Covariates Are the Key to Accurate Probabilistic Building Energy Forecasting with Time Series Foundation Models

Alexander Kreusel
Karlsruhe Institute of Technology
Karlsruhe, Germany
alexander.kreusel@student.kit.edu

Matthias Hertel
Karlsruhe Institute of Technology
Karlsruhe, Germany
matthias.hertel@kit.edu

Moritz Noskiewicz
Karlsruhe Institute of Technology
Karlsruhe, Germany
moritz.noskiewicz@kit.edu

Frederik Zahn
Karlsruhe Institute of Technology
Karlsruhe, Germany
frederik.zahn@kit.edu

Heiko Maaß
Karlsruhe Institute of Technology
Karlsruhe, Germany
heiko.maass@kit.edu

Ralf Mikut
Karlsruhe Institute of Technology
Karlsruhe, Germany
ralf.mikut@kit.edu

Veit Hagenmeyer
Karlsruhe Institute of Technology
Karlsruhe, Germany
veit.hagenmeyer@kit.edu

Abstract

Probabilistic building energy demand forecasting is critical for advanced energy management but remains challenging due to building heterogeneity, data scarcity, and non-stationary behavior. Although Time Series Foundation Models (TSFMs) promise scalable generalization from large-scale pretraining and are intended for zero-shot forecasting without additional training, their default settings often fail to reach satisfactory accuracy. As a result, prior work frequently relies on fine-tuning to achieve competitive performance, thereby undermining the out-of-the-box applicability of the models by increasing computational cost, data requirements, and sensitivity to model configuration. Furthermore, successful covariate integration in TSFMs remains largely unexplored, limiting their effectiveness in settings dominated by external influences such as weather conditions or calendar effects. In this work, we aim to demonstrate that covariate integration in zero-shot settings constitutes a more robust and scalable alternative to fine-tuning for building energy parameter forecasting. We evaluate three TSFMs on two real-world building energy demand time series from an office building, with electrical load and cooling demand as target series that exhibit challenges like external influences, limited data and concept drift. Zero-shot and fine-tuned TSFMs are benchmarked against multiple baseline models. Our results show that covariate integration yields substantial performance gains for Chronos-2 and Chronos across both target series, outperforming all baseline models, while MOIRAI does not benefit from future covariates due to insufficient pretraining. Fine-tuning proves to be highly sensitive to the target series and model configuration, while comparable or superior performance can already be achieved in zero-shot settings by substantially extending the context length, which also increases robustness to concept drift. These findings highlight the potential of covariate-informed zero-shot forecasting with TSFMs as a practical and scalable alternative in real-world building energy applications.

ACM Sustainability Week Companion '26, Banff, AB, Canada

ACM ISBN 979-8-4007-2199-1/26/06
https://doi.org/10.1145/3765611.3815507

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Forecasting**.

Keywords

Time Series Foundation Model, Multivariate Model, Building Energy Management, Time Series Forecasting, HVAC

ACM Reference Format:
Alexander Kreusel, Matthias Hertel, Moritz Noskiewicz, Frederik Zahn, Heiko Maaß, Ralf Mikut, and Veit Hagenmeyer. 2026. Covariates Are the Key to Accurate Probabilistic Building Energy Forecasting with Time Series Foundation Models. In *ACM Sustainability Week 2026 (ACM Sustainability Week Companion '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3765611.3815507

1 Introduction

Probabilistic forecasting of building energy demand is critical for advanced energy management but remains challenging due to heterogeneous building types, individual usage patterns, and the interaction of different energy domains [4, 18, 36]. These challenges are amplified by limited data, especially in cold-start scenarios, and by non-stationary conditions caused by concept drifts in real-world operation. As a result, building-specific models are computationally expensive, data-intensive, and therefore poorly scalable [1, 4, 21]. Time Series Foundation Models (TSFMs) offer a promising solution to this problem by enabling out-of-the-box generalization through large-scale pretraining on high-volume, heterogeneous time series datasets covering multiple temporal resolutions and domains, including the energy sector [27, 40]. Most TSFMs are built on the Transformer architecture [47] and use a context window of past observations as input, where attention mechanisms selectively weight the most informative parts of the history to generate predictions of the future. In order to improve predictive performance, TSFMs can be fine-tuned by continuing training on task-specific data to adapt pretrained representations to a particular dataset and forecasting setting [27]. While existing research often relies on fine-tuning to achieve competitive results, it contradicts the out-of-the-box paradigm of TSFMs, as fine-tuning introduces high computational costs, requires additional data, and is sensitive to model configuration. Furthermore, very little research to date has shown beneficial integration of covariate information, despite some TSFMs explicitly offering support for their incorporation. This limits forecasting performance in many applications, as external influences such as weather conditions or operational changes cannot be fully accounted for. With this work, we aim to demonstrate that the integration of covariates in a zero-shot setting represents a more promising research direction than fine-tuning in univariate scenarios, particularly with regard to widespread applicability across a broad range of tasks. Our main contributions are as follows:

- We evaluate three state-of-the-art TSFMs (Chronos-2 [5], Chronos [6], and MOIRAI [49]) on two real-world office building time series (electrical load and cooling load) and compare zero-shot and fine-tuned performance against several baseline forecasting models.
- We investigate the impact of incorporating continuous physical drivers and encoded calendar events as covariates, and assess whether covariate-enhanced zero-shot forecasting can outperform fine-tuned univariate approaches.
- We systematically analyze the effect of substantially extended context lengths and the stability of fine-tuning under different training configurations

This study is structured as follows. Section 2 reviews related work. Section 3 introduces the data and data splits. Section 4 presents the TSFMs and baseline models considered in this study. Section 5 describes the methodology, including experimental design, model configurations, and evaluation metrics. Section 6 reports the results, which are discussed in Section 7. Section 8 summarizes our key findings and provides an outlook into possible research directions.

2 Related Work

The application of TSFMs for univariate forecasting of electrical load in buildings has been widely studied. Prior work, including [16, 35, 37] and [40], examines the effects of forecast horizon and context length in zero-shot settings and consistently reports declining accuracy for longer forecast horizons and improved performance with longer contexts. Despite this evidence, many studies restrict the context to a few days or one week (e.g., [28, 29, 35, 37, 38, 45]), while others extend the context up to about three weeks (e.g., [16, 26]). In comparison to other models, zero-shot performance of TSFMs varies substantially across studies. While [35] and [45] report accuracy comparable to task specific Transformer models, [38] and [40] report only marginal improvements over simple baselines and statistical methods, indicating strong dataset dependence. Moreover, zero-shot benchmarking of TSFMs may be biased by data leakage from large pretraining corpora, as discussed in [35] and [37]. To improve univariate forecasting performance, the effect of fine-tuning on specific datasets is studied in [16, 26, 39] and [12]. While fine-tuning consistently outperforms zero-shot inference, the gains vary widely with dataset, fine-tuning strategy, and hyperparameter choices, complicating cross-study comparisons.

In contrast to the existing literature on electrical building load forecasting, no study to date investigates the application of TSFMs for cooling demand or cooling load forecasting. Instead, existing data driven studies in this domain utilize a variety of approaches such as large language model based time series forecasting frameworks [50], hybrid approaches combining wavelet decomposition and LSTM networks [48], or feedforward neural networks whose parameters are optimized using the Teaching–Learning-Based Optimization (TLBO) metaheuristic [46]. One possible explanation for this research gap is the strong dependence of cooling demand on external drivers such as weather conditions, making the effective integration of covariates essential. Despite recent advances in TSFMs with native covariate support, such as MOIRAI [49], ChronosX [41], and Chronos-2 [5], as well as adaptations of univariate models to multivariate settings like [14] and [42], successful covariate integration into TSFMs for building energy forecasting remains largely unexplored. For instance, [38] reports that MOIRAI does not show improved performance with covariate integration in electrical load forecasting, and [33] highlights similar limitations in electricity price prediction. Given the strong influence of external factors on building energy parameters [17, 23], further research to effectively integrate covariate information into TSFMs remains necessary [35].

3 Data

This section presents the two target time series used in this study, the *Building Load* series and the *Cooling Demand* series. Both time series are real-world data from an office building at KIT Campus North in Germany. Since the dataset is not publicly available to date, the common issue of data leakage [35] can be excluded, allowing for an unbiased comparison of model performance. All data is neither standardized nor normalized, as all employed TSFMs perform internal standardization [5, 6, 49]. For verification and reproducibility of our results, the dataset will be published on https://github.com/KIT-IAI/Probabilistic-Building-Energy-Forecasting-TSFMs.

3.1 Building Load Time Series

The *Building Load* time series contains the total electrical power consumption of the building from January 2024 to May 2025. It has an hourly resolution and exhibits pronounced daily and weekly patterns. Irregularities occur on public holidays and during vacation periods. To capture these effects, a single covariate is used that combines a sinusoidal encoding of the daily cycle with a binary encoding of public holidays, as depicted in Appendix A Figure 12. Depending on the building type and usage characteristics, this encoding can be adapted or extended to represent alternative operational schedules or event structures.

3.2 Cooling Demand Time Series

The *Cooling Demand* time series contains electrical energy consumption data of a Panasonic R410A split-type air-conditioning (AC) unit located in the server room of the building, covering the period from February 23 to May 31, 2025. The series exhibits a daily pattern, since the coefficient of performance (COP) of the unit is strongly influenced by outdoor air temperature, as shown in Figure 1. Therefore, outdoor air temperature information measured at a weather station at KIT Campus North is incorporated as a covariate to improve forecasting accuracy. The short duration of the dataset introduces concept drift due to seasonal effects, as the series begins in the colder month of February and ends in the warmer month of May (see Appendix A Figure 13).

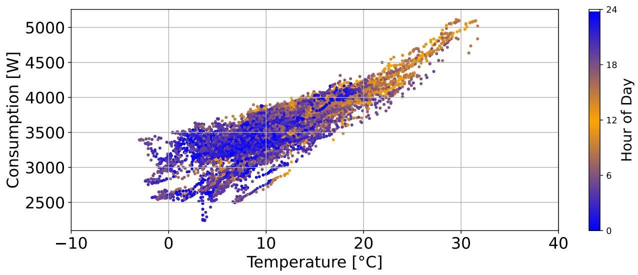


Figure 1: Electrical energy consumption of the split type AC unit versus outdoor air temperature.

3.3 Dataset Splits

For fine-tuning, the data is split into train, validation, and test sets. Since the target series differ in length, two splits are used. For the *Building Load* time series, the year 2024 is used for training, January to March 2025 for validation, and April to May for testing. For the *Cooling Demand* time series, data from February 23 to the end of March is used for training, April for validation, and May for testing.

4 Models

This section presents the TSFMs and baseline models considered in this study. Due to the limited number of TSFMs supporting covariates, we selected three widely used models that employ distinct strategies for integrating covariate information.

MOIRAI. MOIRAI [49] is an encoder-only Transformer that represents time series as non-overlapping patches. These patches are processed through Multi-Patch-Size Input Projection layers, with patch sizes determined by the data frequency. MOIRAI is a multivariate model, supporting the integration of both past and future covariates. It is trained on the LOTSA dataset [49], containing around 27B observations from over 100 uni- and multivariate datasets. The model is available in Small, Base, and Large configurations, ranging from 14M to 311M parameters. This study employs `MOIRAI-1.1R-small`. MOIRAI generates a set of forecast samples from which point predictions are computed as the mean of these samples, following the approach in [10], and quantiles for probabilistic forecasts are computed empirically from the sample distribution at each time step. MOIRAI has two direct successors in MOIRAI-MoE [31] and MOIRAI 2.0 [30]. However, as demonstrated in [25], MOIRAI-MoE does not show improved performance over MOIRAI on cooling load data, particularly on high-resolution time series, while MOIRAI 2.0 is limited to univariate forecasting [30] and therefore not suitable for this study.

Chronos. Chronos [6] is a time series forecasting framework that adapts language models by scaling and quantizing values into a fixed vocabulary for tokenization. It is built on the T5 encoder-decoder Transformer architecture [44] but can also be used with decoder-only architectures such as GPT-2 [43]. Originally a univariate model, Chronos incorporates covariates via a preceding regression model and forecasts the residuals between regression predictions and ground truth [9]. Chronos is available in different sizes, and this study employs the `Chronos-Bolt-Mini` [7] variant with 21M parameters. The model generates multiple forecast samples, from which point forecasts are the sample mean and quantiles are computed as empirical quantiles across samples at each forecast step [6].

Chronos-2. Chronos-2 [5] is a multivariate, encoder-only Transformer based on the T5 encoder [44]. After scaling and tokenization of the input series, it employs a time attention layer to model temporal dependencies within each series, and a group attention layer, which enables in-context learning by sharing information across related series or covariates within a group. The model produces direct multi-step probabilistic forecasts via a quantile head, from which point forecasts are derived as the median of the predicted quantiles. To date, Chronos-2 is available in a single 120M-parameter configuration and is trained on large-scale univariate time series from the Chronos [6] and GIFT-Eval [2] training corpora, supplemented with synthetic multivariate time series [5].

SARIMAX. The seasonal autoregressive integrated moving average with exogenous variables (SARIMAX) model is a statistical time-series forecasting method that extends the ARIMA [13] framework by incorporating both seasonal dynamics and external regressors [3]. In this study, SARIMAX serves as a baseline model to provide a reference for evaluating the TSFMs. The non-seasonal component is specified as ARIMA(1,1,1), while the seasonal component is defined as (1,0,1) with a seasonal period of 24, corresponding to daily seasonality in hourly data. The context length is set to 672 hours (four weeks) of historical data, and model parameters are estimated using maximum likelihood. Quantiles α_i are obtained using the SARIMAX implementation provided by *sktime* [24, 32], which directly supports quantile forecasting.

Last Day. As a second baseline for the *Building Load* time series, forecasts are obtained from the value at the same time on the most recent comparable day, distinguishing between weekdays and weekends. This exploits the strong daily and weekly patterns while accounting for differing consumption behavior.

Linear Regression. As a second baseline for the *Cooling Demand* time series, a linear regression model is fitted on the training data, capturing the relationship between the energy demand and the outdoor air temperature. Forecasts are obtained by applying the model to the temperature data in the forecast window.

5 Methodology

This section presents the methodology used for the main experiments, including experiment design, evaluation metrics and the hardware and software used to conduct the experiments.

5.1 Experiment Design

All three TSFMs are evaluated using a rolling evaluation with stride 1 on the test sets of both target time series. The forecast horizon is fixed to three days, corresponding to 72 values for hourly data. The models are compared using one and four weeks of context, with and without covariate integration, and in zero-shot[1] as well as fine-tuned settings. The model-specific fine-tuning parameters are summarized in Table 1. Originally, MOIRAI is pre-trained with the Negative Log-Likelihood loss [49]. However, as fine-tuning showed no improvement with this loss, it is changed to the MSE. Furthermore, we used a patience of 100, as fine-tuning improvements with MOIRAI often occur after longer periods of stagnation. Due to the limited length of the *Cooling Demand* time series, fine-tuning is performed only with a one-week context, as a longer context would otherwise significantly reduce the size of the training set. As described in Section 4, Chronos incorporates covariate information via a preceding regression model, which is applied consistently during both fine-tuning and inference. For the *Building Load* time series, an Extreme Gradient Boosting model is used, while a Linear Regression model is employed for the *Cooling Demand* time series.

Table 1: Fine-tuning parameters for MOIRAI, Chronos, and Chronos-2, applied to both target series.

Parameter	MOIRAI	Chronos	Chronos-2
Learning Rate	10^{-4}	10^{-5}	10^{-5}
Batch Size	32	32	32
Patch Size	32[2]	1	16
Loss Function	MSE	Cross-Entropy	Quant. Regress.
Optimizer	AdamW	AdamW	AdamW
Patience	100	–	–
Fine-tuning Steps	–	5,000	5,000

5.2 Evaluation Metrics

Mean Absolute Error:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \,. \tag{1}$$

The Mean Absolute Error (MAE) is the average of the absolute differences between the actual values y_i and predicted values $\hat{y}_i$, with N denoting the number of forecast windows times the forecast length. Using absolute errors prevents over- and underestimations from cancelling out. MAE is computationally efficient, easy to interpret due to its consistent units, and robust to outliers [20].

Continuous Ranked Probability Score:

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} \left(F(z) - \mathbf{1}_{\{y \le z\}}\right)^2 \mathrm{d}z. \tag{2}$$

The Continuous Ranked Probability Score (CRPS) is a scoring rule for evaluating probabilistic forecast and is defined as in Equation 2. Here, $F(z)$ denotes the predicted cumulative distribution function, $\mathbf{1}_{\{y \le z\}}$ is an indicator function that equals 1 if $y \le z$ and 0 otherwise, and $\mathbb{R}$ represents the domain of all real numbers over which the integral is evaluated [22, 34]. To limit computational cost, the CRPS is commonly approximated as

$$\mathrm{CRPS}(F, y) \approx \frac{2}{M} \sum_{i=1}^{N} \Lambda_{\alpha_i}(q_{\alpha_i}, y), \tag{3}$$

where $\alpha_i \in (0, 1)$ are the quantile levels, q_{α_i} are the corresponding quantile predictions, and $\Lambda_\alpha(q, y)$ is the quantile loss function

$$\Lambda_\alpha(q, y) = (\alpha - \mathbf{1}_{\{y<q\}})(y - q). \tag{4}$$

This approximation is used, for example, in [11], [25] and [49], and is applied in this study. For CRPS calculation, we use $M = 9$ equally spaced quantiles $\alpha_i \in \{0.1, 0.2, \ldots, 0.9\}$ for MOIRAI, Chronos and the SARIMAX baseline. Chronos-2 outputs 21 quantiles $\alpha_i \in \{0.01, 0.05, 0.10, \ldots, 0.95, 0.99\}$, but since Eq. (3) requires evenly spaced α_i, we exclude the two extreme levels and set $M = 19$ with $\alpha_i \in \{0.05, 0.10, \ldots, 0.95\}$.

5.3 Hardware and Software

The experiments are conducted on a workstation equipped with an Intel Core i7-10700 CPU and an NVIDIA RTX 3070 GPU with 8 GB VRAM. They are performed in a *Python 3.11* environment using libraries such as *Pandas*, *NumPy*, *Matplotlib* and others.

6 Experiment Results

Tables 2 and 3 present the test-set results for both target time series, as well as the respective fine-tuning durations (FTD). On the *Building Load* time series (Table 2), Chronos achieves the overall best performance in fine-tuned setting with a four-week context and integrated covariate, closely followed by the equivalent zero-shot configuration. Chronos-2 attains its best result in zero-shot setting with a four-week context window and covariate integration, performing only slightly worse than the best Chronos configuration. MOIRAI achieves its best performance for point forecasts in fine-tuned setting with a four-week context and without covariate, while its best performance for probabilistic forecasts is obtained in the equivalent zero-shot configuration. In all settings, MOIRAI exhibits a substantially higher error than the best Chronos and Chronos-2 configurations, and only marginally outperforms the SARIMAX baseline in its best-performing setup. Across all models, a longer context of four weeks leads to notably improved performance in zero-shot setting. Regarding the effects of fine-tuning, all models exhibit substantial improvements when using a one-week context window. In contrast, the impact of fine-tuning with a four-week context window is mixed. With respect to covariate integration, Chronos and Chronos-2 benefit significantly from the holiday encoding, whereas MOIRAI shows slightly degraded performance. On the *Cooling Demand* time series (Table 3), Chronos-2 achieves the overall best performance in a zero-shot setting with a four-week context window and covariate integration. Chronos, in its best configuration under the same zero-shot setup, performs only slightly worse in terms of point forecasts, while the performance gap is more pronounced for probabilistic forecasts. MOIRAI also attains its best results in a zero-shot setting with a four-week context and covariate integration, but exhibits a significantly higher error than Chronos and Chronos-2. Although MOIRAI slightly outperforms the SARIMAX baseline in probabilistic forecasting, it clearly falls short to the baseline in point forecast accuracy. In contrast to the *Building Load*, extending the context from one week to four

[1]As the definition of zero-shot sometimes differs in related work, our understanding is defined in Appendix A.
[2]A detailed description of the patch size selection can be found in Appendix A.

Table 2: Results on the *Building Load* time series.

Model	Cov.	Context [Weeks]	Type	MAE [kW]	CRPS [kW]	FTD [min]
Last Day	No	-	-	1.42	-	-
SARIMAX	Yes	4	-	1.38	1.35	-
MOIRAI	No	1	ZS	1.92	1.43	**0**
MOIRAI	No	4	ZS	1.37	1.08	**0**
MOIRAI	Yes	1	ZS	1.98	1.46	**0**
MOIRAI	Yes	4	ZS	1.48	1.12	**0**
Chronos	No	1	ZS	1.29	1.03	**0**
Chronos	No	4	ZS	1.15	0.93	**0**
Chronos	Yes	1	ZS	1.18	0.94	**0**
Chronos	Yes	4	ZS	1.05	0.84	**0**
Chronos-2	No	1	ZS	1.48	1.17	**0**
Chronos-2	No	4	ZS	1.17	0.93	**0**
Chronos-2	Yes	1	ZS	1.28	0.99	**0**
Chronos-2	Yes	4	ZS	1.09	0.86	**0**
MOIRAI	No	1	FT	1.37	1.13	175.10
MOIRAI	No	4	FT	1.36	1.15	110.45
MOIRAI	Yes	1	FT	1.56	1.22	99.41
MOIRAI	Yes	4	FT	1.45	1.17	79.33
Chronos	No	1	FT	1.17	0.93	2.93
Chronos	No	4	FT	1.16	0.93	3.23
Chronos	Yes	1	FT	1.08	0.85	3.00
Chronos	Yes	4	FT	**1.03**	**0.82**	3.21
Chronos-2	No	1	FT	1.15	0.88	10.32
Chronos-2	No	4	FT	1.27	1.04	11.52
Chronos-2	Yes	1	FT	1.13	0.87	6.97
Chronos-2	Yes	4	FT	1.19	0.98	10.18

weeks only leads to small improvements across all models, and fine-tuning significantly degrades test-set performance. By comparison, incorporating temperature information as a covariate substantially improves the forecast accuracy of Chronos-2 and Chronos, and only configurations with covariate integration are able to outperform the covariate-informed baseline in both error metrics. Similar to the *Building Load* time series, MOIRAI only benefits little or not at all from covariate integration.

Comparing the FTD of all models across both time series, MOIRAI on average exhibits substantially longer fine-tuning times, reaching over one hour, whereas Chronos and Chronos-2 only take a few minutes.

Figure 2 displays the development of inference runtime per forecast window for all evaluated TSFMs on the test set of the *Building Load* time series across different context lengths. All models are evaluated in zero-shot setting, both with and without covariate integration, and each reported value represents the mean of a sample of size $N = 5$. MOIRAI consistently achieves the shortest average inference runtime across all context lengths, ranging around 0.02 seconds. Chronos-2 shows slightly higher runtimes of up to 0.03 seconds, while Chronos requires substantially longer runtimes across all context lengths, reaching just above one second. From a context length of 13 weeks onward, the Chronos plot is displayed transparently, as the model's maximum context length of 2,048 is reached.

Table 3: Results on the *Cooling Demand* time series.

Model	Cov.	Context [Weeks]	Type	MAE [W]	CRPS [W]	FTD [min]
Lin. Reg.	Yes	-	-	213.41	-	-
SARIMAX	Yes	4	-	127.00	148.38	-
MOIRAI	No	1	ZS	192.91	150.17	**0**
MOIRAI	No	4	ZS	183.09	141.35	**0**
MOIRAI	Yes	1	ZS	183.35	142.94	**0**
MOIRAI	Yes	4	ZS	174.53	136.77	**0**
Chronos	No	1	ZS	179.55	143.17	**0**
Chronos	No	4	ZS	173.77	137.56	**0**
Chronos	Yes	1	ZS	101.90	82.90	**0**
Chronos	Yes	4	ZS	96.81	78.05	**0**
Chronos-2	No	1	ZS	207.09	163.75	**0**
Chronos-2	No	4	ZS	184.91	145.27	**0**
Chronos-2	Yes	1	ZS	105.88	85.18	**0**
Chronos-2	Yes	4	ZS	**95.22**	**75.32**	**0**
MOIRAI	No	1	FT	217.46	171.71	71.96
MOIRAI	Yes	1	FT	242.28	185.70	64.17
Chronos	No	1	FT	196.86	158.98	2.90
Chronos	Yes	1	FT	102.68	82.57	2.83
Chronos-2	No	1	FT	220.06	199.99	6.67
Chronos-2	Yes	1	FT	136.20	110.45	6.81

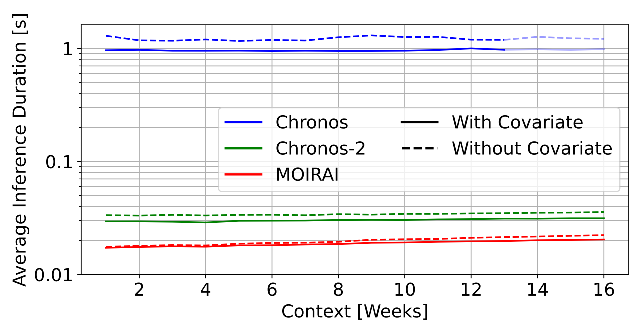


Figure 2: Average inference duration per forecast versus context length on the *Building Load* time series. All models are used in zero-shot setting.

Any excess context is processed but no longer contributes to prediction generation. As runtime differences across models may partly arise from slight differences in the respective inference pipelines, the figure primarily serves to assess how runtime scales with context length and covariate integration within each model. While MOIRAI shows a slight increase in inference runtime with longer context lengths, no such effect is visible for Chronos and Chronos-2. Additionally, incorporating covariates leads to a modest increase in runtime for all models. While the runtime difference between using covariates and not using them slightly grows with increasing context length for MOIRAI, this is not the case for Chronos and Chronos-2.

7 Discussion

This section discusses and interprets the results presented in Section 6. To illustrate differences in model performance, selected figures showing individual forecast windows from the rolling evaluation on the test sets are presented. For reasons of clarity, the figures do not display all confidence intervals used in the computation of the CRPS, but only the central 80 % prediction interval. The section is structured into two subsections, discussing the impact of fine-tuning and covariate integration on model performance.

7.1 Impact of Fine-tuning

Fine-tuning exhibits strongly heterogeneous effects across time series and context lengths, but limited variation across model architectures, suggesting that some conclusions generalize across models. On the *Building Load* time series, fine-tuning yields substantial performance improvements across all TSFMs for a one-week context length. This is likely caused by the pronounced weekly seasonality, which is weakly captured (without repetitions) in zero-shot setting with short context but effectively learned during fine-tuning. When a four-week context is applied, the weekly pattern is represented more explicitly in the input, resulting in smaller or marginal fine-tuning gains compared to the equivalent zero-shot setting. In contrast, none of the models benefit from fine-tuning on the *Cooling Demand* series due to pronounced concept drift, which prevents learned relationships from the training set from generalizing to the test set. This highlights the practical requirement for continuous drift monitoring when deploying fine-tuned TSFMs, potentially necessitating retraining strategies such as rolling fine-tuning, which introduce substantial operational overhead.
Independent of possible performance gains, fine-tuning consistently increases model complexity and implementation effort, raising the question of whether the additional cost is justified, particularly when comparable performance can be achieved in zero-shot settings using longer context lengths. Moreover, fine-tuning is not always straightforward in practice. For MOIRAI, fine-tuning with the default loss function was unsuccessful, requiring a modified loss that resulted in highly volatile training dynamics, delayed improvements, and increased patience requirements. Chronos and Chronos-2 exhibit significantly shorter fine-tuning durations than MOIRAI, as shown in Tables 2 and 3, although optimal performance is often reached very early, sometimes within the first epoch. This suggests that further gains may be achievable through hyperparameter optimization. However, this would introduce additional time series– and configuration-specific overhead, counteracting the intended out-of-the-box usability of TSFMs.

7.2 Impact of Covariate Integration

Tables 2 and 3 demonstrate that Chronos and Chronos-2 benefit from covariate integration across both time series in zero-shot and fine-tuned settings. The relative performance gains are substantially larger for the *Cooling Demand* series than for the *Building Load* series, since the covariate in the *Building Load* task represent infrequent events with limited contribution to the mean test error, enabling some non-covariate models to outperform the baselines. In contrast, no model without covariate information surpasses the SARIMAX point forecast accuracy on the *Cooling Demand* series.

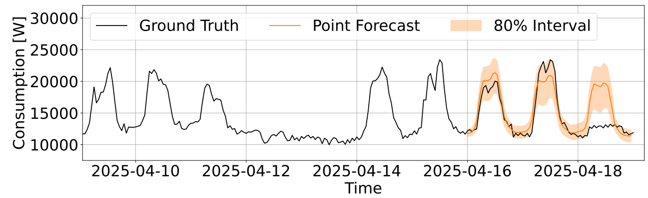


Figure 3: *Building Load* prediction with Chronos in zero-shot setting, one week context and <u>no</u> covariate.

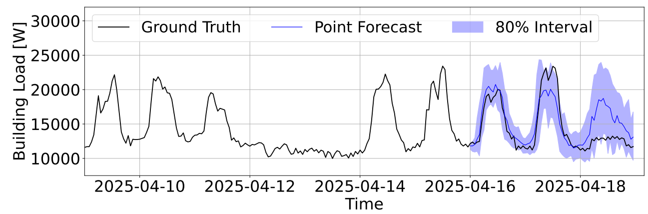


Figure 4: *Building Load* prediction with MOIRAI in zero-shot setting, one week context and covariate.

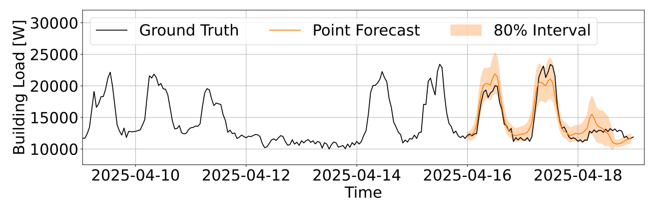


Figure 5: *Building Load* prediction with Chronos in zero-shot setting, one week context and covariate.

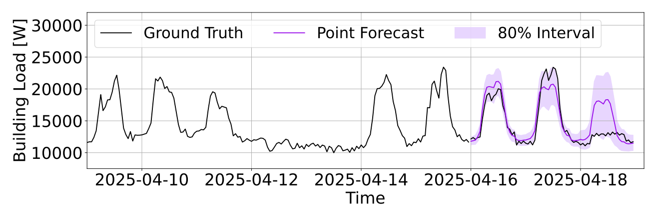


Figure 6: *Building Load* prediction with Chronos-2 in zero-shot setting, one week context and covariate.

To illustrate the effect of holiday encoding on the *Building Load* time series, Figures 4, 5, and 6 present forecasts generated by the three models in a zero-shot setting with covariate integration and a context length of one week. The third day of the forecast horizon corresponds to a public holiday, which cannot be inferred from the context data alone in the absence of external information. In contrast, Figure 3 shows a forecast of the same prediction window generated by Chronos without covariate integration, resulting in a substantial error on the third day. Looking at Figure 4, MOIRAI fails to predict the reduced load on the third forecast day, despite the inclusion of holiday encoding. While the prediction does indicate a lower load compared to the preceding days, this reduction mirrors the pattern observed in the previous week, indicating that the model primarily relies on the inferred seasonality. In contrast, Chronos (Figure 5) produces a substantially adjusted forecast when provided with additional information. Chronos-2 (Figure 6), in the presented model configuration, exhibits behavior similar to MOIRAI. Although it predicts a lower load on the final forecast day, the associated error remains large. This behavior can be explained by the absence of a comparably encoded event within the context window, with the most recent similar event occurring approximately 15 weeks earlier, as marked by the red arrow in Figure 7. When this extended context is made available to Chronos-2, the model adapts its prediction for the final forecast day. However, this is not the case for MOIRAI, as shown in Appendix A Figure 14.

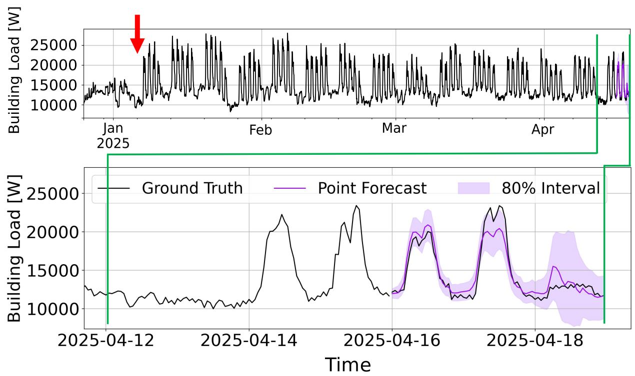


Figure 7: *Building Load* prediction with Chronos-2 in zero-shot setting, 16 weeks context and covariate. The red arrow marks a public holiday in the context.

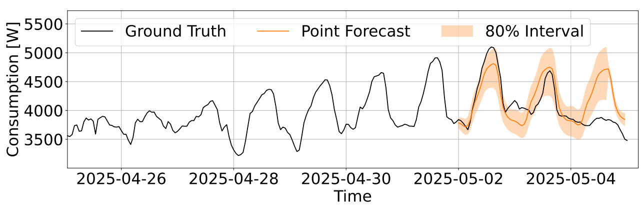


Figure 8: *Cooling Demand* prediction with Chronos in zero-shot setting, one week context and <u>no</u> covariate.

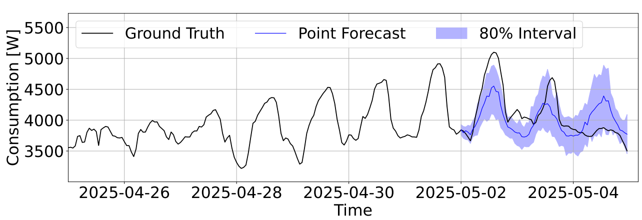


Figure 9: *Cooling Demand* prediction with MOIRAI in zero-shot setting, one week context and covariate.

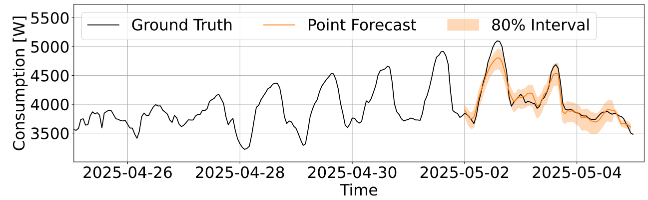


Figure 10: *Cooling Demand* prediction with Chronos in zero-shot setting, one week context and covariate.

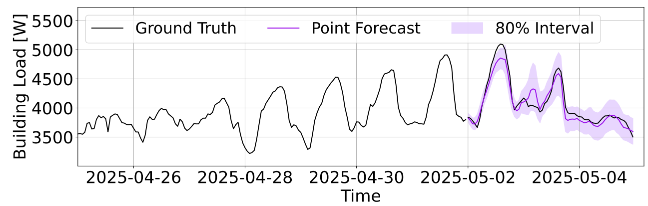


Figure 11: *Cooling Demand* prediction with Chronos-2 in zero-shot setting, one week context and covariate.

Figures 9, 10 and 11 present forecasts on the *Cooling Demand* time series generated in a zero-shot setting with a one-week context and outdoor temperature as covariate. For comparison, Figure 8 shows a Chronos forecast produced without covariate integration. On the third day of the forecast horizon, a pronounced drop in cooling demand occurs, directly caused by a change of outdoor air temperature. Without access to external information, this behavior cannot be predicted, leading to a substantial prediction error. Consistent with the findings for the *Building Load* time series, MOIRAI does not benefit from covariate integration, whereas Chronos exhibits a markedly improved forecast accuracy when temperature information is provided. In contrast to the *Building Load* series, Chronos-2 already shows a substantial accuracy improvement with only a one-week context window.

The presented results reveal pronounced differences in the capability of the compared TSFMs to capture relationships between the target variable and its covariate. Although MOIRAI is designed as a multivariate model with native support for past and future covariates [49], the presented results demonstrate that it does not benefit from the provided future covariates. This observation is consistent with the findings reported in [38] and [33], where similar behavior is observed but no further explanation is provided. While MOIRAIs capability of future covariate integration is strongly indicated in [49], the LOTSA training set contains only past covariates [49], suggesting that the model was never trained to utilize future information. This interpretation is further supported by the absence of any implementation of future covariates in both the pre-training and fine-tuning scripts provided for `MOIRAI-1.1R-small`. Although future covariates are technically passed to the model during inference and flattened together with the target, this processing likely accounts for the slight differences observed in the resulting predictions rather than reflecting an effective use of future covariate information. Consequently, while the implication in [49] is technically correct, the classification of MOIRAI as a model with

effective support for future covariates should be critically reconsidered. This assessment does not extend to past covariates, as their effect was not evaluated in the presented experiments and both the pre-training and fine-tuning scripts include an implementation for past covariates.
The observed performance improvements of Chronos resulting from covariate integration are consistent with its architectural design, as covariates are incorporated via an upstream regression model. The effectiveness of this approach is demonstrated by the comparison between the linear regression baseline on the *Cooling Demand* time series and Chronos combined with the linear regression model, where both regression models are trained on identical data. The substantial gain over the regression-only baseline highlights the added value of residual modeling with Chronos. Notably, these improvements are already evident for short context lengths, regardless of whether covariates are continuous parameters or encodings of singular events. Since covariate integration occurs during the fitting of the regression model rather than within the context window, the training dataset must include the relevant singular events for their effects to be captured and exploited during forecasting. However, this strategy increases complexity and requires domain expertise, since regression model suitability is highly time-series dependent. Training the regression model also weakens the cold-start capability of the model and increases its susceptibility to concept drift. While refitting the regression model on the context window at inference time could mitigate these issues, it would introduce substantially higher computational costs and inference latency.

Chronos-2 emerges as the only evaluated TSFM with beneficial native integration of future covariate information. At the same time, it achieves predictive performance on par with Chronos while retaining key advantages of TSFMs, including out-of-the-box applicability and cold-start capability. However, the effective use of covariates depends strongly on their type and requires different integration strategies. Relationships between the target variable and continuous physical drivers are already captured reliably with relatively short context, while encoding of discrete events requires careful selection of the context length. In particular, rare events may necessitate substantially extended contexts to ensure that relevant historical occurrences are included. This raises questions regarding the impact of extended context lengths on forecasting performance, particularly in prediction windows that do not contain any encoded rare events. While long historical contexts may be necessary to capture infrequent events, they may also amplify the influence of concept drift by introducing outdated or irrelevant information, thereby adversely affecting predictive accuracy. To investigate this, we conduct follow-up experiments with Chronos-2, systematically analyzing the effects of long context lengths on forecast quality. The results are presented in Appendix B.

8 Conclusion

Time Series Foundation Models have already shown promising results for building energy parameter forecasting, but prior work has been limited to univariate settings and often required fine-tuning to achieve competitive accuracy. However, modeling building energy dynamics purely through autoregressive patterns is inherently limiting, as key drivers such as weather, occupancy, and operational schedules are not accounted for. In this work, we investigate the influence of covariate integration into TSFMs for building energy parameter forecasting. We evaluate three TSFMs on two real-world office-building energy demand time series, electrical load and cooling demand, under external influences, limited training data, and concept drift. For each target time series, we include covariate information, evaluate their respective impact on predictive accuracy and systematically compare zero-shot and fine-tuned model configurations. Furthermore, we study the effect of context lengths up to multiple months on accuracy and inference cost. Our key findings can be summarized as follows:

- Incorporating covariates into TSFMs yields large accuracy gains for the analyzed energy time series forecasts, and is essential to achieve competitive performance in scenarios driven by external factors. Among the evaluated TSFMs only Chronos-2 provides reliable native covariate support. The results further indicate that continuous covariates and discrete event indicators require different handling, including appropriate choice of context length and event encoding.
- Zero-shot performance improves substantially with long contexts spanning multiple weeks to months. For Chronos-2, no meaningful degradation is observed even when concept drift occurs within the context, and increased context lengths only marginally affect inference time. However, as pointed out in [12], the usable context length of some TSFMs is explicitly limited (e.g., TimesFM [15] to 512 time steps), which makes multi-month contexts infeasible for hourly data. This limitation should be considered both when selecting TSFMs for a given application and when developing new TSFMs.
- Fine-tuning is highly sensitive to model configuration and the target series, limiting out-of-the-box robustness. On the tested datasets, comparable or superior accuracy is achieved in zero-shot setting by extending the context length, reducing the practical benefit of fine-tuning.

This work provides the basis for future research. First, the generalization of our findings to upcoming TSFMs with future covariate support, such as COSMIC [8] and TabPFN [19], can be tested. Second, motivated by the results presented in Section 6 and the follow-up experiments presented in Appendix B, event-aware dynamic context selection should be investigated, in which context length is chosen to ensure sufficient encoded event coverage without adding uninformative history.

Acknowledgments

This project is funded by the Helmholtz Association under the Program "Energy System Design" and the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI.

Declaration on the use of generative AI and AI-assisted technologies

During the preparation of this work the authors used ChatGPT in order to improve clarity and fluency in the writing process. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] A. Ahmadi, H. Zareipour, and H. Leung. 2025. Globalization for Scalable Short-term Load Forecasting. doi:10.48550/arXiv.2507.11729

[2] T. Aksu, G. Woo, J. Liu, X. Liu, C. Liu, S. Savarese, C. Xiong, and D. Sahoo. 2024. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393* (2024). doi:10.48550/arXiv.2410.10393

[3] F. R. Alharbi and D. Csala. 2022. A seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) forecasting model-based time series approach. *Inventions* 7, 4 (2022), 94. doi:10.3390/inventions7040094

[4] K. Amasyali and N. M. El-Gohary. 2018. A review of data-driven building energy consumption prediction studies. 81 (2018), 1192–1205. doi:10.1016/j.rser.2017.04.095

[5] A. F. Ansari, O. Shchur, J. Küken, A. Auer, B. Han, P. Mercado, S. S. Rangapuram, H. Shen, L. Stella, X. Zhang, et al. 2025. Chronos-2: From Univariate to Universal Forecasting. (2025). doi:/10.48550/arXiv.2510.15821

[6] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. Sundar Rangapuram, S. Pineda Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. Gordon Wilson, M. Bohlke-Schneider, and Y. Wang. 2024. Chronos: Learning the Language of Time Series. doi:/10.48550/arXiv.2403.07815

[7] A. F. Ansari, C. Turkmen, O. Shchur, and L. Stella. 2024. Fast and accurate zero-shot forecasting with Chronos-Bolt and AutoGluon. https://aws.amazon.com/blogs/machine-learning/fast-and-accurate-zero-shot-forecasting-with-chronos-bolt-and-autogluon/

[8] A. Auer, R. Parthipan, P. Mercado, A. F. Ansari, L. Stella, B. Wang, M. Bohlke-Schneider, and S. S. Rangapuram. 2025. Zero-Shot Time Series Forecasting with Covariates via In-Context Learning. doi:/10.48550/arXiv.2506.03128

[9] AutoGluon. 2025. *Forecasting with Chronos*. AutoGluon. https://auto.gluon.ai/1.1.1/tutorials/timeseries/forecasting-chronos.html AutoGluon documentation (v1.1.1), tutorial page. Licensed under Apache 2.0..

[10] M. Beichter, N. Friederich, J. Pinter, D. Werling, K. Phipps, S. Beichter, O. Neumann, R. Mikut, V. Hagenmeyer, and B. Heidrich. 2025. Decision-focused fine-tuning of time series foundation models for dispatchable feeder optimization. *Energy and AI* 21 (2025), 100533. doi:10.1016/j.egyai.2025.100533

[11] J. Berrisch and F. Ziel. 2024. Multivariate probabilistic CRPS learning with an application to day-ahead electricity prices. 40, 4 (2024), 1568–1586. doi:10.1016/j.ijforecast.2024.01.005

[12] S. Bose, Y. Li, A. Van Sant, Y. Zhang, and K. Kim. 2024. From RNNs to Foundation Models: An Empirical Study on Commercial Building Energy Consumption. doi:/10.48550/arXiv.2411.14421

[13] G. Box. 2013. Box and Jenkins: Time Series Analysis, Forecasting and Control. In *A Very British Affair*. Palgrave Macmillan. doi:10.1057/9781137291264_6

[14] A. Dange and S. Sarawagi. 2025. TFMAdapter: Lightweight Instance-Level Adaptation of Foundation Models for Forecasting with Covariates. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 498–507. doi:/10.1145/3746252.3761272

[15] A. Das, W. Kong, R. Sen, and Y. Zhou. 2024. A decoder-only foundation model for time-series forecasting. doi:/10.48550/arXiv.2310.10688

[16] Y. Gu. 2025. Deep Learning Methods for Built Environment Operational Management. (2025). https://vtechworks.lib.vt.edu/server/api/core/bitstreams/4339afc2-459c-4670-b3b9-53a2afaa459d/content

[17] F. Haldi and D. Robinson. 2011. The impact of occupants' behaviour on building energy demand. *Journal of Building Performance Simulation* 4, 4 (2011), 323–338. doi:/10.1080/19401493.2011.558213

[18] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman. 2016. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. doi:/10.1016/j.ijforecast.2016.02.001 ISSN: 0169-2070 Issue: 3 Pages: 896-913 Publication Title: International Journal of Forecasting Volume: 32.

[19] S. B. Hoo, S. Müller, D. Salinas, and F. Hutter. 2026. From Tables to Time: Extending TabPFN-v2 to Time Series Forecasting. doi:/10.48550/arXiv.2501.02945

[20] A. Jadon, A. Patil, and S. Jadon. 2024. A Comprehensive Survey of Regression-Based Loss Functions for Time Series Forecasting. In *Data Management, Analytics and Innovation*, N. Sharma, A. C. Goje, A. Chakrabarti, and A. M. Bruckstein (Eds.). Springer Nature Singapore, Singapore, 117–147.

[21] R. Kaur Jagait, M. N. Fekri, K. Grolinger, and S. Mir. 2021. Load Forecasting Under Concept Drift: Online Ensemble Learning With Recurrent Neural Network and ARIMA. *IEEE Access* 9 (2021), 98992–99008. doi:10.1109/ACCESS.2021.3095420

[22] A. Jordan, F. Krüger, and S. Lerch. 2019. Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software* 90 (2019), 1–37. doi:10.18637/jss.v090.i12

[23] J. Kang and D. M. Reiner. 2022. What is the effect of weather on household electricity consumption? Empirical evidence from Ireland. 111 (2022), 106023. doi:10.1016/j.eneco.2022.106023

[24] F. Király, M. Löning, T. Bagnall, M. Middlehurst, A. Ray, S. Ganesh, M. Walter, G. Oastler, J. Lines, ViktorKaz, B. Heidrich, L. Mentel, Jigyasu, S. Mishra, chrisholder, D. Bartling, A. Shakir, L. Tsaprounis, RNKuhns, C. Gilbert, F. Angelim, M. Baichoo, F. H. Nshuti, H. Akmal, X. Wu, alex-gregory-ds, Guzal, P. Prajapati, and T. Owoseni. 2025. sktime/sktime: v0.40.1. doi:10.5281/zenodo.17705304

[25] A. Kreusel, M. Hertel, M. Noskiewicz, H. Maaß, R. Mikut, and V. Hagenmeyer. 2025. Evaluating Time-Series Foundation Models for Cooling Demand Forecasting with Little Data. In *Proceedings - 35. Workshop Computational Intelligence: Berlin, 20. - 21. November 2025 (AIA Proceedings)*, H. Schulte, F. Hoffmann, and R. Mikut (Eds.). Karlsruher Institut für Technologie (KIT), 301–322. doi:10.5445/IR/1000186052

[26] R. Liang, Y. Deng, D. Xie, F. He, and D. Wang. 2024. Enabling Time-series Foundation Model for Building Energy Forecasting via Contrastive Curriculum Learning. doi:/10.48550/arXiv.2412.17285

[27] Y. Liang, H. Wen, Y. Nie, Y. Jiang, M. Jin, D. Song, S. Pan, and Q. Wen. 2024. Foundation Models for Time Series Analysis: A Tutorial and Survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2024). ACM, 6555–6565. doi:10.1145/3637528.3671451

[28] N. Lin, D. Yun, W. Xia, P. Palensky, and P. P. Vergara. 2024. Comparative Analysis of Zero-Shot Capability of Time-Series Foundation Models in Short-Term Load Prediction. doi:/10.48550/arXiv.2412.12834

[29] X. Lin, A. Prabowo, I. Razzak, H. Xue, M. Amos, S. Behrens, and F. D. Salim. 2024. Exploring Capabilities of Time Series Foundation Models in Building Analytics. doi:10.48550/arXiv.2411.08888

[30] C. Liu, T. Aksu, J. Liu, X. Liu, H. Yan, Q. Pham, D. Sahoo, C. Xiong, S. Savarese, and J. Li. 2025. Moirai 2.0: When Less Is More for Time Series Forecasting. (2025). doi:/10.48550/arXiv.2511.11698

[31] X. Liu, J. Liu, G. Woo, T. Aksu, Y. Liang, R. Zimmermann, C. Liu, S. Savarese, C. Xiong, and D. Sahoo. 2024. Moirai-MoE: Empowering Time Series Foundation Models with Sparse Mixture of Experts. (2024). doi:/10.48550/arXiv.2410.10469

[32] M. Löning, A. Bagnall, S. Ganesh, V. Kazakov, J. Lines, and F. J. Király. 2019. sktime: A unified interface for machine learning with time series. (2019). doi:10.48550/arXiv.1909.07872

[33] G. Marchesi, A. Ballarino, A. Brusaferri, G. Marchesi, A. Ballarino, and A. Brusaferri. 2025. Assessing Time Series Foundation Models for Probabilistic Electricity Price Forecasting: Toward a Unified Benchmark. 18, 23 (2025). doi:10.3390/en18236269

[34] F. Marchesoni-Acland, R. Alonso-Suárez, A. Herrera, J. Kherroubi, J.-M. Morel, and G. Facciolo. 2024. A CRPS Loss for Deep Probabilistic Regression. In *2024 IEEE URUCON*. 1–4. doi:10.1109/URUCON63440.2024.10850406

[35] M. Meyer, D. Zapata, S. Kaltenpoth, and O. Müller. 2024. Benchmarking Time Series Foundation Models for Short-Term Household Electricity Load Forecasting. https://arxiv.org/abs/2410.09487

[36] E. Mocanu, P. H. Nguyen, M. Gibescu, and W. L. Kling. 2016. Deep learning for estimating building energy consumption. *Sustainable Energy, Grids and Networks* 6 (2016), 91–99. doi:10.1016/j.segan.2016.02.005

[37] O. B. Mulayim, P. Quan, L. Han, X. Ouyang, D. Hong, M. Bergés, and M. Srivastava. 2024. Are Time Series Foundation Models Ready to Revolutionize Predictive Building Analytics?. In *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (New York, NY, USA, 2024) *(BuildSys '24)*. Association for Computing Machinery, 169–173. doi:10.1145/3671127.3698177

[38] O. B. Mulayim, P. Quan, L. Han, X. Ouyang, D. Hong, M. Bergés, and M. Srivastava. 2025. Can Time-Series Foundation Models Perform Building Energy Management Tasks? doi:10.48550/arXiv.2506.11250

[39] S. Park, S. Jung, S. Jung, S. Rho, and E. Hwang. 2021. Sliding window-based LightGBM model for electric load forecasting using anomaly repair. 77, 11 (2021), 12857–12878. doi:10.1007/s11227-021-03787-4

[40] Y.-J. Park, F. Germain, J. Liu, Y. Wang, T. Koike-Akino, G. Wichern, N. Azizan, C. Laughman, and A. Chakrabarty. 2025. Probabilistic forecasting for building energy systems using time-series foundation models. 348 (2025), 116446. doi:10.1016/j.enbuild.2025.116446

[41] S. Pineda Arango, P. Mercado, S. Kapoor, A. F. Ansari, L. Stella, H. Shen, H. Senetaire, C. Turkmen, O. Shchur, D. C. Maddix, M. Bohlke-Schneider, Y. Wang, and S. S. Rangapuram. 2025. ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables. doi:/10.48550/arXiv.2503.12107

[42] G. Qin, Z. Chen, Y. Liu, Z. Shi, H. Liu, X. Huang, J. Wang, and M. Long. 2025. CoRA: Covariate-Aware Adaptation of Time Series Foundation Models. (2025). doi:/10.48550/arXiv.2510.12681

[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog* 1, 8 (2019), 9. https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf

[44] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 67 pages. http://jmlr.org/papers/v21/20-074.html

[45] Harish Kumar Saravanan, Shivam Dwivedi, and Pandarasamy Arjunan. 2024. Analyzing the Performance of Time Series Foundation Models for Short-term Load Forecasting. In *Proceedings of the 11th ACM International Conference on*

Systems for Energy-Efficient Buildings, Cities, and Transportation (New York, NY, USA, 2024) *(BuildSys '24)*. Association for Computing Machinery, 237–238. doi:10.1145/3671127.3698708 event-place: Hangzhou, China.

[46] M. H. Sulaiman and Z. Mustaffa. 2024. Chiller energy prediction in commercial building: A metaheuristic-Enhanced deep learning approach. 297 (2024), 131159. doi:/10.1016/j.energy.2024.131159

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017) *(NIPS'17)*. Curran Associates Inc., 6000–6010.

[48] Y. Wang, H. Cheng, H. Chen, M. Ye, Y. Ren, and C. Yang. 2025. A hybrid model based on wavelet decomposition and LSTM for short-term energy consumption prediction of chillers. 99 (2025), 111539. doi:/10.1016/j.jobe.2024.111539

[49] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. 2024. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2178, 25 pages.

[50] C. Zhu, Y. Yang, H. Chen, and M. Zeng. 2025. An Enhanced, Lightweight Large Language Model-Driven Time Series Forecasting Approach for Air Conditioning System Cooling Load Forecasting. *Mathematics* 13, 23 (2025). doi:10.3390/math13233887

APPENDIX

A Data & Definitions

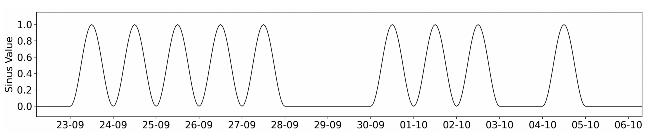


Figure 12: Sinusoidal encoding of public holidays.

Figure 12 illustrates the holiday encoding scheme using a public holiday on October 3rd as an example. Weekdays are represented by a sinusoidal encoding, while weekends and public holidays remain at zero.

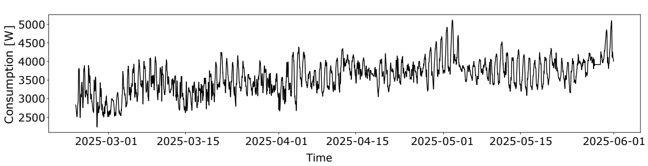


Figure 13: *Cooling Demand* time series.

Figure 13 displays the *Cooling Demand* time series. A drift of electrical energy consumption is apparent across the entire series.

Definition Zero-shot. The TSFM is used without any task-specific training of its own weights on the target dataset. In this work, Chronos with a separately fitted regression model for covariates is still considered zero-shot, as long as Chronos itself is not fine-tuned on the residual series, even though the regression model is trained on the training set.

Patch Size MOIRAI. MOIRAI allows flexible selection of patch sizes that have been pretrained on different frequency patterns of time series data. Available options are [8, 16, 32, 64, 128, "auto"], where "auto" automatically determines a suitable patch size based on the frequency of the input time series [49]. Table 4 presents the results of MOIRAI using different patch sizes. Notably, the "auto" function does not achieve the optimal result and does not correspond to any of the explicitly available patch sizes, leaving it unclear which patch size is internally selected by the model.

Table 4: MAE [W] of MOIRAI on both target series using different patch sizes. The experiments are conducted using a prediction length of 3 days, a context length of 7 days, and <u>no</u> covariate.

Patch Size	Building Load [kW]	Cooling Demand [W]
8	4.17	323.34
16	4.58	358.87
32	1.92	192.91
64	2.25	199.84
128	3.30	227.90
"auto"	2.30	200.31

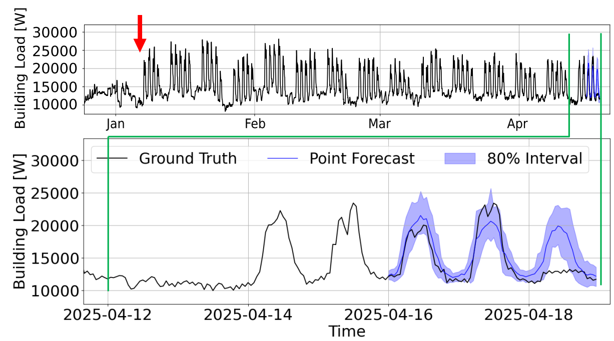
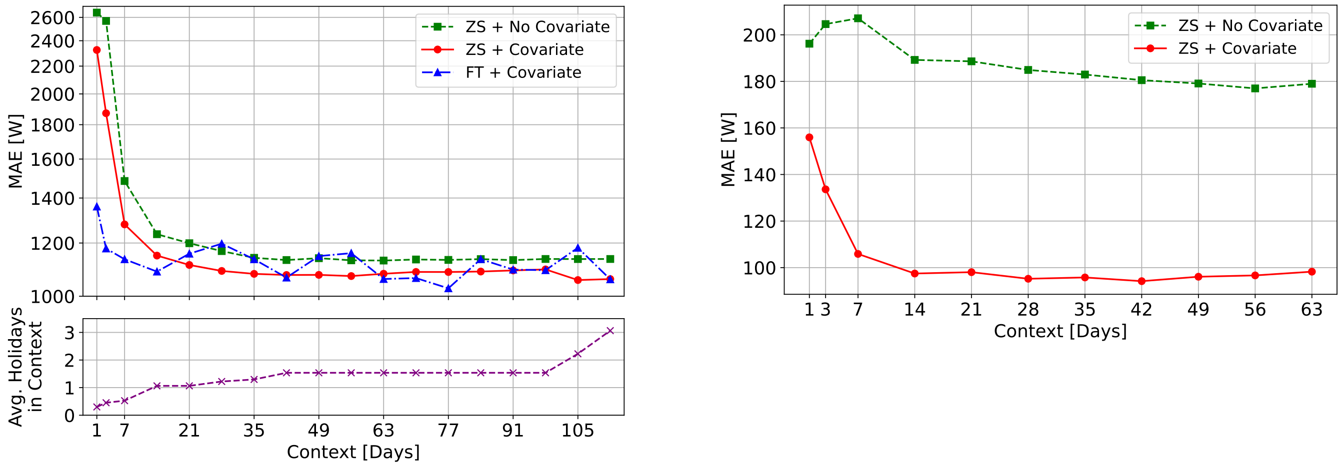


Figure 14: *Building Load* prediction with MOIRAI in zero-shot setting, 16 weeks context and covariate. The red arrow marks a public holiday in the context.

B Follow-up Experiments

Based on the findings discussed in Section 7, this section presents follow-up experiments regarding the effect of substantially increased context lengths on forecast accuracy of Chronos-2. Figure 15a illustrates the development of the MAE of different Chronos-2 configurations on the test set of the *Building Load* time series across different context lengths, reaching from one day to 112 days (16 weeks). The red and green curves denote zero-shot models with and without covariates, while the blue curve represents covariate-informed fine-tuned models. For the fine-tuned results, a separate model is fine-tuned for each context length. All fine-tuning settings are identical across models and follow Table 1, with context length being the only varying parameter. The lower plot (purple) shows the average number of holidays within the context window for each context length, computed exclusively from context windows associated with forecast windows containing at least one holiday that does not fall on a weekend. Thereby, only cases in which holiday encoding is expected to benefit forecast performance are considered. For contexts up to six weeks, both zero-shot configurations exhibit a nearly parallel and steep reduction in MAE, followed by a gradual saturation. Across all context lengths, holiday encoding reduces the



(a) Building Load Series **(b) Cooling Demand Series**

Figure 15: MAE of different Chronos-2 configurations across context lengths for both target time series

MAE, even without accurate holiday prediction (cf. Figure 6), indicating a regularizing effect on the underlying demand structure that is effective already for short contexts. Between 6 and 14 weeks, the average holiday count in the context window remains constant and the non-covariate zero-shot MAE plateaus, indicating that longer contexts add little to non usable autoregressive information. In contrast, the covariate-informed zero-shot MAE increases slightly, which is consistent with a reduced effective contribution (dilution) of the sparse holiday signal when context length grows without increasing event coverage. At context lengths $\geq$ 15 weeks, MAE drops sharply for the covariate-informed model, coinciding with the inclusion of the additional holiday depicted in Figure 7. These findings motivate possible future research regarding event-aware dynamic context selection, in which the context window is chosen to include a minimum number of encoded events while avoiding unnecessary context expansion when no additional events are added. Comparing the zero-shot results to the fine-tuned models, fine-tuning improves MAE mainly for short contexts, but this advantage largely vanishes as context increases and zero-shot models often match or outperform it. Furthermore, the fine-tuned models show a significant MAE volatility across context lengths, indicating strong configuration sensitivity and potentially costly hyperparameter search, whereas zero-shot performance is more stable.
Figure 15b presents a similar analysis for the *Cooling Demand* time series. Fine-tuned configurations are omitted due to the apparent concept drift and the substantially smaller training set, and the maximum evaluated context length is limited to nine weeks by the series duration. The observable effects partly differ from those obtained for the *Building Load* time series. For very short context lengths, the non–covariate-informed model initially exhibits a slight deterioration as the context is extended, suggesting that the most recent observations are the most informative for accurate forecasting. Only beyond a context length of approximately two weeks does the error start to decrease again. In contrast, the covariate informed model shows a similar behavior as on the *Building Load* series, as increasing the context length yields a strong initial reduction in MAE. However, saturation occurs earlier, which is consistent with weaker autoregressive structure and the absence of pronounced weekly seasonality in cooling demand. Notably, the MAE does not deteriorate significantly with longer context lengths, indicating model robustness despite the apparent concept drift in the series.