction horizons. The x-axis represents the training epochs, and the y-axis represents the normalized loss value. Red lines represent the 2-week prediction task, while blue lines represent the 30-minute prediction task. Dotted lines indicate the validation loss, while solid lines indicate the training loss.

4.4. Time Series Foundation Model

The TimesFM demonstrated excellent performance for one-week temperature predictions in building environments. As shown in Figure 4, the model accurately captured actual temperature variation patterns for a specific device during a week in August, showcasing its ability to handle complex temporal dynamics in building systems.

For validation, we focused exclusively on datasets containing five consecutive weeks of data, incorporating both the five-week input context and the prediction horizon. This constraint ensured data continuity while enabling proper evaluation of the model's forecasting capabilities under realistic operational conditions.

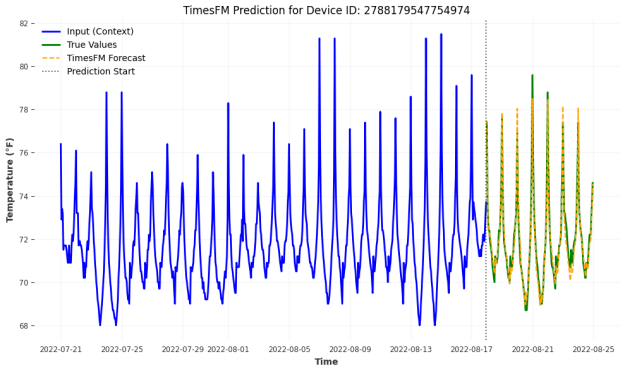

Figure 4. A one-week temperature prediction by TimesFM. The blue line represents the input, the green line represents the ground truth, and the dotted yellow line represents the prediction.

Quantitative evaluation on the entire validation dataset yielded a Mean Absolute Error (MAE) of 1.37 degrees Fahrenheit, demonstrating that TimesFM achieves practical-level accuracy for building temperature prediction tasks. This performance indicates the model's effectiveness in capturing the complex patterns inherent in building thermal dynamics.

However, the one-week sliding window prediction approach showed limitations when encountering behaviors that deviated from historical context patterns. Specifically, prediction accuracy degraded during periods of seasonal transitions, unusual weather conditions, or sudden changes in building operational patterns. These limitations suggest that the fixed prediction horizon constraint may require more flexible forecasting strategies to handle exceptional circumstances effectively.

4.5. LLM-based Forecasting

Since the scope of this experiment is different from the other models, we do not compare its performance with the others. Instead, we present here an illustrative example to highlight the unique capabilities of the LLM in time series forecasting.

Figure 5 shows a prediction result for a device, with a 3-day forecast horizon starting from Oct 15, 2022, 00:00. This task is particularly challenging because the context consists of three weekdays, while the forecast period covers a weekend followed by a weekday.

Remarkably, the model accurately predicted that the indoor temperature would drop during the weekend and then rise sharply on Monday morning, Oct 17, despite this behavioral pattern not being present in the given 3-day weekday context. The model's internal reasoning (Chain-of-Thought) was as follows:

Table 2. Mean Absolute Error across different models and different prediction horizons. Boldfaced values indicate the best performance for each horizon. Parentheses of Gemini indicate the performance without audio input.

	5 min	30 min	1 hour	3 hours	6 hours	12 hours	1 day	3 days	1 week	2 weeks
NaiveMean	**0.0704**	0.244	0.418	0.993	1.60	1.83	1.55	1.69	1.69	1.88
LightGBM	0.0834	**0.163**	**0.230**	**0.434**	-	-	-	-	-	-
TiDE	0.0708	0.205	0.264	0.870	**0.869**	**1.01**	**1.05**	**1.29**	**1.27**	**1.41**
TimesFM	-	-	-	-	-	-	-	-	1.37	-
Gemini	-	-	-	-	0.691 (0.730)	-	1.51 (1.43)	1.89 (2.09)	-	-

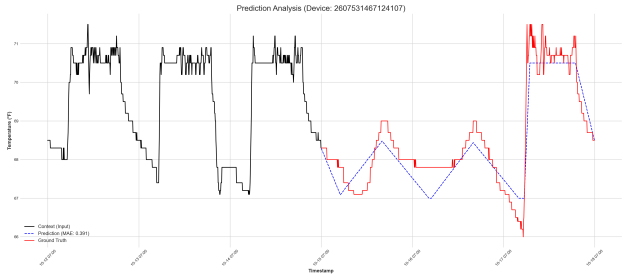

Figure 5. A three-day temperature prediction by LLM-based forecasting. The black line represents the input, the red line represents the ground truth, and the dotted blue line represents the prediction.

The forecast is based on a comprehensive analysis of the provided historical data for indoor temperature, supply air flowrate, and outside air temperature over the past three days. 1) **Weekday Pattern**: The historical data revealed a consistent pattern on weekdays, where the HVAC system maintains the indoor temperature around 70.5°F during occupied hours (approx. 07:00-19:00), followed by a gradual decrease at night, influenced by the drop in outside air temperature. 2) **Weekend Pattern**: The forecast period includes a Saturday and a Sunday. I **assumed** the building would be unoccupied on these days, leading to the HVAC system operating in a minimal or nighttime mode throughout the day. Consequently, I **predicted** that the indoor temperature would not be actively controlled and would instead slowly follow the fluctuations of the outside air temperature, rising slightly during the day and falling at night. 3) **Monday Resumption**: I **predicted** that the regular weekday HVAC pattern would resume on Monday. Specifically, a warm-up sequence would cause a sharp temperature increase in the early morning (around 05:30), stabilizing at 70.5°F during the day before declining again in the evening.

As demonstrated by its reasoning, Gemini inferred the operational status of the HVAC system, successfully predicting that the temperature would track outdoor conditions over the weekend before rising sharply on Monday morning. This result highlights the potential for LLMs to achieve forecasting performance beyond the reach of conventional methods, leveraging their advanced reasoning capabilities.

Furthermore, we hypothesized that augmenting the image-based context with an audio representation of the time series could enhance the LLM's grasp of the data's dynamic characteristics, thereby improving forecast accuracy (Fovino et al., 2024). However, our results did not yield conclusive evidence that including audio improves forecasting performance. The possibility remains that alternative sonification methods or application to longer-term forecasts (e.g., monthly predictions at hourly intervals) could prove beneficial. We leave this as a direction for future work.

5. Conclusion

This paper presented a comparative analysis of five modeling approaches for HVAC temperature prediction, evaluated on the CO-BUILD Smart Buildings Competition dataset across horizons from 5 minutes to 2 weeks.

Our results reveal that model performance is dependent on the prediction horizon:

- **5 minutes**: Naive Mean achieved the lowest error, capitalizing on the high temporal stability of indoor temperatures.
- **30 minutes to 3 hours**: LightGBM consistently outperformed other models, demonstrating its strength in short-term forecasting.
- **6 hours to 2 weeks**: TiDE proved effective for extended horizons, while the multimodal LLM exhibited unique reasoning capabilities for complex operational transitions. TimesFM and Gemini showed potential for longer-term predictions under zero-shot conditions.

This work offers practical insights for selecting appropriate

prediction models based on temporal requirements and computational constraints in smart building implementations.

References

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, Y. Chronos: Learning the language of time series, 2024. URL https://arxiv.org/abs/2403.07815.

Comanici, G., Bieber, E., and Schaekermann. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, July 2025.

Das, A., Kong, W., Leach, A., Mathur, S. K., Sen, R., and Yu, R. Long-term Forecasting with TiDE: Time-series Dense Encoder. *Transactions on Machine Learning Research*, May 2023. ISSN 2835-8856.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

Daswani, M., Bellaiche, M. M. J., Wilson, M., Ivanov, D., Papkov, M., Schnider, E., Tang, J., Lamerigts, K., Botea, G., Sanchez, M. A., Patel, Y., Prabhakara, S., Shetty, S., and Telang, U. Plots Unlock Time-Series Understanding in Multimodal Models, November 2024.

Fovino, L. G. N., Zanella, A., and Grassi, M. Evaluation of the effectiveness of sonification for time series data exploration, February 2024.

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, January 2025.

Herzen, J., Lässig, F., Piazzetta, S. G., Neuer, T., Tafti, L., Raille, G., Van Pottelbergh, T., Pasieka, M., Skrodzki, A., Huguenin, N., Dumonal, M., Kościsz, J., Bader, D., Gusset, F., Benheddi, M., Williamson, C., Kosinski, M., Petrik, M., and Grosch, G. Darts: User-friendly modern machine learning for time series. *J. Mach. Learn. Res.*, 23 (1):124:5442–124:5447, January 2022. ISSN 1532-4435.

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*, volume 30. Curran Associates, Inc., 2017.

OneBuilding. Repository of building simulation climate data from the creators of the epw, 2025. URL https://climate.onebuilding.org/. Accessed: 2025-07-10.

Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656: 5–28, 2010.

Schneider, L., Bischl, B., and Feurer, M. Overtuning in Hyperparameter Optimization, June 2025.

Smith, L. N. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, March 2017. doi: 10.1109/WACV.2017.58.

Time and Date. Past weather in san francisco, california, usa, 2025. URL https://www.timeanddate.com/weather/usa/san-francisco/historic. Accessed: 2025-07-10.