Evaluating Time-Series Foundation Models for Cooling Demand Forecasting with Little Data

Alexander Kreusel, Matthias Hertel, Moritz Noskiewicz, Heiko Maaß, Ralf Mikut, Veit Hagenmeyer

Karlsruhe Institute of Technology, E-Mail: ralf.mikut@kit.de

Abstract

Modern buildings increasingly integrate local energy generation, consumption, and storage, often involving multiple energy carriers such as electricity and thermal energy. This complexity creates opportunities for cost-efficient operation based on accurate forecasts of key energy parameters such as the cooling demand. However, generating such forecasts on a building-level can be challenging, as individual buildings likely exhibit highly specific consumption patterns and often offer only limited historical data.

Time Series Foundation Models (TSFMs) offer a promising solution to this problem due to their ability to generalise across forecasting tasks and adapt to new domains with minimal data via fine-tuning. This study evaluates three state-of-the-art TSFMs (MOIRAI, MOIRAI-MoE and Chronos) in both zero-shot and fine-tuned settings. The models are tested on real-world energy consumption data from a split-type cooling unit used in a server room, spanning a period of less than four months. Outdoor air temperature is included as a covariate to assess its impact on prediction accuracy. Results are compared to two baseline models.

Our findings show that Chronos, when incorporating outdoor air temperature, achieves a substantial improvement in forecasting accuracy for three-day ahead forecasts, reducing the Mean Absolute Percentage Error (MAPE) to as low as 2.57 %. In contrast, MOIRAI and MOIRAI-MoE show no significant benefit from the inclusion of temperature information. Overall, the study demonstrates that TSFMs represent a promising alternative for cooling demand forecasting

in buildings, particularly in scenarios where only limited historical data is available.

1 Introduction

The heating and cooling sector accounts for more than 50 % of Germany's total final energy consumption, making it a central focus of the energy transition [17, 21]. Accurate demand forecasting is essential for efficient energy management, yet a major challenge stems from the limited or non-existent historical consumption data in most buildings. This impedes the training of building-specific forecasting models and constrains their practical applicability. In recent years, TSFMs have emerged as a promising solution to address the challenge of forecasting under limited historical data [11]. These models are pre-trained on diverse collections of time series and can be applied to a wide range of forecasting tasks without task-specific training. They can also be adapted to domain-specific problems with only limited data through fine-tuning. While this make TSFMs well suited for time series forecasting under data-scarce conditions, their application for cooling demand forecasting remains largely unexplored. A major challenge is that external covariates, such as outdoor temperature, strongly influence the efficiency of cooling systems and thus the cooling demand, yet strategies for effectively integrating such auxiliary information into TSFMs remain underdeveloped. This study addresses this gap by systematically evaluating the potential of TSFMs for cooling demand prediction with the integration of covariate information. Three TSFMs are tested on a real-world time series of a split-type AC unit at both 15-minute and hourly resolutions. The models are compared in zero-shot and fine-tuned settings, with and without outdoor temperature as a covariate, and using two context lengths to assess the impact of model configurations on prediction performance. Two baseline models are included for benchmarking.

This study is structured as follows. Section 2 provides an overview of current studies on chiller energy consumption prediction. Section 3 describes the methodology, including the dataset, preprocessing steps, the models used, and the experimental design. Section 4 presents the results, which are discussed in Section 5. Section 6 summarises the findings and outlines future research directions.

2 Related Work

In the prediction of chiller systems, a distinction is commonly made between physics-based and data-driven models [24]. As this work focuses on data-driven approaches, the following provides an overview of recent research in this field. Tien et al. [20] present a comprehensive survey of chiller system prediction across spatial and temporal scales, highlighting the diversity and rapid development of approaches. Wang et al. [22] emphasise the challenges posed by high variability and non-stationarity and propose a hybrid method combining wavelet decomposition, Long Short-Term Memory (LSTM) networks, RAdam optimisation, and Pearson correlation-based feature selection [22]. Their results, based on factory data from Xiamen, show improved prediction accuracy compared to baseline models such as linear regression and feed-forward networks. Sulaiman et al. [18] apply a fixed forward Neural Network (NN) optimised with the Teaching-Learning-Based Optimisation (TLBO) metaheuristic to predict energy consumption in commercial buildings, demonstrating significant performance gains. In subsequent work, Sulaiman et al. [19] extend this approach using Kolmogorov-Arnold Networks (KANs), which outperform both TLBO-based and conventional NN models. While previous studies focus on electrical consumption, others address the prediction of the Coefficient of Performance (COP), which requires estimating the ratio of cooling capacity to electrical input. Ho et al. [6] employ a hybrid ARIMA-regression model and identify the part load ratio as most relevant, while Deng et al. [5] use a dynamic graph convolutional network and highlight cooling water return temperature as the key variable. These differing findings underline the system-specific nature of COP modelling.

3 Methodology

3.1 Data and Pre-Processing

Cooling Demand Time Series The cooling demand time series contains electrical energy consumption data of a `Panasonic R410A` split type air conditioning (AC) unit in the server room of the (office) Building 445 at Karlsruhe Institute of Technology (KIT) and covers February 23 to May 31, 2025. Based on the electrical energy consumption of the AC unit, the actual cooling demand can be calculated with the temperature dependent COP. The raw series is non-equidistant with a temporal resolution of 4 seconds to 3 minutes and 4 seconds and contains several missing segments. Gaps up to 30 minutes are filled via linear interpolation, while longer gaps are imputed with the value observed 135 minutes earlier, corresponding to one hysteresis-control cycle. A more detailed description of the tested imputation methods, as well as the data pre-processing and analysis, can be found in [10].

The series exhibits a daily pattern and a recurring 135-minute cycle from the unit's hysteresis control. As a prediction of the shorter cycle is not necessary to predict the overall cooling demand, the series is smoothed using a rolling window with a window size of 135 minutes. It is further resampled to 15-minute and hourly resolution. Finally, the time series is divided into training, validation, and test sets, with the training set covering data until March 31, 2025, the validation set comprising April 2025, and the test set comprising May 2025. The processed time series is shown in Figure 1. Despite the short time span, seasonal effects causing rising average outdoor air temperatures introduce a concept drift, complicating the forecasting task.

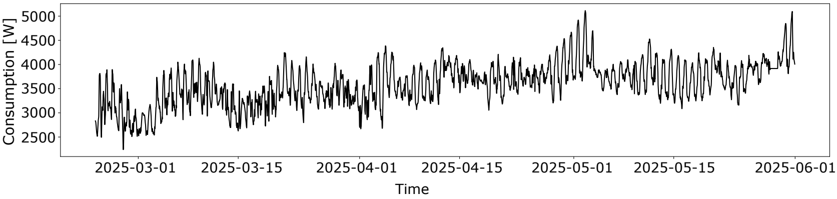

Figure 1: Pre-processed cooling demand time series in **15 minute** resolution.

Outdoor Air Temperature Time Series The Outdoor Air Temperature time series contains measurements recorded every 15 minutes by a weather station on the roof of Building 445. The recording period and the division into training, validation, and test sets are identical to those of the cooling demand series. Different imputation methods were evaluated for the missing segments in the raw series. The selected strategy fills gaps of up to four hours via linear interpolation, while longer gaps are imputed using measurements from the nearby Station 4177 in Rheinstetten provided by the German Weather Service (DWD)[1]. Further information on the compared imputation methods is provided in [10]. Finally, the 15-minute series is resampled to derive an additional hourly-resolution series. Figure 2 shows the relationship between the chiller's energy consumption and the outdoor temperature, with each point colour-coded according to the time of day. At lower temperatures, the relationship exhibits considerably more variance compared to higher temperature ranges.

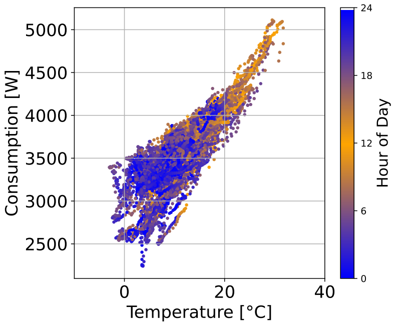

Figure 2: Energy consumption versus outdoor Temperature.

3.2 Compared Models

MOIRAI [23] is built on an encoder-only Transformer and uses a patch-based approach, dividing each time series into non-overlapping segments that are transformed via Multi-Patch-Size Input Projection layers, with patch sizes chosen according to the series frequency (e.g., hourly, daily). MOIRAI is developed as a multivariate model, supporting the integration of both past and future covariates. It is trained on the Large-scale Open Time Series Archive (LOTSA) dataset [23], containing around 27 billion observations from over

[1] Data from DWD Station 4177: `https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/air_temperature/recent/`

100 univariate and multivariate time series. However, the included covariates are limited to past covariates, with no future covariates provided. The model is available in Small, Base, and Large variants with 14 to 311 million parameters. This study employs `MOIRAI-1.1R-small`. MOIRAI is pre-trained with the Negative Log-Likelihood loss [23]. As fine-tuning on the cooling demand time series showed no improvement with this loss function, it is changed to the packed Mean Squared Error (MSE). More details on this can be found in [10]. MOIRAI generates multiple forecast samples, and point predictions are computed as the mean of these samples, as also done in [3].

MOIRAI-MoE [12] is a decoder-only Transformer-based TSFM that extends the MOIRAI architecture. Instead of using multiple heuristic projection layers for different frequencies, it employs a single projection layer and addresses time series variability through a sparse Mixture-of-Experts (MoE) mechanism within the Transformer layers. As MOIRAI, MOIRAI-MoE is developed as a multivariate model, trained on the LOTSA dataset and is available in two sizes (Small and Base). In this study, the `MOIRAI-MoE-1.0R-small` version with 117 million parameters is applied [12]. MOIRAI-MoE generates multiple forecast samples, and point predictions are computed as the median of these samples. This approach is used because MOIRAI-MoE occasionally produces extreme outlier samples on 15-minute data, which are effectively ignored when using the median. Further details can be found in [10].

Chronos [1] is a framework for probabilistic time series forecasting that adapts language models by scaling and quantizing values into a fixed vocabulary for tokenization. It is primarily based on the T5 encoder-decoder transformer architecture [16] but can also be applied to decoder-only architectures such as GPT-2 [15]. Originally developed as a univariate model, covariates are incorporated via a separate regression model predicting the target variable, while Chronos models the residuals between predictions and actual values. Chronos is available in different model sizes, and in this study, the `Chronos-Bolt-Mini` variant with 21 million parameters is used. Chronos generates multiple forecast samples, with point predictions computed by default as the mean of these samples [1].

Linear Regression As the first baseline model, a linear regression is trained on the training set using the relationship between the consumption of the AC unit and the outdoor temperature. The prediction of the target variable is then obtained by applying the model to the outdoor temperature values within the forecasting window.

Last Day As a second baseline model, the value from the previous day at the same time is used for each corresponding time step within the forecasting horizon. This approach is motivated by the pronounced daily patterns in the data.

3.3 Experiment Design

All models are employed to generate rolling forecasts of the cooling demand time series for the next three days, using a stride of one. Two context lengths, one week and four weeks, are used and compared. Since no fine-tuning script for MOIRAI-MoE is available at the time of this study, this model is applied in a zero-shot (ZS) setting only. In addition to zero-shot evaluation, MOIRAI and Chronos are also fine-tuned (FT) on the training set. The model configurations used for fine-tuning are presented in Tables 1 and 2 and are either recommended in the respective paper or were determined experimentally (see [10]). For all fine-tuned setups, the context length is restricted to one week, as extending it to four weeks would drastically reduce the size of the training dataset.

All models are evaluated both with and without the integration of outdoor air temperature as covariate, in order to quantify the impact of additional information. While the MOIRAI paper suggests support for incorporating future covariates during training, this functionality is not implemented in the available fine-tuning script. Consequently, only Chronos is fine-tuned with covariate information.

Table 1: MOIRAI fine-tuning settings

Parameter	Value
Learning Rate	1×10^{-6}
Batch Size	32
Patch Size[2]	32
Batches per Epoch	100
Patience	100
Loss Function	PackedMSE
Optimizer	AdamW

Table 2: Chronos fine-tuning settings

Parameter	Value
Learning Rate	1×10^{-6}
Batch Size	32
Max. Training Steps	10,000
Regression Model[3]	Linear
Loss Function	Cross-entropy
Optimizer	AdamW

3.4 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 the predicted values $\hat{y}_i$, with N denoting the number of data samples times the forecast length[4]. Taking each error without sign prevents over- and underestimations from cancelling out. MAE is computationally efficient, relatively robust to outliers, and easy to interpret due to its consistent units. However, it is non-differentiable at zero, complicating gradient-based optimisation [8].

Mean Absolute Percentage Error

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{2}$$

The Mean Absolute Percentage Error (MAPE) measures forecasting accuracy by averaging the absolute percentage differences between predicted and actual

[2] This patch size is also used during inference.

[3] The linear regression model is applied for all model settings with covariate integration (zero-shot and fine-tuned).

[4] The notation of N, y_i and $\hat{y}_i$ will remain the same for all error metrics and will therefore not be explained again in the following.

values. Errors are relative to the actual value and taken without sign. MAPE is intuitive and allows easy comparison across time series. However, it is sensitive to small denominators, as values near zero can produce undefined or excessively large errors [8].

Mean Squared Error

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

The MSE measures the average squared deviation between the actual value and the prediction. By squaring errors, it smooths the gradient for small errors, aiding optimisation. However, this also makes MSE highly sensitive to outliers, which can significantly affect model performance and cause large gradient jumps during backpropagation. The MSE is always positive and ranges from 0 to infinity [8].

Continuous Ranked Probability Score

$$\mathrm{CRPS}(F,y) = \int_{\mathbb{R}} (F(z) - \mathbf{1}_{\{y \leq z\}})^2 \, \mathrm{d}z \tag{4}$$

The Continuous Ranked Probability Score (CRPS) is a scoring rule for probabilistic predictions and is defined as in Equation 4. Here, $F(z)$ denotes the predicted cumulative distribution function, $\mathbf{1}_{\{y \leq z\}}$ is the indicator function that equals 1 if $y \leq z$ and 0 otherwise, and $\mathbb{R}$ represents the domain of all real numbers over which the integral is evaluated [9] [13]. To reduce computational effort, the CRPS is commonly approximated as

$$\mathrm{CRPS}(F,y) \approx \frac{2}{N}\sum_{i=1}^{N} \Lambda_{\alpha_i}(q_{\alpha_i}, y), \tag{5}$$

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

$$\Lambda_\alpha(q,y) = (\alpha - \mathbf{1}_{\{y<q\}})(y-q). \tag{6}$$

This approximation is used, for example, in [4] and in [23], and is applied in this study.

3.5 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*, and *Matplotlib*.

4 Experiment Results

Table 3: Forecasting Results of the **15 minute** *Cooling Demand* time series on the test set.

Model	**Cov.**	**Context**	**Type**	**MSE**	**MAPE**	**MAE**	**CRPS**
		[Weeks]		[W^2]	[%]	[W]	[W]
Lin. Regression	Yes	-	-	62,866.47	5.42	214.53	-
Last Day	No	-	-	81,026.19	5.42	208.69	-
MOIRAI	No	1	ZS	131,741.85	7.56	284.33	220.89
MOIRAI	No	4	ZS	94,476.82	6.51	245.08	187.00
MOIRAI	Yes	1	ZS	91,996.97	6.36	241.90	189.68
MOIRAI	Yes	4	ZS	91,862.53	6.27	240.02	186.06
MOIRAI-MoE	No	1	ZS	125,425.10	7.35	280.98	232.11
MOIRAI-MoE	No	4	ZS	131,008.77	7.49	278.64	227.19
MOIRAI-MoE	Yes	1	ZS	101,789.40	6.66	252.11	237.80
MOIRAI-MoE	Yes	4	ZS	107,645.66	6.62	249.50	218.22
Chronos	No	1	ZS	71,034.98	5.15	195.99	167.68
Chronos	No	4	ZS	69,672.95	4.98	189.32	162.04
Chronos	Yes	1	ZS	18,477.34	2.73	103.03	89.17
Chronos	Yes	4	ZS	**17,956.99**	**2.67**	**100.50**	**86.43**
MOIRAI	No	1	FT	82,778.91	5.89	222.62	168.31
Chronos	No	1	FT	73,438.70	5.13	195.44	168.69
Chronos	Yes	1	FT	20,095.80	2.88	108.41	93.87

Tables 3 and 4 present the test set results for the 15-minute and hourly time series. On the 15-minute time series (Table 3), Chronos achieves the lowest error across all evaluation metrics in the zero-shot setting when using a four-week context and the temperature covariate. Shortening the context to one week under otherwise identical settings results in only a slightly higher error. Compared to

Table 4: Forecasting Results of the **hourly** *Cooling Demand* time series on the test set.

Model	**Cov.**	**Context** [Weeks]	**Type**	**MSE** [W^2]	**MAPE** [%]	**MAE** [W]	**CRPS** [W]
Lin. Regression	Yes	-	-	62,008.75	5.39	213.41	-
Last Day	No	-	-	79,763.28	5.37	207.12	-
MOIRAI	No	1	ZS	66,448.49	5.10	196.72	153.25
MOIRAI	No	4	ZS	60,577.64	4.92	189.15	146.59
MOIRAI	Yes	1	ZS	59,829.38	4.87	188.52	147.26
MOIRAI	Yes	4	ZS	55,855.61	4.69	180.85	142.14
MOIRAI-MoE	No	1	ZS	76,899.71	5.36	207.35	165.91
MOIRAI-MoE	No	4	ZS	65,786.63	5.03	194.23	152.68
MOIRAI-MoE	Yes	1	ZS	63,030.09	4.92	189.85	151.20
MOIRAI-MoE	Yes	4	ZS	59,359.82	4.81	185.51	145.52
Chronos	No	1	ZS	57,716.64	4.72	179.55	143.17
Chronos	No	4	ZS	52,007.28	4.57	173.77	137.56
Chronos	Yes	1	ZS	17,943.55	2.70	101.90	82.90
Chronos	Yes	4	ZS	**16,225.05**	**2.57**	**96.81**	**78.05**
MOIRAI	No	1	FT	75,852.56	5.56	211.82	160.01
Chronos	No	1	FT	59,535.35	4.80	182.18	145.20
Chronos	Yes	1	FT	17,944.00	2.69	101.68	81.98

the linear regression baseline model, the best Chronos configuration reduces the MSE by up to 71 %. Fine-tuning Chronos results in a slightly decreased performance, compared to the respective zero-shot counterpart. All Chronos models without covariate integration perform significantly worse and fail to surpass the linear regression baseline in both zero-shot and fine-tuned settings.

MOIRAI achieves its best performance in the fine-tuned setting using a one-week context without covariates, although it does not outperform either baseline model. In the zero-shot setting, performance with a one-week context and no covariates is significantly worse, but increasing the context length improves the results. Adding temperature as a covariate leads to a substantial performance gain for the one-week context and a modest improvement for the four-week context. Nonetheless, none of the MOIRAI configurations surpass the weakest Chronos model.

MOIRAI-MoE does not outperform either baseline model in any configuration and performs on par with the worst MOIRAI model. Increasing the context length reduces the error only marginally while incorporating the covariate noticeably lowers the error. Still, the performance remains worse than that of the best MOIRAI model and all Chronos configurations.

The results on the hourly time series (Table 4) largely confirm those observed on the 15-minute data. Chronos achieves the best performance in the zero-shot setting with covariates and a four-week context, closely followed by the same configuration with a one-week context and the fine-tuned model with a one-week context and covariate. All configurations without covariate perform significantly worse, but unlike in the 15-minute setting, they still outperform both baseline models. Interestingly, fine-tuning both with and without covariates leads to a decrease in model performance.

MOIRAI achieves its best performance in the zero-shot setting using a four-week context combined with covariates. All zero-shot configurations outperform both baseline models, and within this setting, extending the context length and incorporating temperature as a covariate both contribute positively to forecasting accuracy. Compared to Chronos, all zero-shot MOIRAI settings reach similar performance levels to Chronos configurations that exclude covariates. In contrast, fine-tuning with a one-week context results in a significant drop in accuracy, aligning with the degradation observed in Chronos under the same conditions.

Moirai-MoE performs worse on average than both MOIRAI and Chronos. In its best configuration, which uses a four-week context and includes covariates, it outperforms both baseline models. As observed with MOIRAI, extending the context length and incorporating temperature as a covariate lead to slight improvements in model performance.

Table 5: Fine-tuning duration of the **15 min** time series.

Model	**Cov.**	**Context**	**Learning Rate**	**Total**	**Best**	**Best**
		[Weeks]		[min]	[min]	[Steps]
MOIRAI	No	1	1×10^{-6}	65.35	1.40	300
Chronos	No	1	1×10^{-6}	7.79	0.37	600
Chronos	Yes	1	1×10^{-6}	6.55	1.35	2,200

Table 6: Fine-tuning duration of the **hourly** time series.

Model	Cov.	Context	Learning Rate	Total	Best	Best
		[Weeks]		[min]	[min]	[Steps]
MOIRAI	No	1	1×10^{-6}	77.97	11.14	1,700
Chronos	No	1	1×10^{-6}	7.20	0.00	100
Chronos	Yes	1	1×10^{-6}	6.85	0.95	1,500

Tables 5 and 6 present the training times and number of training steps for all evaluated model configurations. Notably, both models complete very few training steps on the 15-minute time series when covariates are not included, while Chronos fails to show any training improvement in the corresponding configuration of the hourly time series. The reasons for this behaviour will be discussed in the following section. It is also noteworthy that MOIRAI requires significantly more training time than Chronos, even when the number of training steps is similar.

5 Discussion

The following section presents a detailed analysis of various effects to support the interpretation of the results in Tables 3 and 4. To illustrate these effects, figures showing forecasts from different model configurations are provided and compared.

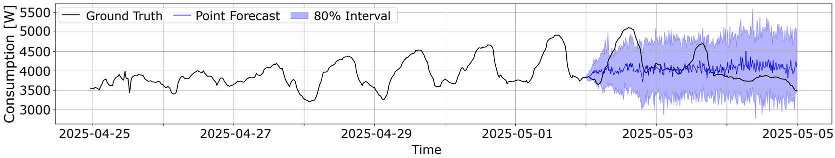

Figure 3: **Resolution:** 15 min; **Model:** MOIRAI; **Covariate:** No; **Type:** ZS.

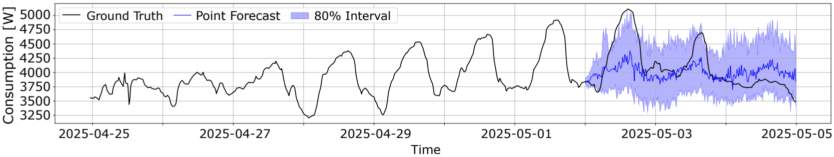

Figure 4: **Resolution:** 15 min; **Model:** MOIRAI; **Covariate:** No; **Type:** FT.

Resolution-Specific Effects (15-Minute Data) Figure 3 and Figure 4 show predictions of the 15-minute time series generated by MOIRAI in the zero-shot and fine-tuned settings, respectively. In both plots, the context length is set to one week and no covariate is used. In the zero-shot scenario, both MOIRAI and MOIRAI-MoE fail to capture the daily pattern of the time series, resulting in the high errors presented in Table 3. Instead, their predictions resemble noise fluctuating around the mean of the input window, lacking any meaningful temporal structure. The limited performance of the two models is likely due to the small proportion of 15-minute data in the LOTSA dataset, which contains only four time series at this resolution [23]. However, this finding contradicts the claim by the authors that MOIRAI-MoE, through its Mixture of Experts approach, should enable frequency-independent forecasting [12]. Accordingly, the daily pattern of the time series should at least be captured in the predictions, since such patterns are present in the model's training data, though at a different temporal resolution.

In contrast to the zero-shot setting, the fine-tuned MOIRAI model (Figure 4) is able to partially learn the daily pattern of the time series, resulting in a slightly lower error. However, the performance improvement achieved through fine-tuning is relatively modest and the model fails to predict the magnitude of the series.

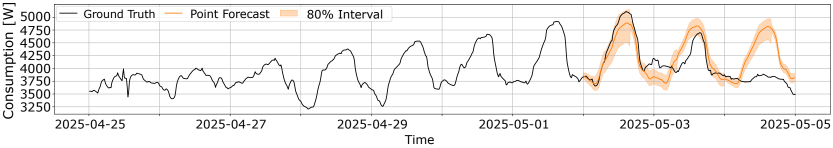

Figure 5: **Resolution:** 15 min; **Model:** Chronos; **Covariate:** No; **Type:** ZS.

Figure 5 displays a prediction of the same forecast window generated by Chronos in the zero-shot setting without covariates. In contrast to MOIRAI and MOIRAI-MoE, Chronos is able to capture the underlying daily pattern of the time series, whereby the predictions are strongly influenced by the temporal dynamics within last days of the input window. This leads to a substantial error on the third day of the forecast horizon, when the time series encounters a structural shift due to a change of outdoor temperature that cannot be inferred from the input data. As a result, MOIRAI and MOIRAI-MoE achieve lower prediction errors on that specific day. When evaluating model performance across the entire test set (see Table 3), this leads to only a small advantage for the Chronos configurations without covariates over the MOIRAI and MOIRAI-MoE variants in terms of average error. However, such aggregated metrics obscure important differences in model behaviour, as the closer examination of individual forecast windows reveals that Chronos is capable of learning and reproducing meaningful temporal patterns, whereas the predictions generated by MOIRAI and MOIRAI-MoE lack clear structure. Notably, despite capturing temporal dynamics well, Chronos tends to be overconfident in its forecasts, as the predicted uncertainty band appears too narrow to adequately reflect the true variability in the data.

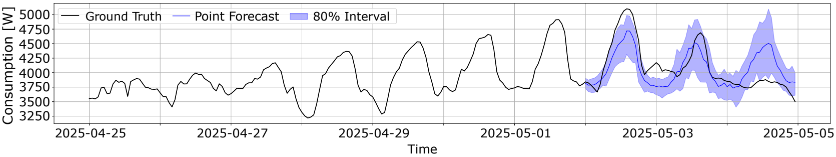

Figure 6: **Resolution:** Hourly; **Model:** MOIRAI; **Covariate:** No; **Type:** ZS.

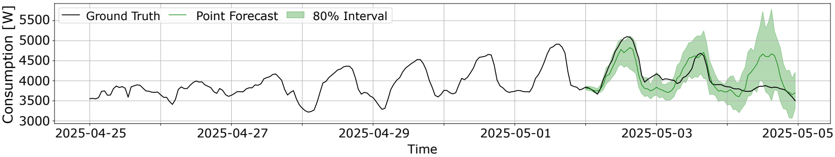

Figure 7: **Resolution:** Hourly; **Model:** MOIRAI-MoE; **Covariate:** No; **Type:** ZS.

Zero-Shot Performance Figures 6 and 7 display predictions generated by MOIRAI and MOIRAI-MoE in zero-shot setting without covariates, applied to the hourly-resolved version of the time series. Both models are now able to capture and reproduce the daily pattern, resulting in a significantly reduced prediction error compared to the 15-minute resolution. However, the models encounter the same issue as Chronos on the third day of the forecast horizon, where an abrupt change in the time series leads to noticeable deviations from the ground truth. Chronos benefits slightly from the lower temporal resolution, likely because resampling filters out high-frequency noise, leading to a modest reduction in prediction error.

Impact of Fine-Tuning As the results in Table 4 demonstrate, fine-tuning MOIRAI and Chronos without the inclusion of covariates does not lead to improved performance, but rather results in a decline in prediction accuracy. This suggests that the time series lacks meaningful intrinsic autoregressive structure, which is further supported by the very low number of training steps observed for the model variants without covariates (see Tables 5 and 6). Since the efficiency of the AC unit, and therefore its energy consumption, is strongly affected by temperature, which follows a diurnal cycle, the visible autocorrelations reflect external periodicity rather than genuine internal dependencies. Without access to temperature data, the model cannot effectively exploit this structure, making fine-tuning without exogenous inputs largely ineffective.

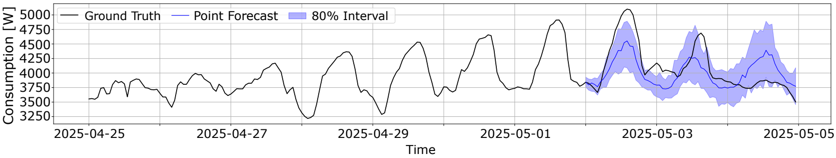

Figure 8: **Resolution:** Hourly; **Model:** MOIRAI; **Covariate:** Yes; **Type:** ZS.

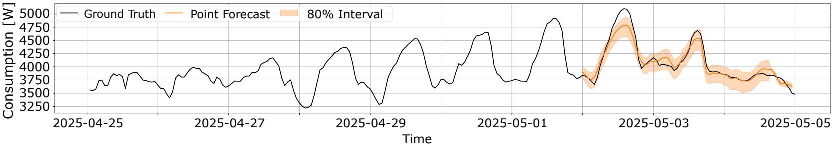

Figure 9: **Resolution:** Hourly; **Model:** Chronos; **Covariate:** Yes; **Type:** ZS.

Effect of Covariate Integration Figures 8 and 9 present predictions from MOIRAI and Chronos in a zero-shot setting, using one week of context and temperature as a covariate. Compared to the equivalent setup without covariates, MOIRAI's forecasts exhibit noticeable differences, accompanied by a slight reduction in average prediction error across the entire test period. A similar effect is observed for MOIRAI-MoE, indicating that both models incorporate the provided future covariate information into their forecasts despite not having encountered such covariates during pre-training or fine-tuning. Nevertheless, the exact way in which this information is utilised remains unclear. While incorporating the covariate results in a slightly improved performance on the cooling demand series, both models show decreased performances on other time series tested in [10].

In contrast to MOIRAI, Chronos achieves a significant improvement in forecasting accuracy when incorporating the temperature covariate. Several noteworthy observations emerge when evaluating model performance. First, the comparison to the linear baseline is of particular interest, as Chronos incorporates the covariate using a linear regression model that is fit on the training set. Chronos considerably outperforms this baseline, suggesting that it captures additional patterns within the context length, besides the temperature dependency. These patterns may arise from latent relationships involving other variables, such as interactions between the cooling system's energy consumption and the server room load, which could induce autoregressive structures in the target time series. Secondly, fine-tuning Chronos with covariate integration leads to decreased performance compared to the zero-shot setting, showing the model's limited ability to generalise the learned temperature dependency across different seasonal conditions. The linear regression model used to incorporate the temperature covariate is trained on data from the colder months of February and March. During this period, the correlation between outdoor temperature and cooling energy consumption exhibits higher variance (see Figure 2), indicating a less stable linear relationship. In contrast, the test period in May features warmer temperatures and a more consistent correlation pattern, requiring the model to extrapolate beyond the temperature range seen during training. As fine-tuning is conducted on the residuals between the ground truth and the linear regression model's predictions, forecasting accuracy may decline if the residual patterns learned during training do not generalise to the test period. As a result, the model may overfit to residual patterns specific to the training data, which are not representative of the test set.

6 Conclusion

Directly forecasting cooling demand rather than estimating it as part of total building energy consumption provides several advantages, as it enables more targeted energy management, can act as a proxy for larger cooling systems, and offers a lever for demand-side management or grid-friendly cooling strategies. Against this background, this study investigates the use of TSFMs for cooling demand forecasting with outdoor air temperature as covariate. Three TSFMs and two reference models are evaluated on real-world data with less than four months of history, reflecting realistic conditions where only limited data is available. Since none of the models handle preprocessing or missing values internally, and no alternative TSFM that does is available to date, the data is first preprocessed and missing values filled. Afterwards, three-day-ahead forecasts are generated with different model configurations. The results show that Chronos benefits significantly from covariate integration, clearly outperforming baseline models. In contrast, MOIRAI and MOIRAI-MoE do not gain from future covariate information, most likely due to the absence of such data in their pre-training corpora. This highlights both the potential of TSFMs for cooling demand forecasting and the need for improved approaches to incorporating covariates.

This study provides a foundation for further research. Future experiments on a larger dataset, where training data better reflects the temperature ranges of the validation and test sets, could enable fine-tuning with covariates to outperform zero-shot performance. Moreover, while the current evaluation uses measured temperature data as covariate input, this does not reflect a realistic application scenario. Using weather forecast data as covariate input would allow assessment of its impact on model performance, introducing additional uncertainty. Fine-tuning the models on weather forecast data is also of interest to investigate its effect on probabilistic predictions under realistic operating conditions. Furthermore, Chronos was originally developed as a univariate model and incorporates covariates via a tabular regression approach, meaning it does not directly learn relationships between variables. However, recent advances such as ChronosX [14], which extends Chronos to capture inter-variable dependencies, and COSMIC [2], a multivariate pre-trained model supporting zero-shot forecasting with covariates, have emerged as promising alternatives. Consequently, comparisons with these models, as well as with TabPFN [7], which demonstrates strong performance even on relatively small datasets, is of particular interest.

References

[1] 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. Chronos: Learning the Language of Time Series, 2024. URL: `https://arxiv.org/abs/2403.07815`.

[2] A. Auer, R. Parthipan, P. Mercado, A. F. Ansari, L. Stella, B. Wang, M. Bohlke-Schneider, and S. S. Rangapuram. Zero-Shot Time Series Forecasting with Covariates via In-Context Learning, 2025. URL: `https://arxiv.org/abs/2506.03128`.

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

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

[5] Q. Deng, Z. Chen, W. Zhu, Z. Li, Y. Yuan, and W. Gui. A performance prediction method for on-site chillers based on dynamic graph convolutional network enhanced by association rules. *Building Simulation*, 17:1213–1229, 2024. doi:10.1007/s12273-024-1136-3.

[6] W.T. Ho and F.W. Yu. Predicting chiller system performance using ARIMA-regression models. *Journal of Building Engineering*, 33:101871, 2021. doi:10.1016/j.jobe.2020.101871.

[7] S. B. Hoo, S. Müller, D. Salinas, and F. Hutter. From Tables to Time: How TabPFN-v2 Outperforms Specialized Time Series Forecasting Models, 2025. URL: https://arxiv.org/abs/2501.02945.

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

[9] A. Jordan, F. Krüger, and S. Lerch. Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90:1–37, 2019. doi:10.18637/jss.v090.i12.

[10] A. Kreusel. Integrating Covariates into Time Series Foundation Models for Building Energy Forecasting. Master's thesis, Karlsruhe Institute of Technology, 2025.

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

[12] X. Liu, J. Liu, G. Woo, T. Aksu, Y. Liang, R. Zimmermann, C. Liu, S. Savarese, C. Xiong, and D. Sahoo. Moirai-MoE: Empowering Time

Series Foundation Models with Sparse Mixture of Experts, 2024. URL: https://arxiv.org/abs/2410.10469.

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

[14] 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. ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables, 2025. URL: https://arxiv.org/abs/2503.12107.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

[17] Statista. Verteilung des Endenergieverbrauchs zu Wärmezwecken nach Verbrauchssektoren im Jahr 2022, 2024. Accessed: 19. January 2025. URL: https://de.statista.com/statistik/daten/studie/1463468/umfrage/verteilung-waermeverbrauch-nach-sektoren/.

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

[19] M. H. Sulaiman, Z. Mustaffa, M. S. Saealal, M. M. Saari, and A. Z. Ahmad. Utilizing the Kolmogorov-Arnold Networks for chiller energy consumption prediction in commercial building. *Journal of Building Engineering*, 96:110475, 2024. doi:10.1016/j.jobe.2024.110475.

[20] P. W. Tien, S. Wei, J. Darkwa, C. Wood, and J. Kaiser Calautit. Machine Learning and Deep Learning Methods for Enhancing Building Energy

Efficiency and Indoor Environmental Quality – A Review. *Energy and AI*, 10:100198, 2022. doi:10.1016/j.egyai.2022.100198.

[21] Umweltbundesamt. Energieverbrauch für fossile und erneuerbare Wärme, 2024. Accessed: 19. January 2025. URL: https://www.umweltbundesamt.de/daten/energie/energieverbrauch-fuer-fossile-erneuerbare-waerme.

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

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

[24] F.-W. Yu, W.-T. Ho, and C.-F. Jeff Wong. Improved energy management of chiller system with AI-based regression. *Applied Soft Computing*, 150:111091, 2024. doi:10.1016/j.asoc.2023.111091.