XGBoost with physics-informed features and residual regressor for the SBCS benchmark

Vladimir Pyltsov
Department of Mechanical Engineering
Columbia University
v.pyltsov@columbia.edu

Abstract

This study proposes an enhanced XGBoost model with physics-informed features and a residual regressor. The model is used in the long-term forecasting setting of the Smart Buildings Control Suite (SBCS) benchmark, with a context of building energy dataset with a large exogenous matrix and similar lengths of training and test sets. The results show improvements over more than 10% across the majority of the selected horizons compared to the baseline XGBoost model without any modifications. A notable error improvement includes the horizon of the full test set. The proposed model can be used as an initial step towards further advancements in the capabilities of tree-based models in long-term forecasting and building energy setting.

1 Introduction

The operation of building accounts for 30% of the global energy consumption [Ali et al., 2024]. This implies that improving the efficiency of the building management systems can result in significant energy savings. A large variety of research has been applied across academic field exploring optimal control strategies to improve the building energy dynamics. Recent advances include complex models of Model Predictive Control (MPC), Reinforcement Learning (RL), and Deep Learning (DL) [Arun et al., 2024, Nguyen et al., 2024, Stoffel et al., 2024].

One of the ways to improve control strategies is the addition of forecasting techniques. Predicting the values based on the taken strategy can potentially shift the dynamics of actions in order to maximize the overall efficiency in the long run. This work explores such a setting, the forecasting model for which is developed and experimented.

1.1 Context

The setting is provided by the Smart Buildings Control Suite (SBCS) benchmark [Goldfeder et al., 2024] as a part of *UrbanAI 2025 Contest*. The dataset contains control and observational variables of the building devices for year 2022. The control variables, such as temperature, supply, and damper set points, are considered exogenous and available ahead of time. The observational variables of temperature sensors in building rooms are considered endogenous. The contest splits the first half of 2022 as the training set and the second half of 2022 as testing. Using only exogenous variables (i.e., endogenous variables, their lags or moving averages, are not available in the test set), the goal is to predict the readings of the temperature sensors.

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

1.2 Related work

By the time of this study submission, multiple works proposed various models for the setting. The submissions are available via *ACM e-Energy AI DEEDS Workshop* [SIG, 2025] and *ICML 2025 CO-BUILD Workshop* [Goldfeder et al., 2025]. A short summary of the works is provided in Table 1.

Table 1: Summary of the proposed models and results for the Co-build benchmark

Work	Model	Test Horizon	Result (MAE)
Ko [2025]	Lasso Regression	-	1.75
Jiang et al. [2025]	PI-ModNN	Full Period	5.71°C
Guerra Trigo [2025]	XGBoost	-	1.74
He and Guo [2025]	Soft-MoE	Full Period	1.18
Sun et al. [2025]	XGBoost	2 Weeks	4.20
Pyltsov [2025]	XGBoost	1 Week to Full Period	1.17
Arisaka et al. [2025]	TiDE	5 min to 2 Weeks	1.41
Saha and Shinde [2025]	XGBoost	1 Day to 4 Weeks	1.79
Neogi [2025]	Two-layer LSTM	1 Day to Full Period	4-20°C
Gokhman [2025]	DLinear	10,563 test samples	0.22
Sourirajan [2025]	Transformer CL	1 Week	1.61

Table 1 provides a summary using the following assumptions:

- (-) in the "Horizon" column indicates that the horizon was not explicitly stated. The assumption is that the full period was tested.
- MAE results are assumed to be degrees Fahrenheit unless otherwise specified.
- If multiple models were tested, the best-performing model is reported.
- If multiple horizons were tested, the result for the longest period is reported.

It can be seen that the majority of the tested and explored models are tree-based, showing excellent performance over short- and long-term horizons. This can be explained as the setting has an overall regressive nature. With a large exogenous matrix, the absence of a sliding window and lagged variables, the setting can be quantified as $y = f(X)$ rather than a more typical time-series setting of $y_t = f(y_{t-1}, ..., y_{t-b})$. Although regression models can offer a solution, they have challenges in quantifying nonlinear relationships and computing large numbers of variables. Hence, tree-based models provide a computationally efficient way to find relationships with a large exogenous matrix by recursively partitioning the feature space. This makes them an excellent candidate for high-dimensional data settings such as building device measurements. Notably, the summary indicates that even the same tree-based models show different results across the dataset. The discrepancy can potentially be explained by different data processing techniques and challenges (while periods might be the same or similar, the number of test samples might not necessarily match) and hyperparameter choice.

This work builds on the findings of previous studies to improve long-term forecasting. In particular, the key idea is to build on the tree-based model with additional feature engineering and the addition of a residual regressor. The main objective of the study is the following : **to improve the long-term horizon benchmark by modifying a tree-based baseline model**. The code is available in the following repository: `https://github.com/starship204/Urban-AI-2025-Contest`.

2 Methodology

2.1 Data & Processing

The data contains 51,852 samples in a training set and 53,292 samples in a test set. The measurements are at a 5-minute frequency for the entire year 2022. The endogenous matrix contains invalid and mismatched unit readings. The processing involved eliminating all rows of the entire data matrix, which contained invalid temperature values. This was done because for most of the timestamps, the invalid readings were present for all sensors. At the same time, certain periods of invalid readings were relatively long, making interpolation potentially challenging. After the elimination procedure, the training set was reduced to 35,502 samples, and the test set was reduced to 33,877 samples. For the varying units, a threshold of 273 was applied: if the temperature reading exceeded the value of 273, the conversion from units of Kelvin was applied. The conversion was applied to all variables that contained the string 'temperature' in the name key word.

2.2 Feature Engineering

2.2.1 Temporal Features

Temporal features were created as variables that are known in advance. They included a one-hot encoded vector for the time of the day (6-12 is morning, 12-18 is day, 18-22 is evening, and other hours are night), season, hour of the day, weekend indicator, and day of the week. This allows the model to have useful indicators to capture temporal patterns of the dynamics.

2.2.2 Physics-Informed Features

Various exogenous variables were manipulated to create physics-informed features. The list with pseudo-formulas and a short description is provided below:

- Setpoint spread - difference between the cooling setpoint and the heating setpoint. It can be written as `setpoint_spread = cooling_sp - heating_sp`.
- Control effort - defined as the average of the supply air damper percentage command and the heating water valve percentage command. It can be written as `control_effort = (valve_pos + damper_pos) / 2`.
- Thermal effectiveness - a proxy indicator of a relation between discharge temperature and temperature setpoints. For cooling effectiveness, the formula is `max(0, cooling_sp - discharge_temp) / max(1, cooling_sp - 50)` and for heating effectiveness is `max(0, discharge_temp - heating_sp) / max(1, 80 - heating_sp)`. The values of 50 and 80 for each of the formulas were chosen as approximate indicators as reference temperatures.
- Flow-normalized control - control effect multiplied by the flow rate. It can be written as `flow_normalized_control = control_effort * flow_rate / 1000`.
- Temporal momentum features - difference of the heating and cooling setpoints and flow rate between the observed value at `t` and the value at `t-1`. This feature is based on the following **assumption**: the exogenous variables at previous time stamps are available at the prediction time stamp. The features can be written as `cooling_sp_change = cooling_sp - prev_cooling_sp`, `heating_sp_change = heating_sp - prev_heating_sp`, and `flow_change = flow_rate - prev_flow`.
- Neighbor features - difference of the cooling and heating setpoints and the flow rate between the observation value and the mean of the values of adjacent rooms. This can be written as `cooling_sp_deviation = cooling_sp - avg_neighbor_cooling`, `heating_sp_deviation = heating_sp - avg_neighbor_heating`, and `flow_deviation = flow_rate - avg_neighbor_flow`. The adjacent spaces were determined through the floor plan and can be seen on Figure 1.

2.3 Model

The initial model is a simple XGBoost model [Chen and Guestrin, 2016], which is trained on a full exogenous matrix with the addition of temporal and physics-informed features.

2.3.1 Self-correction

After the initial predictions are made in the training phase, they are self-corrected before training the residual regressor. The main idea is that the predictions cannot deviate temporally and spatially by a large amount.

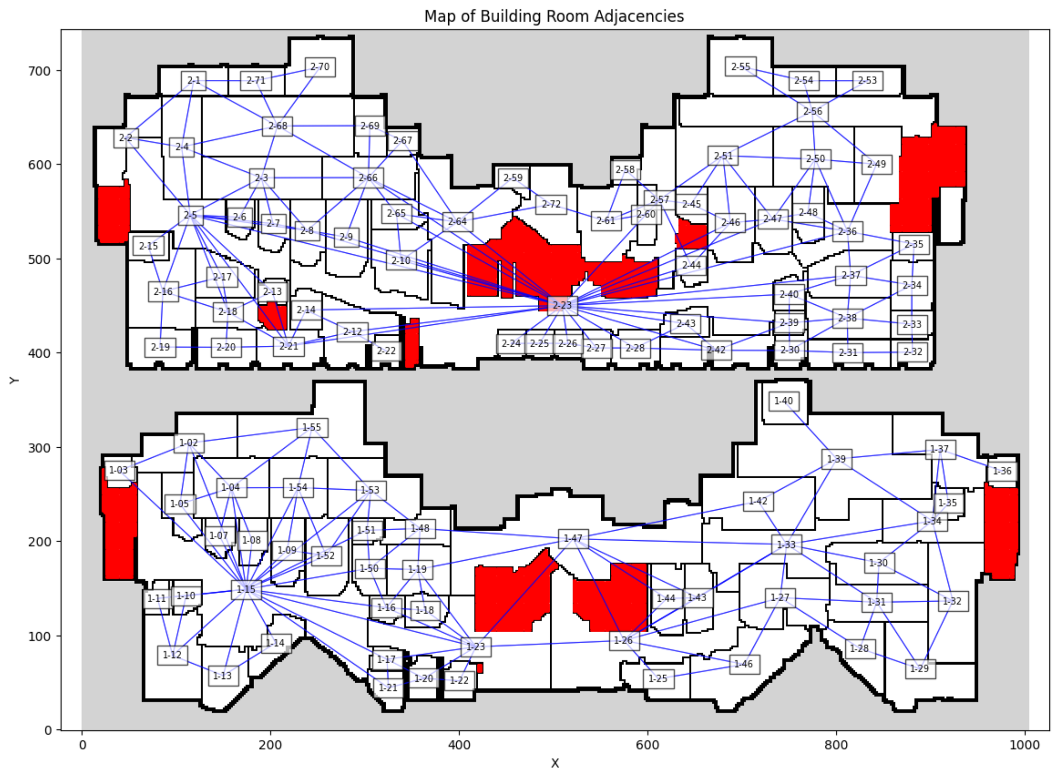

Figure 1: Floor plan with connected rooms.

Temporal Correction:

$$\hat{y}_{i,t}^{\text{corrected}} = \begin{cases} 0.7 \times \hat{y}_{i,t-1} + 0.3 \times \hat{y}_{i,t} & \text{if } |\hat{y}_{i,t} - \hat{y}_{i,t-1}| > 2.0 \\ \hat{y}_{i,t} & \text{otherwise} \end{cases} \tag{1}$$

Spatial Correction:

$$\hat{y}_{i,t}^{\text{corrected}} = \begin{cases} 0.7 \times \hat{y}_{i,t} + 0.3 \times \bar{y}_{\mathcal{N}_i,t} & \text{if } |\hat{y}_{i,t} - \bar{y}_{\mathcal{N}_i,t}| > 5.0 \\ \hat{y}_{i,t} & \text{otherwise} \end{cases} \tag{2}$$

where $\bar{y}_{\mathcal{N}_i,t}$ denotes the mean of the neighboring room predictions. The weight coefficients of 0.7 and 0.3 were chosen arbitrarily.

Corrections are applied for the residual regressor to learn the remaining unexplained potential signal. Hence, the chosen bounds are rather conservative.

2.3.2 Residual regressor

An additional XGBoost model is then trained on the residuals, which is the difference between the actual values and the corrected predictions. The final predictions can then be written as:

$$\hat{\mathbf{y}}_{\text{final}} = \hat{\mathbf{y}}_{\text{initial}} + 0.5 * \hat{\mathbf{y}}_{\text{residual}} \tag{3}$$

The coefficient 0.5 was also arbitrarily chosen.

2.3.3 Training

The initial XGBoost model is trained with 20 Optuna hyperparameter trials. The residual regressor has fixed parameters with no hyperparameter optimization. For comparison, a baseline XGBoost

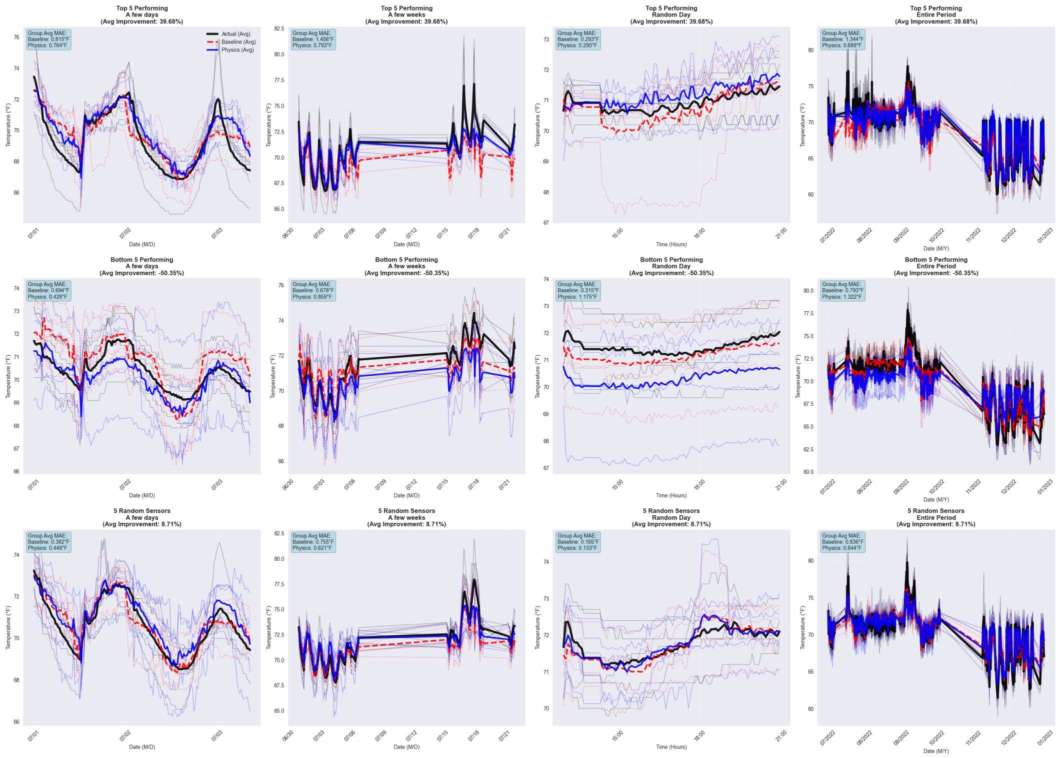

Figure 2: Predictions for various horizons and sensor groups.

model was trained on the exogenous matrix without additional features and a residual regressor. Limitations and potential enhancements are described in the Discussion 4 section. The experiments were run with an NVIDIA GeForce RTX 3070 GPU.

3 Results

3.1 Main results

The MAE results are presented over horizons of the remaining clean dataset (i.e., 2 weeks is not July 1 to July 14, but rather the first 2 weeks of the encountered points in the modified dataset).

Table 2: Main results (MAE)

Horizon	Baseline	Modified	Improvement
1 Week	1.0725	0.9516	11.27%
2 Weeks	1.2657	1.1019	12.94%
1 Month	1.1453	1.0108	11.74%
3 Months	1.1599	1.0482	9.63%
Full Period	1.2086	1.1062	8.47%

The main results are presented in Table 2. The modified model shows improvements in predictions on all horizons, with improvements of more than 10% compared to the XGBoost baseline model for several horizons. The full period result presents the highest accuracy compared to all previous tree-based attempts, improving the benchmark by 5.75% (i.e., 1.1062 MAE compared to 1.1737).

3.2 Overall dynamics

Several visual examples of model prediction and ground truth are presented in Figure 2.

Overall, both models provide visually satisfactory predictions with rather rare jumps and accurate following of both daily and seasonal patterns. Some quick summary statistics are provided in Table 3

Table 3: Summary statistics

Description	Statistic
Number of improved sensors	92 (74.8%)
Number of degraded sensors	31 (25.2%)
Best sensor improvement	42.28%
Worst sensor improvement	-87.38%
Baseline model second stat. moment of MAE	0.3501
Modified model second stat. moment of MAE	0.3466

Most of the sensors have improvements in terms of prediction values. A notable observation is that the best improvement is twice as small as the worst improvement. This implies that for several cases, the baseline model provides much better estimates of exogenous relationships. In terms of standard deviation, both models have an almost identical spread in terms of predictions.

4 Discussion

4.1 Ensemble

One of the potential improvements could be the ensemble of both the baseline and the modified method. As shown in Table 3, some of the sensors have poorer predictions with a modified model. Additional motivation for the idea can also be seen in Figure 3.

It can be seen that for most periods, the modified model performs better. The baseline model performs better in the night period of August, the morning period of July, the evening period of August, and the morning period of December. Hence, developing an ensemble model that can potentially use a better-performing model during certain periods can make the predictions better.

Two simple cases are developed to assess the potential impact. The first case is an oracle scenario, which chooses the best prediction (between baseline and modified models) for every individual sample with perfect hindsight. The second scenario is a simple averaging between predictions. This is one of the most common ensemble methods. The results are presented in Table 4.

Table 4: Results for ensemble scenarios (MAE)

Horizon	Baseline	Modified	Oracle	Ensemble (Averaging)
1 Week	1.0725	0.9516	0.6725	0.9132
2 Weeks	1.2657	1.1019	0.8356	1.1040
1 Month	1.1453	1.0108	0.7432	0.9957
3 Months	1.1599	1.0482	0.7951	1.0310
Full Period	1.2086	1.1062	0.8392	1.0915

It can be seen that selecting a model with the best prediction (oracle scenario) can potentially significantly reduce the error. The challenge is to develop a model selection strategy. The selection has a high decision-making span in terms of spatial and temporal dimensions. Spatially, every sensor has its own model decision. From Figure 3, it is evident that the individual curves of both the baseline and modified model predictions can be lower than average. Temporally, models can have different performance even within the same time window. For instance, in the August evening time window, the modified model has a better performance on average at the beginning of the window, but has a worse performance later on. This situation could also be common for individual sensors. A simple ensemble method provides further improvements to the results achieved by the modified model.

4.2 Limitations & Future Work

The main limitation of this work is the overall choice of the residual regressor and self-correction after training. Although the main idea is to learn the true residuals, self-correction can potentially be

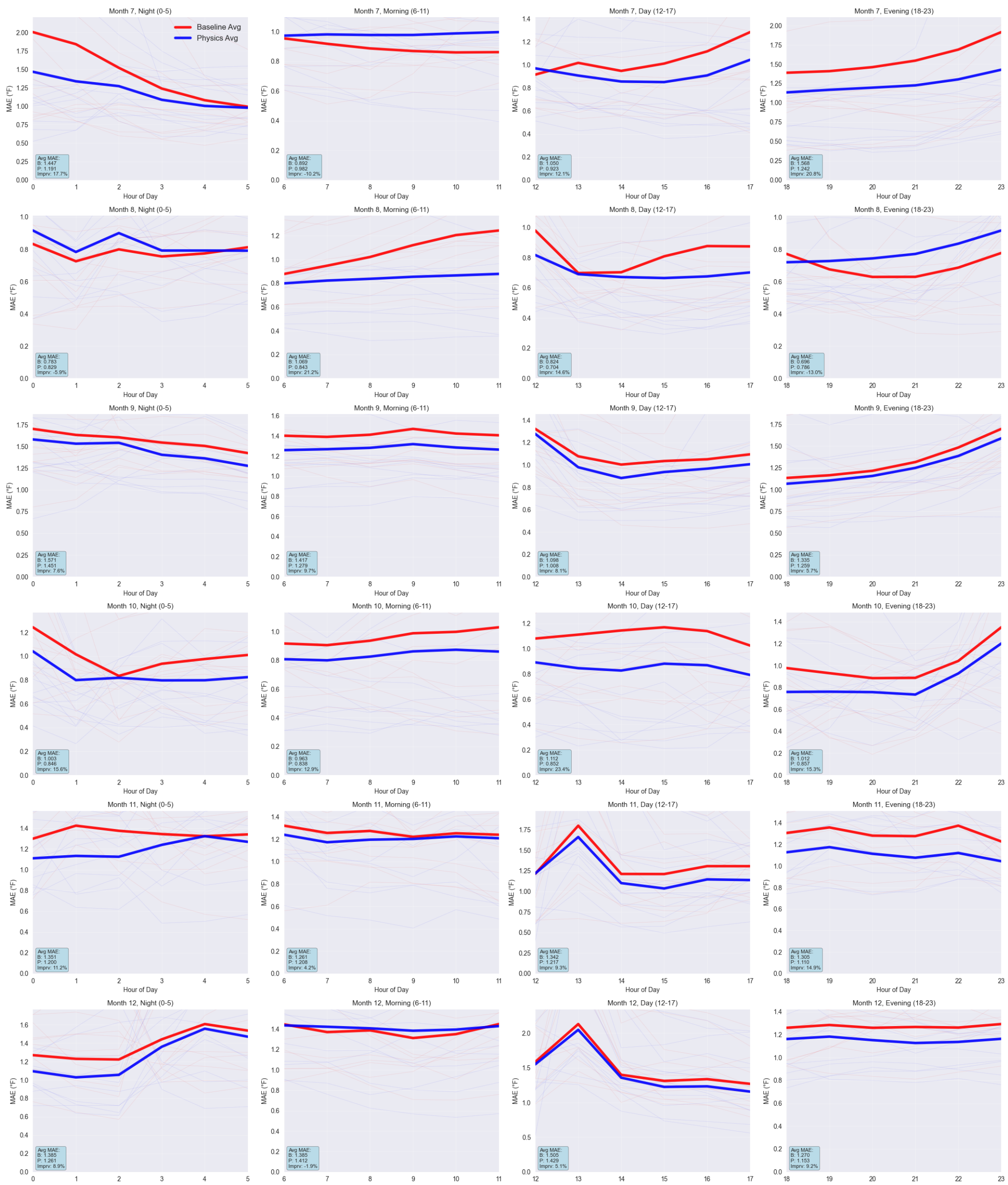

Figure 3: MAE of models for different periods.

eliminated to learn the initial model's overall error or embedded in the training or prediction process of the main model. The self-correction and residual prediction weight coefficients were not optimized. The impact of a different choice of the residual regressor model can also be explored and tested.

Lastly, as discussed in the Discussion 4 section, the main vision is to explore the ensemble models for the setting. It is evident that substantial improvements can be made by looking at individual sensor dynamics and more granular temporal deviations. While a simple averaging strategy improves the predictions, more complex strategies can bridge the gap with the oracle scenario benchmark. Another potential avenue could be to develop individual models for each of the sensors. The concern would be computational expenses and the architecture complexity.

5 Conclusion

This work proposes a modified XGBoost model with physics-informed features and a residual regressor. The results show improvements compared with the baseline model and previously achieved results. The study also takes a closer look at the overall prediction dynamics. The results reveal that the modified model improves the predictions for most sensors, but not all. Moreover, the baseline model also achieved a better performance during certain time windows on average for all sensors. This suggests that future improvements can explore ensembling approaches, which capture the individual sensor and temporal dynamics more robustly.

References

Usman Ali, Sobia Bano, Mohammad Haris Shamsi, Divyanshu Sood, Cathal Hoare, Wangda Zuo, Neil Hewitt, and James O'Donnell. Urban building energy performance prediction and retrofit analysis using data-driven machine learning approach. *Energy and Buildings*, 303:113768, 2024.

Sohei Arisaka, Eikichi Ono, Hiroyasu Miura, Yutaka Shoji, Yangayang Li, and Kuniaki Mihara. Co-build smart buildings competition: An empirical comparison of hvac temperature prediction models. In *ICML 2025 CO-BUILD Workshop on Computational Optimization of Buildings*, 2025.

M Arun, Gokul Gopan, Savithiri Vembu, Dilber Uzun Ozsahin, Hijaz Ahmad, and Maged F Alotaibi. Internet of things and deep learning-enhanced monitoring for energy efficiency in older buildings. *Case studies in thermal engineering*, 61:104867, 2024.

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

Ruslan Gokhman. Forecasting building temperature time series with exogenous variables: Icml co-build challenge. In *ICML 2025 CO-BUILD Workshop on Computational Optimization of Buildings*, 2025.

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

Judah A. Goldfeder, Philippe M. Wyder, J. Nathan Kutz, John Sipple, Victoria Dean, Hod Lipson, Na Li, and Bing Dong. Icml 2025 workshop on computational optimization of buildings (co-build). `https://icml.cc/virtual/2025/workshop/39975`, 2025. Workshop held at the International Conference on Machine Learning (ICML), East Ballroom A, July 18, 2025.

Gabriel Guerra Trigo. 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*, pages 941–943, 2025.

Kanxuan He and Hongshan Guo. A temporal features-enhanced mixture-of-experts approach for indoor temperature prediction. In *ICML 2025 CO-BUILD Workshop on Computational Optimization of Buildings*, 2025.

Zixin Jiang, Xuezheng Wang, and Bing Dong. 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*, pages 924–927, 2025.

Elvin Ko. 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*, pages 921–923, 2025.

Pinaki Prasad Guha Neogi. Multi-scale lstm networks for long-term building temperature prediction: A simplified approach to complex thermal dynamics. In *ICML 2025 CO-BUILD Workshop on Computational Optimization of Buildings*, 2025.

Anh Tuan Nguyen, Duy Hoang Pham, Bee Lan Oo, Mattheos Santamouris, Yonghan Ahn, and Benson TH Lim. Modelling building hvac control strategies using a deep reinforcement learning approach. *Energy and Buildings*, 310:114065, 2024.

Vladimir Pyltsov. Icml 2025 co-build contest: Xgboost iterations. In *ICML 2025 CO-BUILD Workshop on Computational Optimization of Buildings*, 2025.

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

Proceedings of the 16th ACM International Conference on Future Energy Systems (E-Energy '25), New York, NY, United States, June 2025. SIGEnergy, Association for Computing Machinery. ISBN 979-8-4007-1125-1. Conference held June 17–20, 2025.

Vaibhav Sourirajan. Benchmarking forecasting models for long-horizon prediction of temperature distribution in smart buildings. In *ICML 2025 CO-BUILD Workshop on Computational Optimization of Buildings*, 2025.

Phillip Stoffel, Max Berktold, and Dirk Müller. Real-life data-driven model predictive control for building energy systems comparing different machine learning models. *Energy and Buildings*, 305: 113895, 2024.

Liping Sun, Yucheng Guo, Siliang Lu, and Zhenzhen Li. Time-series forecast for indoor zone air temperature with long horizons a case study with sensor-based data from a smart building. In *ICML 2025 CO-BUILD Workshop on Computational Optimization of Buildings*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract presents the new results for the benchmarked setting. This is the main contribution of the work; the setting and scope is also described in the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations and future work are described in the discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This work does not provide theoretical proofs, assumptions, or contributions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The code is available via a Github link; the reproduction instructions can be found there.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is available via a Github link.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The procedures are described in the methodology section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The work describes the standard deviation of the result.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The code is available via a Github link; the compute details can be found there.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No harm was done when conducting the paper.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of the work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The proposed model does not pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The authors of the original XGBoost model and the datset benchmark are cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is provided via a Github link.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing or research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.