ICML 2025 CO-BUILD contest: XGBoost iterations

Vladimir Pyltsov [1]

Abstract

This document is intended to report methodology and results for the ICML 2025 CO-BUILD contest submission. The report describes the approach, including preliminary data analysis, review of the previous submissions, model selection and comparison, and discusses the findings of the results. The method approaches the contest from a data analytical perspective rather than a simulation viewpoint, treating data as a time series forecasting task. Various models are iterated, and a few data processing techniques are explored. The results show improved accuracy compared to the benchmarked value. The limitations and implications are discussed.

1. Introduction

Buildings play a crucial role in the overall energy dynamics of modern systems. According to IEA, operations of buildings make up 30% of the global total energy consumption (International Energy Agency, 2025). Hence, optimizing building load operations is crucial within the scope of increasing energy demand in the future. One of the strategies to obtain efficiency gains is performing optimal control. The popular technique, gaining traction in recent years, is reinforcement learning (RL). This approach allows for controlling devices without explicitly modeling the environment (model-free), making it flexible to be applicable and adjustable to a wide range of settings. A variety of research has been conducted applying RL strategies to building device controls, including both simulations and experimental work (Vázquez-Canteli & Nagy, 2019).

One potential enhancement of the RL approaches is the addition of forecasting. It allows for making more informed decisions by considering future outcomes. Forecasting helps in identifying the optimal sequence of actions over time to maximize cumulative rewards. The actions can be adjusted in real-time according to the predicted scenarios. The ICML 2025 CO-BUILD contest is a direct application of forecasting to building data, with the implications of potentially improving the RL controls.

1.1. Context

The data is provided by the established benchmark from (Goldfeder et al., 2024), which previously also performed RL simulations for building control. The overall context is readings from sensors and set points from the measurement and control devices. These include but are not limited to temperature, humidity, CO_2 sensors, and temperature, airflow, and fan setpoints. The building contains two floors and is separated into multiple zones, each containing a varying number of devices. The endogenous variables are temperature reading sensors, and the exogenous variables are all other variables.

The objective of the contest is to predict the temperature readings for the second half of the year 2022, given data for the first half of 2022. There are no restrictions in terms of adding external exogenous data. The contest also allows for flexibility of the prediction horizon. The models forecasting for a longer horizon with higher quality are described as being better evaluated.

1.2. Other Submissions

By the time of submissions, four submissions were available via ACM e-Energy AI DEEDS workshop (acm, 2025). The short overview is provided below, and the best result is selected as the benchmark. The comparison of the model accuracy and the selected models is presented in Table 1.

Table 1. Comparison of the other submissions (*indicates MSE).

SUBMISSION	MODEL	MAE
(KO, 2025)	LASSO REG.	1.75
(JIANG ET AL., 2025)	PI NEURALNET	5.71
(GUERRA TRIGO, 2025)	XGBOOST	1.74
(SOURIRAJAN, 2025)	SEQCAST	373.02*

Ko (2025) makes predictions over the entire time horizon using the regression models. The preprocessing steps include

[1]Department of Mechanical Engineering, Columbia University, New York, USA. Correspondence to: Vladimir Pyltsov <v.pyltsov@columbia.edu>.

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

invalid value processing, feature scaling, and the addition of temporal features. From the results, out of the three models of Lasso Regression, Ridge Regression, and XGBoost, Lasso Regression gives the best predictions. Jiang *et al.* (2025) propose to use Physics-Informed Modularized Neural Network (PI-ModNN) to conduct predictions. The authors compare the model to the LSTM and provide results over various prediction horizons. They suggest that increasing error over time might be potentially due to the missing data and outliers in July. Guerra Trigo (2025) performs analysis over the entire prediction horizon. The approach involves adding zonal and temporal features, with the best-performing model being XGBoost. Sourirajan (2025) constructs a SeqCast LSTM model and compares it to the Vanilla LSTM. The predictions were conducted for one week using one month of training data.

2. Methodology

2.1. Data

The data provided contains 123 endogenous temperature sensor readings and over 1000 exogenous variables. The frequency is 5 minutes with 51,852 time stamps in the training set and 53,292 time stamps in the testing (referred to as validation in the instructions) set.

2.2. Preprocessing

The dataset contains invalid and inconsistent units for the variables, mainly for temperature sensor readings. For the Kelvin to Fahrenheit conversion, a similar threshold as in (Ko, 2025) of 273 was applied.

The removal of invalid temperature readings was performed by fully excluding rows with data containing even one invalid sensor reading. This resulted in the reduction of the training set to 35,502 time stamps and of the testing set to 33,877 time stamps.

Several temporal features were added as exogenous variables. The features are dummy variable (or one-hot encoded) vectors representing: hour, time of the day, season, weekday/weekend, and day of the week.

2.3. Models

As the XGBoost (Chen & Guestrin, 2016) model performed the best in previous results, the baseline model is the XGBoost model, which is optimized with 3 Optuna search trials.

The other choice of model iterations is a binned architecture. The choice is the binned XGBoost models. The idea is the following. Let:

- $y_{t,b}$ represent endogenous variable at time t and bin b
- $X_{t,b}$ represent exogenous features at time t and bin b
- $f_b(\cdot)$ represents function trained for bin b

The model is trained on the training set for each bin:

$$y_{t,b} = f_b(X_{t,b}) + \varepsilon_{t,b} \tag{1}$$

where $\varepsilon_{t,b}$ is the residual error.

After the models are trained on separate bins, the predictions are stacked to form a full prediction sequence:

$$\hat{Y}_t = \{\hat{y}_{t,0}, ..., \hat{y}_{t,b}\} \tag{2}$$

where $\hat{y}_{t,b} = f_b(X_{t,b})$.

Firstly, the following approach does not restrict the training function specifically to XGBoost architecture: the models could range from simple OLS models to much more complex ML architectures. In this case, the model choice is motivated by multiple factors. XGBoost models effectively capture non-linear relationships; at the same time, the models do not require overwhelming feature handling, such as scaling. The simple architecture also allows for quicker training with easier hyperparameter handling. Within the context of the contest with a large dataset, both in terms of exogenous features and timestamps and variously scaled features, XGBoost models are an attractive choice.

Secondly, the approach allows for the bin selection. The overall idea of binning is to isolate the temporal patterns through independent model training. Instead of excessively including temporal features or time series decomposition in the architecture, the models "focus" on the exogenous variable dependencies and their relationship across the overall data length. Intuitively, more bins should allow for more flexibility and parameter optimization, but can come at excessive computational cost. For the contest domain, the choice for the majority of the iterated models is hourly bins for the computational considerations (for one iteration, the bin is 15 minutes). The baseline XGBoost model is indicated as 1 bin in the results.

Lastly, the architecture allows for a wider parameter tuning through focused hyperparameter optimization for a specific bin. However, this can also cause significant computational costs in the context of the contest with a large dataset. For faster training, 3 to 5 Optuna search trials were chosen.

2.4. Dimensionality

The dataset contains a vast amount of exogenous features. Reducing their input amount can help reduce computational intensity. Nevertheless, transformation of the exogenous

features requires careful handling in order to capture both spatial (in this case, zonal) and temporal variations.

The choice of the studied approach is simple PCA component analysis. The method is straightforward by retaining a few of the components, which explain the majority of the variation, as the exogenous variables. The PCA transformers are fitted on the training data and are then applied to both training and test data. Multiple study cases are created: 1) no grouping (N) - the PCA is applied to the whole exogenous variable matrix (iterations of various number of components were performed); 2) variable type grouping (V) - the exogenous variables are grouped into categories, and PCA of 15 components is applied separately to each group; 3) variable type and zonal grouping (VZ) - the exogenous variables are grouped into categories and zones, and PCA of 3 components is applied separately to each group. The categories for the variable types are air temperatures, water temperatures, setpoints, control commands, environmental, flow pressure, and other. The variables are categorized into those groups according to the keywords in the variable name. The zonal categorization is performed based on the zonal metadata information. The full exogenous input without PCA processing is indicated as (-).

2.5. Metrics

The chosen focus is long-term horizon predictions. Hence, the benchmark is the best-achieved accuracy of $1.74°F$ MAE. Additionally, the accuracy of the same model was calculated for other ranges: 1 week, 2 weeks, 1 month, 3 months, and the full period.

3. Results

3.1. Main Results

The results for the entire period are presented in Table 2. For the no-grouping PCA case, the iteration with the best results of 100 components is presented. The other iterations with 5, 20, 50 components for that case gave MAE of 2.12, 2.09, 2.07, respectively.

Table 2. MAE Results (entire period) .

BINS	PCA	# OF FEATURES	MAE
1	-	1092	1.17
HOURLY	-	1092	1.29
HOURLY	N	117	1.98
HOURLY	V	119	1.79
HOURLY	VZ	809	1.23
15-MINS	VZ	809	1.33

The best result with the lowest MAE is the baseline XGBoost model of a single bin. The second-best result is the VG case of hourly bins. The result represents the improvement of more than 0.5 $°F$ MAE compared to the benchmarked result. There are a few notable observations. Firstly, the results are better for the hourly model with the variable and zonal PCA grouping compared to the one with the full feature input. There are potentially two explanations for this. One of them is the stochastic nature of the hyperparameter optimization. The other reason is the potential positive impact of PCA processing, which removed some of the noisy and irrelevant features despite the retention of a large number of components. Secondly, the binned models perform worse than the baseline model. This suggests that isolating temporal behavior does not provide prediction improvements. Thirdly, the variable and zonal grouping hourly model has better results than the models with less robust PCA approaches. The variable grouping performs better than the no-grouping approach, but provides fewer overall improvements. This suggests that zonal variation plays a significant role in quantifying the overall predictions.

The results for the three best-performing models across different periods are provided in Table 3. The predictions of the hourly binned PCA-processed model have lower error across all selected periods compared to those of the non-processed model. This potentially supports the hypothesis that PCA removes some of the noise features, making the performance of the models more robust.

Table 3. MAE test results by period.

PERIOD	HOURLY (-)	HOURLY (VZ)	1 (-)
1 WEEK	1.42	1.20	0.97
2 WEEKS	1.42	1.20	0.97
1 MONTH	1.79	1.38	1.06
3 MONTHS	1.34	1.23	1.09
ENTIRE PERIOD	1.29	1.23	1.17

The first month predictions have the highest MAE. This can potentially be explained by the presence of missing values in the month of July. Because of a smaller number of values, a few inaccurate predictions can more significantly contribute to the overall error.

3.2. Detailed Comparison

The section is primarily focused on a more close-up comparison of the Hourly (VZ) and 1 (-) models.

For the Hourly (VZ) model, the MAE of the best-performing sensor is 0.67 $°F$, and the MAE of the worst-performing sensor is 2.45 $°F$. The plots for the predictions across different periods of the best-performing sensor are depicted in Figures 1,3, and 5. For the 1 (-) model, the MAE of the best-performing sensor is 0.47 $°F$, and the MAE of the worst-performing sensor is 2.80 $°F$. The plots for the

predictions across different periods of the best-performing sensor are depicted in Figures 2,4, and 6.

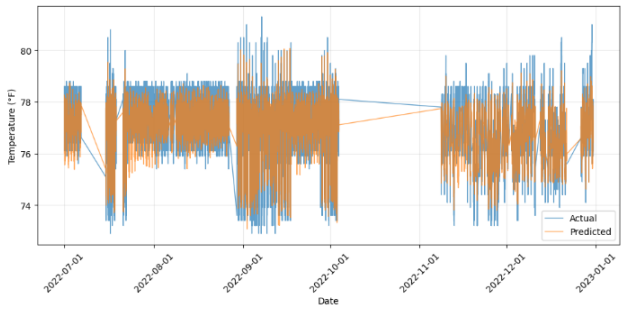

Figure 1. The actual and predicted values for the best-performing sensor for the entire prediction horizon - Hourly (VZ) case.

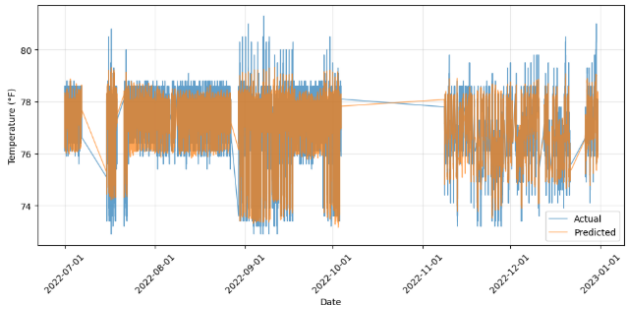

Figure 2. The actual and predicted values for the best-performing sensor for the entire prediction horizon - 1 (-) case.

The overall pattern of the predictions, depicted in Figures 1 and 2, is satisfactory. There are no predictions that violate any physical range, and the predictions retain robustness across the entire period. The notable observation is that the hourly model puts more emphasis on the spikes and drops; the predictions align with that behavior, for instance, in the month of September. The baseline model tends to be much more conservative in terms of the prediction of range. The difference can be seen in the period of late July and early August. The hourly model overestimates the drops, potentially causing large errors in predictions; the baseline model does not estimate the drops, yielding higher accuracy.

Figures 5 and 6 reveal granular patterns of each model's predictions for the first week. A cyclical pattern can be recognized in the hourly model predicted values. The pattern follows a daily pattern with values dropping in the night period and ramping in the morning. This pattern makes sense as temperature goes up during the day and drops during the night, and it was potentially captured in the model training. Nevertheless, the actual reading does not follow that pattern exactly. This mismatch creates significant errors in predictions (like, for instance, July 2 - the predicted values drop during the night, however, the actual values do not). The baseline model makes more stochastic predictions with a less prominent cyclical trend.

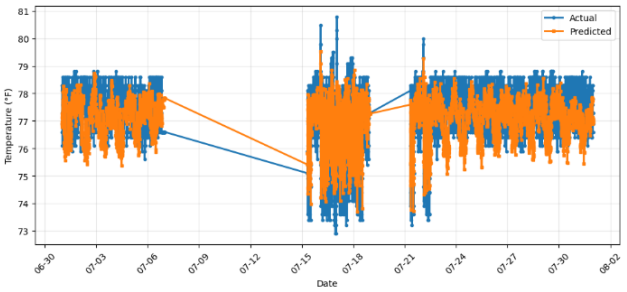

Figure 3. The actual and predicted values for the best-performing sensor for the first month of the prediction horizon - Hourly (VZ) case.

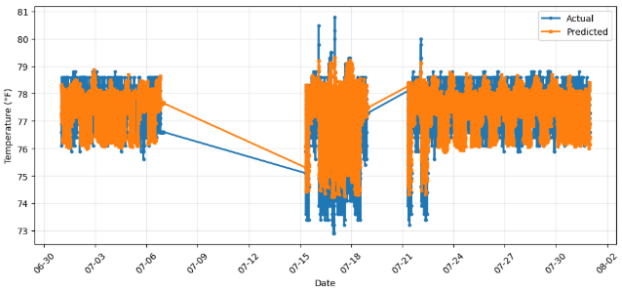

Figure 4. The actual and predicted values for the best-performing sensor for the first month of the prediction horizon - 1 (-) case.

The MAE for the best-performing sensor of the hourly architecture across all months is provided in Table 4. The results show that the error is the highest in the month of July, aligning with the previous result of the highest error for the first month of predictions.

Table 4. MAE of the best-performing sensor for different months - Hourly (VZ) case.

MONTH	TIMESTAMPS	MAE
JULY	6428	0.765
AUGUST	7528	0.633
SEPTEMBER	8547	0.732
OCTOBER	864	0.614
NOVEMBER	4231	0.633
DECEMBER	6279	0.543

The lowest number of timestamps in the month of October is also evident in Figure 1. The lowest error is observed for the month of December.

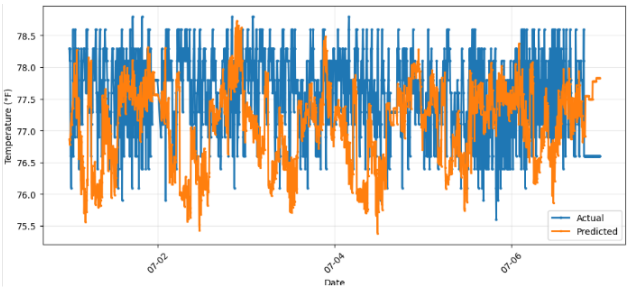

Figure 5. The actual and predicted values for the best-performing sensor for the first week of the prediction horizon - Hourly (VZ) case.

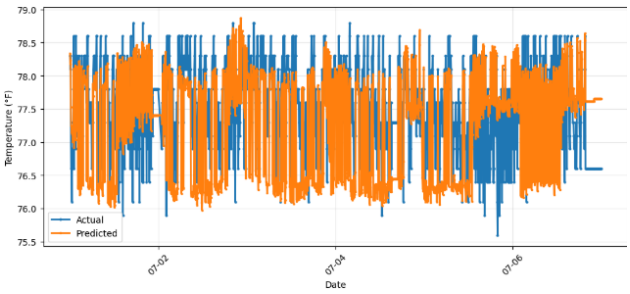

Figure 6. The actual and predicted values for the best-performing sensor for the first week of the prediction horizon - 1 (-) case.

The distribution of the prediction errors is depicted in Figures 7 and 8. The median is slightly lower than the mean for both cases.

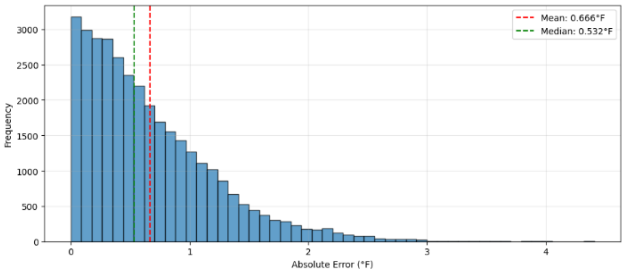

Figure 7. The distribution of prediction errors for the best-performing sensor - Hourly (VZ) case.

Results and discussion on hourly model training and an additional model are provided in the Appendix.

4. Discussion

The best MAE results are lower than the established benchmark. Nevertheless, it is not clear whether the overall predictions of either approach would be satisfactory for the RL tasks. For hourly models, the mismatch of the ramping and dropping of the readings can potentially create false signals for the actions of the RL algorithm. For instance, the downfall of the temperature on July 2 is predicted around the afternoon. This could be treated as a signal to reduce the airflow into the room, since additional cooling is not anticipated. However, the actual values stay the same, requiring the airflow to ramp up. This can potentially reduce the efficiency of the operations. The baseline model provides more stochastic predictions. However, whether the model would provide better signals for control mechanisms is unclear.

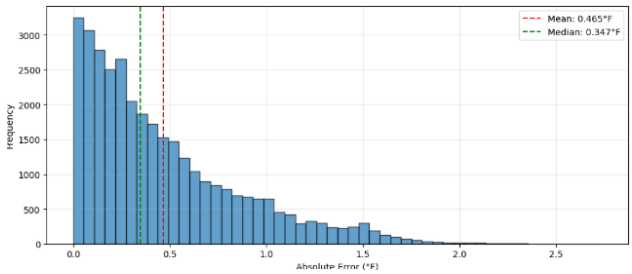

Figure 8. The distribution of prediction errors for the best-performing sensor - 1 (-) case.

The data processing techniques become one of the crucial steps in the approach. The choice of handling zeros/missing values can potentially alter the results. This work uses a straightforward approach of removing the entire rows of data with invalid readings. However, other approaches could be considered. One of the potential ideas could be backward, forward, and mean-value filling. The indicators of the value filling could potentially provide helpful input for the models to treat those values or even sensors differently. The issue is the mismatch between the missing values in the training and validation sets; hence, the creation of individual vector indicators can be challenging, as certain sensors might have missing values in the training but not the validation set. The aggregated vectors are likely to lose some important information, while creating vectors for all sensors can become computationally intense. The handling of exogenous variables is also significant. The reduction in dimensions while preserving all the information is challenging. Several iterations of this work show that retaining spatial and variable type variations is crucial. Hence, the techniques, which handle variations across multiple dimensions, need to be considered.

The baseline approach is suitable for rapid estimations and benchmarks. XGBoost model, being a decision tree-based approach, is flexible in terms of feature handling and is suitable for retaining long-term predictions without collapsing. If some cyclical behavior is observed, the binning technique

can potentially help capture some of the temporal patterns. Neither of the models formulates the physics explicitly, relying on the provided data of historical observations.

5. Conclusion

This work iterates through different models and data processing techniques to forecast temperatures for the CO-BUILD contest. The main observations are the following: 1) a simple XGBoost model can provide robust long-term predictions, which can be used as benchmarked values for other techniques; 2) exogenous variable processing should take into account both spatial and variable type variations; 3) binned models do not provide better forecasts within the scope of the contest data. Future work could include more robust exogenous feature handling, more model iterations, and comparison with a physics-informed model.

Software and Data

The code is available via https://github.com/starship204/ICML2025-COBUILD-contest.

Impact Statement

The forecasting in the energy sector, including building modeling, is of crucial importance for the optimal operation of energy systems. Making accurate predictions can help optimize the control of the building device and the management of larger systems. This can provide economic and resource efficiency gains with a broad societal impact. The selected models show robust forecasts over the prediction horizons for the contest domain. At the same time, the binned models have not received much traction in theoretical or practical research. The idea of isolating temporal variations through binning can potentially be applied on a broader scope. Hence, there is also an opportunity for impact in the broader ML community by investigating the applicability of the approach for time-series forecasting beyond the energy domain.

References

Proceedings of the 16th ACM International Conference on Future Energy Systems, Rotterdam, Netherlands, 2025. Association for Computing Machinery. URL https://dl.acm.org/doi/proceedings/10.1145/3679240. Accessed: 2025-07-07.

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

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, 2024.

Guerra Trigo, G. Predicting building zone air temperatures using xgboost and feature engineering: A smart buildings challenge submission. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*, pp. 941–943, 2025.

International Energy Agency. Buildings – energy system, 2025. URL https://www.iea.org/energy-system/buildings. Accessed: 2025-07-07.

Jiang, Z., Wang, X., and Dong, B. Physics-informed modularized neural networks for building dynamic modeling: A smart buildings hackathon case study. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*, pp. 924–927, 2025.

Ko, E. Temperature prediction with feature engineering and multiple regression: Smart buildings hackathon submission. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*, pp. 921–923, 2025.

Sourirajan, V. Seqcast: Sequence-based forecasting of temperature distribution in smart buildings. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*, pp. 959–961, 2025.

Vázquez-Canteli, J. R. and Nagy, Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied energy*, 235:1072–1089, 2019.

A. Model training

The training error by hourly bin for two best performing models is provided below.

Table 5. MAE training results by bin.

BIN	HOURLY (-)	HOURLY (VZ)
HOUR 0	1.25	1.29
HOUR 1	1.33	1.47
HOUR 2	1.37	1.38
HOUR 3	1.60	1.68
HOUR 4	1.58	1.63
HOUR 5	1.50	1.50
HOUR 6	1.55	1.58
HOUR 7	1.57	1.54
HOUR 8	1.63	1.65
HOUR 9	1.65	1.73
HOUR 10	1.75	1.86
HOUR 11	1.96	1.99
HOUR 12	1.89	1.76
HOUR 13	1.56	1.75
HOUR 14	1.35	1.34
HOUR 15	1.14	1.15
HOUR 16	1.21	1.21
HOUR 17	1.24	1.22
HOUR 18	1.31	1.29
HOUR 19	1.28	1.34
HOUR 20	1.38	1.33
HOUR 21	1.52	1.31
HOUR 22	1.46	1.34
HOUR 23	1.29	1.41

The models are mainly similar in terms of training error. The slight difference is in training for the evening and night hours. For the evening hours, the PCA processed model has a slightly lower error; for the night hours, the non-processed model has a slightly lower error. Moreover, for both models, predicting the temperatures around midday is the most challenging period. Within the scope of the building dynamics, this aligns with the lunch period in terms of occupant dynamics or the period with the highest sun altitude in terms of solar radiation exposure. However, the connection is not entirely clear and would require a more robust investigation.

The overall observation is that the training error for each of the models is larger than the overall error for the testing period. A possible explanation is that the training set potentially has a larger overall variation, making training challenging. Nevertheless, a rather good performance of models on the testing set suggests the flexibility and robustness of the approach.

B. Additional model and plots

An additional model with more temporal features was run. The additional features included temporal cyclical encoding, various indicators of on/off set points through keyword search and categorization. The total amount of exogenous input was 1152 features. The results did not yield improvements. They are provided in the table below.

Table 6. Additional model MAE test results by period.

PERIOD	1 (-)
1 WEEK	1.08
2 WEEKS	1.08
1 MONTH	1.38
3 MONTHS	1.19
ENTIRE PERIOD	1.18

The additional plots are regarding the best-performing sensors. The majority of the iterations revealed the best-performing sensor under the ID of 16286830034440683520. However, the simple explanation of why it is the best-performing sensor is potentially constant reading fluctuations. Below are plots of observations of different sensors (the plots do not have the rigorous implementation of the timestamp handling, i.e., the predictions were just plotted continuously without the gaps with the removed rows).

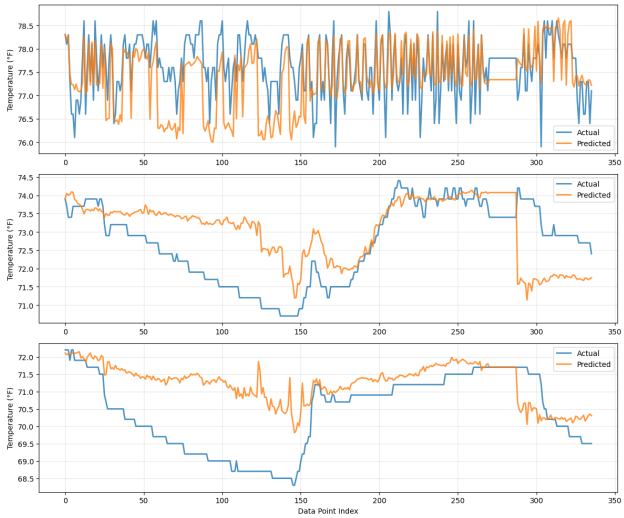

Figure 9. The actual and predicted values for the first two weeks. The top plot is the best-performing sensor, the middle plot is the worst-performing sensor, and the bottom plot is a random sensor.

From Figure 9, it can be seen that the top plot best-performing sensor has continuous fluctuations, while the other two sensors have slower variations. Figure 10 reveals the idea further. For the best-performing sensor, the fluctuations stay almost the same over the entire prediction period. However, for the other two sensors, the variation range is much higher, especially with a significant drop towards the end (which is winter months - this makes sense as the

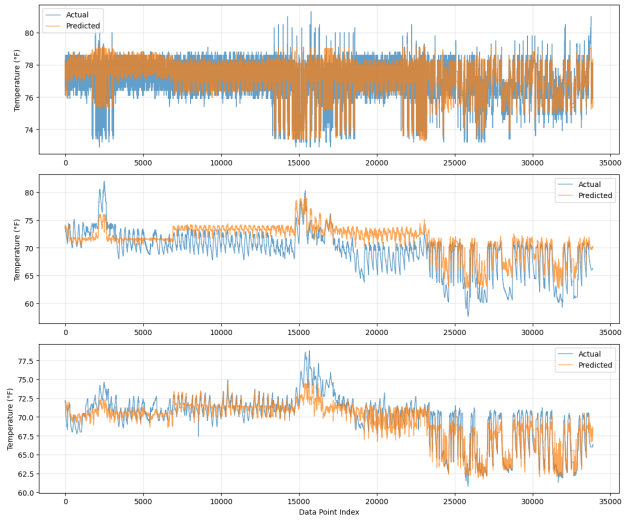

Figure 10. The actual and predicted values for the entire period. The top plot is the best-performing sensor, the middle plot is the worst-performing sensor, and the bottom plot is a random sensor.

temperatures go down).

Figure 10 also shows how the worst-performing predictions overestimate the sensor fluctuations across summer and fall months. For the winter months, the predictions align much better. The bottom plot, which depicts a random sensor, shows an overall satisfying trend of the predictions. The exception is the sudden surge in temperatures in the middle.