2026 European Conference on Computing in Construction
Corfu, Greece
July 12–15, 2026

CAUSALITY-INFORMED TIME-SERIES FOUNDATION MODELS FOR BUILDING PERFORMANCE PREDICTION

Dennis Grauer and Joern Ploennigs
University of Rostock

Abstract

Time-Series Foundation Models, pretrained on large amounts of time-series data, offer strong zero-shot forecasting capabilities. Recent TSFM versions provide improved support for covariate forecasting. This renders them a promising alternative to address current limitations of data-driven building performance prediction approaches: they are highly dependent on data availability, are mostly case-specific, and often favor correlation over causation. In this paper, we compare the predictive performance of two TSFMs with graph neural network approaches and a statistical baseline. We also present an approach for configuring causal covariates for TSFM predictions from graph representations. All models are evaluated using two real-world building datasets.

Introduction

Given the challenges of climate change, it is imperative to improve the energy performance of buildings. Accurate predictions of indoor environmental quality (IEQ) and building energy use can help identify strategies to reduce the environmental impact of buildings (Ali et al., 2024) while maintaining a healthy indoor environment for occupants. With an increasing number of Internet of Things (IoT) sensors deployed in modern buildings, machine learning (ML) approaches that leverage historical data have emerged as an alternative to traditional physical simulations.

Although these approaches perform well on IEQ and building energy management (BEM) tasks, they face two major challenges. First, models often overfit to spurious correlations in the data due to the high redundancy of sensors in buildings (Chen et al., 2025). For example, it is easier for models to learn correlations among temperature sensors than dependencies on heating and cooling systems. Second, models are often limited by the availability and quality of sensor data (Ogundiran et al., 2024; Semasinghe et al., 2025). Consequently, models in the building domain are highly case-specific because they need to be configured for the available sensor data of each building (Ogundiran et al., 2024). Both issues make it difficult to transfer models between buildings, as they do not sufficiently learn generalizable physical causal relationships but instead capture case-specific correlations.

Recently, Time-Series Foundation Models (TSFMs) such as Chronos (Ansari et al., 2024, 2025), TimesFM (Das et al., 2024), Moment (Goswami et al., 2024), and Moirai (Woo et al., 2024) have emerged as an option to address the data availability challenge. They pose to require less training data, as they are pretrained on a variety of time-series datasets from different domains, similar to the transformer architectures of Large Language Models (LLMs) that are trained on large amounts of natural language data. This enables them to make zero-/few-shot forecasts by leveraging a short context window of previous time steps. With increased support for co- and multivariate forecasting in newer model versions, TSFMs promise to incorporate external factors and to predict multiple time series simultaneously.

In this paper, we investigate whether their out-of-the-box performance provides a suitable alternative for short-term forecasting tasks related to Indoor Environmental Quality (IEQ) and Heating, Ventilation, and Air Conditioning (HVAC) loads. We compare the predictive performance of (i) two state-of-the-art TSFMs, (ii) the Autoregressive Integrated Moving Average (ARIMA) model as a strong statistical baseline for time-series forecasting, and (iii) Graph Neural Networks (GNNs) combined with sequence-to-sequence learning methods as a simple way to encode physical causal relationships into machine learning models. By evaluating forecasting accuracy across four tasks, split across two real-world building datasets and varying forecasting horizons, we aim to explore not only the applicability of TSFMs in general but also whether they can be combined with causality-informed approaches such as GNNs. To this end, we *propose a causality-informed TSFM approach* that automatically configures TSFM features from graphs. Our evaluation is conducted on two publicly available datasets: a residential building with sparse HVAC data coverage and no occupancy ground truth, and an office building with diverse HVAC operational data and indoor environment measurements across multiple thermal zones.

This paper is structured as follows. We start with a review of related work on Time-Series Foundation Models and data-driven indoor environment and building energy performance prediction. We then describe our methodology by defining the models and forecasting tasks, followed by an introduction to the scenarios. After presenting and discussing the results, the paper concludes with an outlook.

Background and related works

Indoor environment predictions, HVAC and energy efficiency

Indoor environmental quality is a key factor in ensuring occupants' comfort and health. IEQ is largely governed by heating, ventilation, and air conditioning (HVAC) systems, which also account for a major share of energy use in modern buildings. This relevance makes the resulting complex relationships an important subject of research. Data-driven models that learn from historical data without requiring detailed building information have been employed for various tasks in the IEQ and energy-efficiency domain, including data collection and processing, fault and anomaly detection and analysis, HVAC control management, design optimization, and predictive analysis (Ogundiran et al., 2024). In this paper, we focus on the latter by predicting different IEQ variables and HVAC energy use. HVAC-related predictions mainly involve heating and cooling loads as well as energy consumption, which have been tackled by researchers using deep learning methods (Shi et al., 2024; Liu et al., 2023), neural networks (Borowski and Zwolińska, 2020; Irshad et al., 2024), and graph neural networks (Jia et al., 2024a,b). IEQ prediction is a diverse field, including indoor temperature, CO_2, particulate matter, and thermal comfort. While approaches such as neural networks (Cho and Moon, 2022; Tagliabue et al., 2021; Elmaz et al., 2021) and graph neural networks (Zhang et al., 2023) show convincing results, multiple reviews highlight a key limitation: these approaches remain highly case-specific and are sensitive to the availability and quality of data (Li et al., 2025; Latoń et al., 2025; Ogundiran et al., 2024; Semasinghe et al., 2025). This raises the question of whether zero-/few-shot models could provide an alternative that alleviates this limitation.

Time series foundation models

Time-Series Foundation Models (TSFMs) have recently emerged as a zero-shot alternative for time-series tasks. Built on transformer architectures, they are pretrained on a diverse collection of time-series datasets, enabling them to forecast, detect anomalies, and classify time series. The underlying idea is that TSFMs learn generalizable patterns and correlation structures from these datasets. This can enable zero-shot forecasting, as the model only needs to identify patterns that it has encountered before. In addition, TSFMs support transfer learning by fine-tuning them on new datasets and tasks with fewer data, as the model needs to learn the new task but not the patterns from scratch. Beyond univariate forecasting, models such as Moirai (Woo et al., 2024) and the most recent version of Chronos (Ansari et al., 2025) support integrating past and future covariates into forecasting pipelines.

This suggests that TSFMs could be applicable to different prediction tasks in buildings. However, existing studies report mixed results. Mulayim et al. (2025) evaluate TSFMs for building energy management tasks and find limited zero-shot generalization; TSFMs often fail to outperform classical statistical models. Similarly, Lin et al. (2024) show that TSFMs can handle heterogeneous building data if they are well configured. Extending the analysis to spatio-temporal forecasting, Gupta et al. (2025) report that TSFMs perform competitively at high sampling frequencies but degrade in settings with sparse spatial sensor coverage. Overall, these studies suggest that TSFMs do not necessarily work out of the box but require careful parameterization. This motivates our research question: Can incorporating causal relationships improve TSFM performance in building settings?

Methodology

The aim of this paper is to evaluate the accuracy of TSFMs on different IEQ and energy-consumption tasks and to investigate whether causal information can improve performance. To this end, we perform univariate forecasting with TimesFM and Chronos as two state-of-the-art TSFMs, and covariate forecasting with Chronos using past and future covariates. On the GNN side, we compare two different GNN models as well as our newly proposed approach, which uses GNN graph-representation nodes as covariates for Chronos. We compare these deep learning models with the well-established Autoregressive Integrated Moving Average (ARIMA) model (Newsham and Birt, 2010) as a strong statistical baseline.

Since we want to focus on a comparison of data-driven methods, white-box models are excluded from the scope of this paper. GNN models are favored over other deep learning models, as they allow encoding causal relations in a rather simple manor.

The general task is to predict the next values over a prediction duration P from a set of previous values spanning a context duration D. Since we focus on short-term forecasting in this study, we set $P \in \{1\,\mathrm{h}, 2\,\mathrm{h}, 4\,\mathrm{h}, 8\,\mathrm{h}\}$ to provide an overview of general trends in model forecasting capabilities. The context duration D is defined separately for each model. Depending on the case-specific sampling rate f_s (in minutes; see Scenarios), the context window length C and the horizon length H are defined as $C = (60D)/f_s$ and $H = (60P)/f_s$. In compliance with other work in this domain, we use root mean squared error (RMSE) as the evaluation metric.

Graph neural network models

We employ two types of GNN models: (i) a spatiotemporal GNN (GRUGCN) that uses a Gated Recurrent Unit (GRU) layer to capture temporal regression and a graph convolutional layer, and (ii) a two-layer Graph Convolutional Network with no mechanism to capture temporal dependencies. For both models, we define a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The node set $\mathcal{V}$ denotes a subset $\{v_1, v_2, \ldots, v_i\}$ of available time-series signals from sensor measurements. Each node is associated with a node-feature vector that contains the previous values $\{x_{t-C}, \ldots, x_{t-1}\}$. The edge set $\mathcal{E}$ denotes a set of directed edges $\{e_1, e_2, \ldots, e_i\}$.

Each edge $e = (v_i, v_j)$ represents a directed connection from node v_i to node v_j. We construct these connections based on domain knowledge of physical relationships between the measured signals and dataset exploration, aiming to obtain a graph that captures causal relationships between sensor measurements. Although ontologies like Brick (Balaji et al., 2016) offer schemes for building metadata, they do not cover physical relations and causalities, making automatic graph construction based on causality a complex problem (Langbridge et al., 2023; Grauer and Ploennigs, 2025), not covered in the scope of this paper.

For both models, the context duration is set to $D \in \{4\,\mathrm{h}, 8\,\mathrm{h}\}$. Since we consider different datasets and prediction targets, each requires careful hyperparameter tuning. To this end, we employ the Optuna framework (Akiba et al., 2019) to optimize the learning rate and hidden size of each model layer.

Time series foundation models and ARIMA

We use TimesFM (Das et al., 2024) and Chronos (Ansari et al., 2025) as state-of-the-art TSFMs in our evaluation. We use the most recent versions available at the time of writing, namely TimesFM 2.5-200m and Chronos 2. While TimesFM only supports univariate forecasting, Chronos 2 can incorporate past and future covariates (real-valued or categorical) to support multivariate forecasting. Its in-context learning (ICL) capability shares information within a group of multivariate time series via a group-attention mechanism. We test the effectiveness of this mechanism by adding graph nodes as past covariates to Chronos. Therefore, we define the causal-covariate forecasting task as follows:

$$\hat{y}_{causal} = Chronos(D_{target}, \mathcal{S}) \qquad (1)$$

where D_{target} is defined as the previous time-steps of the prediction target and $\mathcal{S}$ is defined as a subset of $\mathcal{V}$:

$$\mathcal{S} = \{v_i \mid \exists e_{i,target} \in \mathcal{E}\} \quad \forall v_i, v_k \in \mathcal{V} \qquad (2)$$

where $e_{i,target}$ denotes an edge directed from node v_i to the target node v_{target}. The full pipeline for configuring causal covariates from graph representations is given in Algorithm 1.

Besides the causal approach, we define two additional Chronos covariate settings: one that uses only two past covariates for each prediction target, and one that adds past covariates as well as known future covariates. While past-only covariates add information to the context window D, future covariates are incorporated over the prediction horizon H. To create realistic scenarios, we restrict future covariates to quantities that can be estimated (e. g., future temperature or dew point using weather forecasts), are controlled by building operators (e. g., air supply rates or heating setpoints), or follow time patterns (e. g., indicators for working hours or day of the week).

TimesFM serves as a comparison to Chronos, as it is designed and trained for univariate time series only. We set

Algorithm 1 Causal-covariate TSFM forecasting

Require: Directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, target node v_{target}, historical data matrix $\mathrm{X} \in \mathbb{R}^{|\mathcal{V}| \times T}$, context length C, horizon H

Ensure: Forecast $\hat{y}_{t:t+H}$ for v_{target}

```
1: D_target ← X[v_target, t−C : t−1]      ▷ Target history
2: S ← ∅                                   ▷ Causal covariate set
3: for each v_i ∈ V do
4:     if (v_i, v_target) ∈ E then
5:         S ← S ∪ {X[v_i, t−C : t−1]}
6:     end if
7: end for
8: ŷ_{t:t+H} ← Chronos(D_target, S)
9: return ŷ_{t:t+H}
```

$D = 2048$ for both models, as this is the maximum context length used in the first pretraining stage of Chronos (Ansari et al., 2025). Although TimesFM 2.5 supports context lengths of up to 16,000 samples, our exploratory experiments suggest that longer context lengths do not necessarily improve performance while increasing computational cost. For ARIMA, we set $D = 1024$. To simplify forecasting, we use AutoARIMA (Smith et al., 17), which automatically fits the model parameters to the data.

Since the data for the GNN models are split into training and test subsets, TSFM and ARIMA predictions are made only on the test subset to ensure a fair comparison across all model types.

Scenarios

Residential building

We use the dataset presented by Pandey et al. (2025) to evaluate forecasting performance in residential buildings. The dataset contains sensor data capturing occupant behavior, energy use, and IEQ for eight student dorm units over a period of two years. The data are divided into pre- and post-retrofit periods. During the retrofit, heat pumps were installed alongside a heat recovery ventilation (HRV) system, and the insulation of walls and windows was significantly improved. For our prediction tasks, we use the post-retrofit period, as the HVAC and HRV deployment provides more sensor data for the GNN models. Although energy use for HVAC and HRV is available after the retrofit, other operational signals (e. g., HVAC supply air volume or temperature) are missing. The available sensors include outdoor climate conditions (temperature [°C], solar radiation [W/m^2], dew point [°C], humidity [%]), indoor environment variables (relative humidity [%], illuminance, CO_2 [ppm], TVOC [ppb], temperature [lx]), energy use of electrical appliances [kWh], and occupant behavior (door and window state [binary, indicating open or closed]). After data cleaning and preparation, we retain 8.5 months of data sampled at $f_s = 10\,\mathrm{min}$. We use a 70:30 split for training and validation/testing, as recommended for the size of the cleaned dataset (Muraina, 2022). We aim to predict the following target variables:

- **Indoor relative humidity [%]:** We predict the relative humidity in the living room of the student dorm. Input features for the GNN models and causal-covariate setting include outdoor relative humidity [%], dew point temperature [°C], room temperature [°C], HRV power use [kWh], and window state (0 = closed, 1 = open). Outdoor temperature and dew point temperature are used as both past and future covariates for Chronos.

- **HVAC energy use [Wh]:** We predict the combined energy use of electric baseboard heaters in the dorm and heat pumps. Input features include outdoor temperature [°C], relative humidity [%], dew point temperature [°C], indoor temperature [°C], and indoor relative humidity [%]. Outdoor temperature and humidity are used as past and future covariates.

The graph structures used for the GNNs are shown in Figure 2. Due to sparse operational information on ventilation, the graph remains relatively simple. Moreover, as this is a residential building, humidity and HVAC energy use depend strongly on occupant behavior, daily schedules, habits, and personal preferences. In particular, HVAC energy use is a challenging stochastic prediction task: it exhibits spikes at certain time steps and remains close to zero at others, with no clear seasonality (Figure 1).

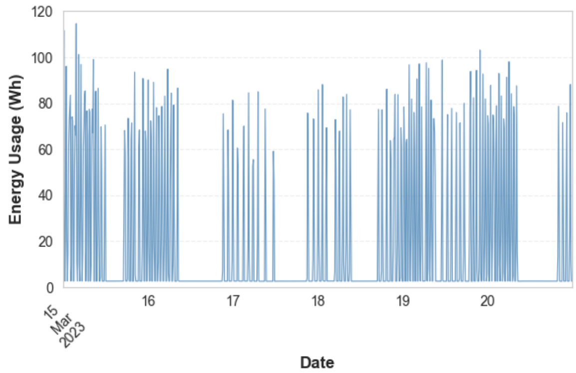

Figure 1: Exemplary HVAC energy usage of student dorm for one week

Office building

We use the Building 59 dataset (Luo et al., 2022) to evaluate forecasting performance in an office building. The records include energy consumption, HVAC operational data, IEQ, and outdoor climate data for a campus building over three years, from January 2018 to December 2020. The IEQ data are separated by thermal zones. For each zone, the zone temperature, cooling and heating temperature setpoints, CO_2 concentration, fan speed for the ventilation air supply, and heating valve position are provided. The building is divided into northern and southern parts, each served by two rooftop units (RTUs). For each RTU, operational data such as supply/return temperature, air flow rates, and setpoints are available. We use data from the southern part of the building, as it provides more detailed measurements. Although the dataset contains a short period of occupancy data, there are no time spans in which both occupancy and indoor-climate data are complete. The following prediction targets are chosen:

- **Zone temperature [°C]:** Zone temperature of an office-space thermal zone. Input features for the graph-based models include temperatures of adjacent zones, outdoor air temperature and humidity, RTU supply-air temperature, and zone fan speeds and heating valve positions. Since office temperatures depend strongly on occupancy, we use a binary indicator for typical working hours (9 am to 5 pm on weekdays) as a past and future covariate for Chronos, alongside outdoor air temperature. The sampling rate is $f_s = 5\,\mathrm{min}$.

- **HVAC energy consumption [kWh]:** HVAC energy consumption is the aggregate of the energy consumption of both RTUs serving the southern part of the building. The graph representation is constructed from each RTU's operational signals (supply/return/mixed-air temperature, supply flow rates), setpoints, fan speeds, heating valve position, temperatures of the thermal zones served by the RTU, and outdoor climate conditions. This yields a complex graph with 115 nodes and 154 edges. For covariate forecasting, we use the working-hours indicator and outdoor air temperature. The sampling rate is $f_s = 15\,\mathrm{min}$, i. e., the original sampling rate of HVAC energy consumption.

Results and discussion

Student dorm - relative humidity

For the task of predicting relative humidity, TSFM and ARIMA significantly outperform the GCN approaches for all horizons (Table 1). With an RMSE below 1 %, these models show superior performance for the shorter prediction horizons of 1 and 2 hours. When the prediction horizon is doubled to 4 and 8 hours, their performance decreases. ARIMA performs at a similar level for the 1-hour prediction but quickly falls behind Chronos and TimesFM as the horizon increases. This is likely due to ARIMA autoregressive stepwise prediction, where errors accumulate for longer horizons. TSFM on the contrary can replicate longer patterns due to the transformer architecture. While adding past-only covariates of outdoor temperature and dew point adds little accuracy, incorporating known future covariates improves performance as the context grows, with particularly strong gains for the 4- and 8-hour horizons. Causal covariates also improve performance but remain behind future covariates.
The GCN model fails to capture the dynamics across horizons and context lengths. Its best performance is achieved

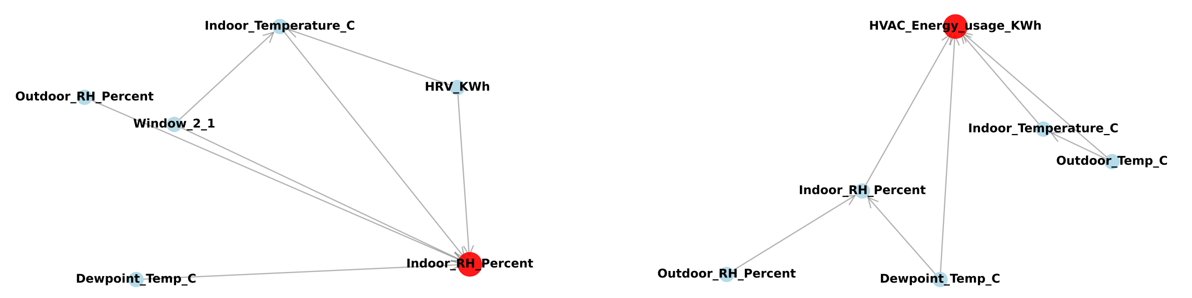

Figure 2: Graph underlying for predictions in the student dorm data set for relative humidity (left) and HVAC energy usage (right).

with the short context length of 4 hours ($C = 24$) for the longest horizon; however, on the 8-hour horizon, performance degrades across all context lengths. The GRUGCN performs best when input and output sequence lengths are equal, yet it still remains behind ARIMA and the TSFMs. These results suggest that TSFMs perform particularly well when predicting variables with a fixed range. Relative humidity, by definition, is bounded between 0 and 100 %; in the test set, values range from 20 % to 57 %, with a mean of 35.1 % and a median of 35.25 %. For machine learning models, the nonlinear relationships between input features can be difficult to learn from scratch, especially given the limited number of samples and meaningful features in the underlying graph.

Table 1: Results for residential humidity forecasting (RMSE in %), best result for each horizon is marked ***bold***

	Horizon P [h]			
Model	1	2	4	8
TimesFM (Univ.)	0.70	1.11	1.80	2.75
Chronos (Univ.)	0.68	1.03	1.65	2.86
Chronos Cov. (past)	**0.66**	1.00	1.60	2.79
Chronos Cov. (p+f)	0.67	**0.93**	**1.24**	**1.79**
Chronos Causal	0.67	0.98	1.52	2.67
ARIMA	0.84	1.59	2.16	3.61
GRUGCN (4 hours)	6.86	7.51	5.01	9.51
GRUGCN (8 hours)	5.24	6.20	6.98	6.25
GCN (4 hours)	9.49	10.81	13.30	8.82
GCN (8 hours)	13.23	8.16	10.55	9.20

Student dorm - HVAC energy usage

Predicting HVAC energy use in the student dorm is a challenging task that none of the models in this study can solve sufficiently well (Table 2). Among all models, Chronos with past covariates achieves the best results for the shortest prediction horizon.

When visualizing 1-hour-ahead predictions over two days in the test set (Figure 3), we find that Chronos can predict peaks in energy use with only a small lag, but it underestimates the magnitude of the energy consumed. The GRUGCN model captures only the baseline, also underestimates peaks, and adds noise around periods with an approximately constant baseline slightly below zero.

For the complex temporal patterns of HVAC energy use, the trend of decreasing accuracy for TSFM and ARIMA forecasts is confirmed, whereas GNN-based models perform similarly across all horizons. For the 4-hour horizon, ARIMA performs best, while the GRUGCN model performs best for the 8-hour horizon.

Because peaks in HVAC energy use are driven by unobserved occupant behavior patterns, all model types struggle to make accurate predictions. Although TSFMs outperform the graph-based models on shorter horizons, the non-seasonal and seemingly random patterns undermine their strengths.

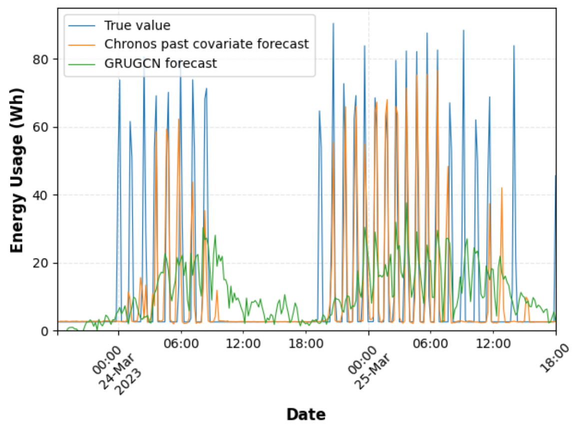

Figure 3: HVAC energy usage of two days in the test set. True value and predictions from GRUGCN with 4 hours context and Chronos past with past covariates for 1 hour horizon are shown.

Office building - zone temperature

Zone temperature in the office building (Table 3) is best predicted by Chronos with causal covariates across all horizons beyond 1 hour. Adding future covariates does not improve longer-horizon predictions at the same scale as in the humidity task: forecasts using only past covariates perform only slightly worse for the 4- and 8-hour horizons, and the RMSE increases with each horizon. The fact that the causal approach performs best for longer horizons

*Table 2: Results for residential HVAC energy forecasting (RMSE in Wh), best result for each horizon is marked **bold***

	Horizon P [h]			
Model	1	2	4	8
TimesFM (Univ.)	24.25	20.25	21.41	22.47
Chronos (Univ.)	15.74	17.09	18.36	19.48
Chronos Cov. (past)	**15.47**	16.92	18.15	19.30
Chronos Cov. (p+f)	15.48	**16.83**	18.12	19.17
Chronos Causal	15.52	16.90	18.15	19.28
ARIMA	16.70	17.30	**17.91**	18.98
GRUGCN (4 h)	17.02	17.93	18.19	18.87
GRUGCN (8 h)	17.53	17.87	18.37	**18.82**
GCN (4 h)	18.85	18.81	19.32	19.87
GCN (8 h)	18.78	18.96	19.41	19.65

(even outperforming predictions that use known future covariates) suggests that Chronos' group-attention mechanism benefits from well-defined causal covariates.

While TimesFM performs best for the 1-hour horizon, it does not appear to scale well for temperature forecasting: it is surpassed by ARIMA for the 2-hour horizon and by both GRUGCN variants for the two longest horizons. Since the sampling rate is doubled compared to the humidity task, TimesFM accuracy decreases substantially for $H > 48$.

As the horizon increases, the GRUGCN model gains ground relative to the TSFMs and ARIMA and surpasses ARIMA for the longest horizon. Compared to humidity forecasting, the gaps between TSFMs and graph-based models shrink. This is likely due to (i) the higher availability of meaningful features for graph learning in the office-building dataset and (ii) the reduced complexity of temperature prediction compared to relative humidity.

*Table 3: Results for office building zone temperature forecasting (RMSE in °C), best result for each horizon is marked **bold***

	Horizon P [h]			
Model	1	2	4	8
TimesFM (Univ.)	**0.160**	0.303	0.611	1.022
Chronos (Univ.)	0.175	0.268	0.365	0.427
Chronos Cov. (past)	0.184	0.302	0.410	0.425
Chronos Cov. (p+f)	0.175	0.284	0.375	0.396
Chronos Causal	0.163	**0.257**	**0.346**	**0.373**
ARIMA	0.183	0.286	0.437	0.696
GRUGCN (4 hours)	0.386	0.418	0.466	0.682
GRUGCN (8 hours)	0.374	0.458	0.496	0.665
GCN (4 hours)	0.600	0.576	0.665	0.766
GCN (8 hours)	0.652	0.734	0.836	0.893

Office building - HVAC energy usage

Among the results for HVAC energy-use forecasting in the office building (Table 4), TimesFM stands out by achieving the best predictions for horizons up to 4 hours. While the GNN-based models and ARIMA perform at a similar level, TimesFM outperforms them but still shows a decrease in accuracy for longer horizons: it falls behind all Chronos variants for the 8-hour horizon. Among the Chronos variants, the past-and-future-covariate setting performs best for the longest horizon. We also find that, in this task adding covariates yields smaller gains than in other cases as the target variable depends on a wide range of factors. In particular, the causal-graph variant, which incorporates a large number of covariates, serves as a scalability test for the in-context learning (ICL) mechanism. The strong drop in performance indicates that this approach may not suitable when many covariates are used. The availability of diverse sensor data highly benefits the graph-based approaches, as they perform similarly to the TSFMs and ARIMA up to the 4-hour horizon. Given the complex graph structure, adding a GRU layer to capture temporal patterns does not improve performance, as results are similar to those of the GCN model.

Although HVAC energy data for the office building are still noisy, the patterns are easier to predict because they are governed by observable behavior represented in the graph.

*Table 4: Results for HVAC energy usage forecasting (RMSE in kWh), best result for each horizon is marked **bold***

	Horizon P [h]			
Model	1	2	4	8
TimesFM (Univ.)	**0.71**	**0.88**	**1.15**	1.50
Chronos (Univ.)	0.94	1.08	1.32	1.44
Chronos Cov. (past)	0.95	1.07	1.30	1.41
Chronos Cov. (p+f)	0.95	1.07	1.29	**1.38**
Chronos Causal	3.58	4.31	4.65	5.62
ARIMA	0.90	1.07	1.32	1.54
GRUGCN (4 hours)	0.96	1.10	1.31	1.56
GRUGCN (8 hours)	0.95	1.14	1.35	1.58
GCN (4 hours)	0.94	1.10	1.31	1.56
GCN (8 hours)	0.96	1.14	1.34	1.58

Conclusion and outlook

In this paper, we compare the predictive performance of two state-of-the-art time-series foundation models with ARIMA and graph neural network (GNN)-based approaches. We also present a new approach for configuring covariates based on causal relationships encoded in graph representations. Using two real-world building case studies of residential and office use, we compare forecasting performance for HVAC energy use, thermal-zone temperature, and relative humidity.

The study finds that TSFMs show strong performance on both examined IEQ tasks, surpassing the GNN-based models. By comparing results across datasets with dense sensor coverage and lower data availability, we conclude that TSFMs are a strong alternative to training machine learning models from scratch, particularly in use cases where high-quality data are scarce. As the number of measurements incorporated for graph learning grows, the performance gap between pretrained TSFMs and the relatively simple GNNs used in our experiments shrinks. In cases with high sensor coverage, we show that Chronos' in-context learning capabilities can benefit from causal covariates and perform better than most Chronos runs with only historical data. However, the HVAC use case also indicates that identifying a refined set of causal variables is important. This suggests that semantic graph models could help identify more effective covariate sets.

Regarding the two TSFMs we compare, Chronos performs best in three out of four experiments and maintains its accuracy better as the prediction horizon increases. Chronos' support for incorporating known future covariates into the prediction horizon shows potential to help maintain performance for longer horizons in the temperature-forecasting task, but does not consistently improve performance across all experiments. As the office-building HVAC results show, it cannot be assumed that covariate forecasting is superior to univariate forecasting in every case. In addition, although covariate support moves beyond purely historical-context forecasting, current prediction pipelines do not incorporate metadata, semantic context, or natural-language inputs, as proposed, for example, by Mulayim et al. (2025).

Overall, while causality-informed TSFMs show promise for building-related forecasting tasks, their performance varies across use cases, and further research is needed to identify the factors that govern their effectiveness. Beyond predictive accuracy, practical deployment also requires interpretability. Although TSFMs have proven valuable for prediction tasks with sparse data, other aspects remain understudied. Trust in predictions in real-world applications is often undermined by the black-box nature of machine learning models (Chen et al., 2023). While ML and GNN models can be augmented with explainability techniques such as GNNExplainer (Ying et al., 2019), the nature of pretraining and learning solely from time-series data and covariates limits TSFM explainability. For future research, we propose augmenting zero-shot TSFMs with semantic context and exploring synergies that combine the strengths of TSFMs with classical machine learning models and semantically explainable causal information.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In ACM SIGKDD.

Ali, D. M. T. E., Motuzienė, V., and Džiugaitė-Tumėnienė, R. (2024). Ai-driven innovations in building energy management systems: A review of potential applications and energy savings. Energies, 17(17).

Ansari, A. F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S. S., Shen, H., Stella, L., Zhang, X., et al. (2025). Chronos-2: From univariate to universal forecasting. arXiv:2510.15821.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D. C., Mahoney, M. W., Torkkola, K., Gordon Wilson, A., Bohlke-Schneider, M., and Wang, Y. (2024). Chronos: Learning the language of time series. Transactions on Machine Learning Research.

Balaji, B., Bhattacharya, A., Fierro, G., Gao, J., Gluck, J., Hong, D., Johansen, A., Koh, J., Ploennigs, J., Agarwal, Y., et al. (2016). Brick: Towards a unified metadata schema for buildings. In Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments.

Borowski, M. and Zwolińska, K. (2020). Prediction of cooling energy consumption in hotel building using machine learning techniques. Energies, 13(23).

Chen, X., Lv, G., Zhuang, X., Roa, C. D., Schiavon, S., and Geyer, P. (2025). Integrating symbolic neural networks with building physics: A study and proposal. J. of Building Engineering, 111:113033.

Chen, Z., Xiao, F., Guo, F., and Yan, J. (2023). Interpretable machine learning for building energy management: A state-of-the-art review. Advances in Applied Energy, 9.

Cho, J. H. and Moon, J. W. (2022). Integrated artificial neural network prediction model of indoor environmental quality in a school building. J. of Cleaner Production, 344.

Das, A., Kong, W., Sen, R., and Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. In Forty-first International Conference on Machine Learning.

Elmaz, F., Eyckerman, R., Casteels, W., Latré, S., and Hellinckx, P. (2021). Cnn-lstm architecture for predictive indoor temperature modeling. Building and Environment, 206.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. (2024). Moment: A family of open time-series foundation models. arXiv:2402.03885.

Grauer, D. and Ploennigs, J. (2025). Does chatgpt know building physics? exploiting foundation models for building performance prediction with gnns. In EC3 Conf.

Gupta, R., Raina, N., Chen, B., Chen, L., Danilov, C., Eckhardt, J., Bernard, K., and Nahrstedt, K. (2025). No one-model-fits-all: Uncovering spatio-temporal forecasting trade-offs with graph neural networks and foundation models. In buildSys.

Irshad, K., Khan, A. I., Zayed, M. E., Algarni, S., and Alqahtani, T. (2024). Performance analysis of air conditioning system integrated with thermal energy storage using enhanced machine learning modelling coupled with fire hawk optimizer. J. of Building Engineering, 98.

Jia, X., Song, H., Nan, X., and Cai, X. (2024a). Prediction of short-term heat load of office buildings based on gnn-lstm modeling. In 2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC). IEEE.

Jia, Y., Wang, J., Reza Hosseini, M., Shou, W., Wu, P., and Mao, C. (2024b). Temporal graph attention network for building thermal load prediction. Energy and Buildings, 321.

Langbridge, A., O'Donncha, F., Ba, A., Lorenzi, F., Lohse, C., and Ploennigs, J. (2023). Causal temporal graph convolutional neural networks (ctgcn). arXiv:2303.09634.

Latoń, D., Grela, J., Ożadowicz, A., and Wisniewski, L. (2025). Artificial intelligence and machine learning approaches for indoor air quality prediction: a comprehensive review of methods and applications. Energies, 18(19).

Li, Y., Arellano-Espitia, F., Aler, R., Igualada, L., and Corchero, C. (2025). Data-driven methods and their applications to building hvac energy consumption prediction: A review. J. of Building Engineering.

Lin, X., Prabowo, A., Razzak, I., Xue, H., Amos, M., Behrens, S., and Salim, F. D. (2024). Exploring capabilities of time series foundation models in building analytics. arXiv:2411.08888.

Liu, H., Liu, Y., Guo, X., Wu, H., Wang, H., and Liu, Y. (2023). An energy consumption prediction method for hvac systems using energy storage based on time series shifting and deep learning. Energy and Buildings, 298.

Luo, N., Wang, Z., Blum, D., Weyandt, C., Bourassa, N., Piette, M. A., and Hong, T. (2022). A three-year dataset supporting research on building energy management and occupancy analytics. Scientific data, 9:156.

Mulayim, O. B., Quan, P., Han, L., Ouyang, X., Hong, D., Bergés, M., and Srivastava, M. (2025). Can time-series foundation models perform building energy management tasks? arXiv:2506.11250.

Muraina, I. (2022). Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. In 7th international Mardin Artuklu scientific research conference.

Newsham, G. R. and Birt, B. J. (2010). Building-level occupancy data to improve arima-based electricity use forecasts. In buildSys, pages 13–18.

Ogundiran, J., Asadi, E., and Gameiro da Silva, M. (2024). A systematic review on the use of ai for energy efficiency and indoor environmental quality in buildings. Sustainability, 16(9).

Pandey, P. R., Liu, Y., Wilson, N., and Dong, B. (2025). Dataset on occupant behavior, indoor environment, and energy use before and after dormitory retrofit. Scientific Data, 12(1).

Semasinghe, K., Perera, S., Jin, X., Samaratunga, M., and Nanayakkara, S. (2025). Data-driven approaches to improve indoor environmental quality: A systematic literature review. In Digital Frontiers in Buildings and Infrastructure.

Shi, Z., Zheng, R., Shen, R., Yang, D., Wang, G., Liu, Y., Li, Y., and Zhao, J. (2024). Building heating load forecasting based on the theory of transient heat transfer and deep learning. Energy and Buildings, 313.

Smith, T. G. et al. (2017–). pmdarima: Arima estimators for Python. [Online; accessed <today>].

Tagliabue, L. C., Cecconi, F. R., Rinaldi, S., and Ciribini, A. L. C. (2021). Data driven indoor air quality prediction in educational facilities based on iot network. Energy and Buildings, 236.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. (2024). Unified training of universal time series forecasting transformers.

Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. NeurIPR, 32.

Zhang, J., Xiao, F., Li, A., Ma, T., Xu, K., Zhang, H., Yan, R., Fang, X., Li, Y., and Wang, D. (2023). Graph neural network-based spatio-temporal indoor environment prediction and optimal control for central air-conditioning systems. Building and Environment, 242.