Are Time Series Foundation Models Ready to Revolutionize Predictive Building Analytics?

Ozan Baris Mulayim*
omulayim@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Pengrui Quan*
prquan@g.ucla.edu
University of California, Los Angeles
Los Angeles, CA, USA

Liying Han
liying98@ucla.edu
University of California, Los Angeles
Los Angeles, CA, USA

Xiaomin Ouyang†
xmouyang@cse.ust.hk
Hong Kong University of Science and Technology
Hong Kong SAR, Hong Kong

Dezhi Hong‡
hondezhi@amazon.com
Amazon
Seattle, WA, USA

Mario Bergés§
marioberges@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Mani Srivastava§
mbs@ucla.edu
University of California, Los Angeles
Los Angeles, CA, USA

ABSTRACT

Recent advancements in large language models have spurred significant developments in Time Series Foundation Models (TSFMs). These models claim great promise in performing zero-shot forecasting without the need for specific training, leveraging the extensive "corpus" of time-series data they have been trained on. Forecasting is crucial in predictive building analytics, presenting substantial untapped potential for TSFMS in this domain. However, time-series data are often domain-specific and governed by diverse factors such as deployment environments, sensor characteristics, sampling rate, and data resolution, which complicates generalizability of these models across different contexts. Thus, while language models benefit from the relative uniformity of text data, TSFMs face challenges in learning from heterogeneous and contextually varied time-series data to ensure accurate and reliable performance in various applications. This paper seeks to understand how recently developed TSFMs perform in the building domain, particularly concerning their generalizability. We benchmark these models on three large datasets related to indoor air temperature and electricity usage. Our results indicate that TSFMs exhibit marginally better performance compared to statistical models on unseen sensing modality and/or patterns. Based on the benchmark results, we also provide insights for improving future TSFMs on building analytics.

*Both authors contributed equally to this research.
†The work was done when the author was at UCLA.
‡Work unrelated to Amazon
§Authors hold concurrent appointments as Amazon Scholars, and as Professors at their respective universities, but work in this paper is not associated with Amazon.

BUILDSYS '24, November 7–8, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0706-3/24/11.
https://doi.org/10.1145/3671127.3698177

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

time series foundation models, building analytics, forecasting foundation models

ACM Reference Format:
Ozan Baris Mulayim, Pengrui Quan, Liying Han, Xiaomin Ouyang, Dezhi Hong, Mario Bergés, and Mani Srivastava. 2024. Are Time Series Foundation Models Ready to Revolutionize Predictive Building Analytics?. In *The 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BUILDSYS '24), November 7–8, 2024, Hangzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3671127.3698177

1 INTRODUCTION

Building on the rapid advancement of large language models, time-series foundation models (TSFMs) have also experienced significant development as of late. Recent advancements in TSFMs via pretraining on large and diverse time-series datasets, such as `MOMENT` [8] and `TimesFM` [4], have shown promising results across various applications. Their main promise lies in their ability to perform zero-shot forecasting without requiring specific training due to the vast "corpus" of time-series data they are trained on.

However, the presence of a large corpus alone might not warrant these models' ability to generalize effectively. Unlike language data, which can be effectively leveraged due to uniform grammatical rules and consistent linguistic structures, time-series data is inherently *heterogeneous*. Aggregated time-series datasets often consist of sequences from vastly disparate sources, and thus the approaches to representing language as tokens uniformly may not work for various numerical time series data [9, 14]. These differences pose significant challenges for TSFMs, as they must learn to adapt to a wide range of temporal and contextual attributes unique to each

dataset. Consequently, while the vast corpus of time-series data provides a rich resource, it might not guarantee the same level of generalizability seen in language models.

The generalizability of these models is further challenged as there are often unique confounding variables to consider in time-series data. With buildings as an example: (1) Controlled Dynamics: temperature and electricity measurements are influenced by the operation of HVAC systems, such as duty-cycling schedule and setpoint changes [10], and by occupancy patterns, as buildings are operated differently when they are occupied; (2) Natural Dynamics: external factors such as outdoor air temperature, solar irradience level, and activities of occupants [5] introduce natural multivariate dynamics present in building datasets. Both controlled and natural dynamics introduce significant complexity to predictive modeling in this domain, presenting challenges not typically encountered in language processing.

Even though these complex dynamics present significant challenges, TSFMs could still potentially revolutionize building analytics, which are currently hindered by the ad-hoc model development across diverse buildings. Together with robustness to unfamiliar datasets, the applicability to real-life building analytics would entail stable performance across various metrics and conditions, and handle the diverse confounding effects present in building control scenarios. However, current research on predictive analytics for buildings is still largely based on physics-based and conventional machine-learning models [3, 18], and there is no comprehensive evaluation of the *readiness* of TSFMs for these analytic tasks. We take the first step in addressing this gap by conducting a multifaceted assessment of their readiness across varying context and prediction durations and datasets. Our evaluation framework, therefore, focuses on *generalizability* across datasets and modalities. Specifically, we focus on univariate time-series forecasting with TSFMs that can make zero-shot predictions of two key physical values in predictive building management: electricity usage and indoor air temperature. Despite the known benefits of including covariates in predictions for building analytics, we focus on univariate predictions due to the simple fact that all pre-trained TSFMs available today can perform univariate forecasting while only a few allow including covariates.

We arrived at the following findings through evaluation of TSFMs on forecasting tasks for buildings: (1) Dataset-level familiarity: TSFMs outperform statistical models only on previously seen electricity datasets, with marginal improvement on unseen ones. (2) Modality-level familiarity: On large-scale indoor air temperature data, TSFMs perform better over longer durations, while statistical models, particularly `AutoARIMA`, perform better for shorter durations and nearly match the performance of the top TSFMs on unseen datasets for seen sensor modalities.

2 EXISTING TIME-SERIES FOUNDATION MODELS

A TSFM is a large model pre-trained on massive amounts of time-series data. TSFMs are designed to learn about general time-series patterns and leverage zero-shot or transfer learning techniques to perform time-series analysis on previously unseen datasets [4, 8]. As this is a nascent and evolving field with most models released in 2024, we first review existing TSFMs and summarize their architectures and attributes.

Encoder-based architecture. `MOMENT` uses an encoder and lightweight prediction heads as backbone [8]. The model tokenizes input data using fixed-length patches and employs transformers for prediction, incorporating reversible instance normalization for re-scaling and centering time-series. This approach allows `MOMENT` to be adapted to various downstream tasks. `Chronos` [1] uses `T5` architecture [15] for probabilistic forecasting. `Chronos` tokenizes time-series values using scaling and quantization and then trains existing transformer-based language model architectures on these tokenized time-series using cross-entropy loss. `SimMTM` [6] uses an encoder-based Transformer architecture with modules for masking, representation learning, similarity learning, and reconstruction. `Uni2TS` [21] utilizes encoder-only transformers for multivariate time-series, handling different patches and variates. It addresses cross-frequency learning, covariate handling, and probabilistic forecasting. `UniTime` [12] uses an encoder-based transformer, incorporating semantic instructions through a language encoder to handle domain confusion.

Decoder-only architecture. `LagLlama` [16] uses the Llama architecture [19] for multivariate time-series with a focus on probabilistic forecasting. It employs lag features, data augmentation, and the conventional Llama architecture for robust predictions. `TimesFM` [4] model employs a decoder-based transformer for multivariate time-series, processing patches through residual blocks to generate tokens for forecasting.

Others. `TimeLLM` [11] reprograms an embedding-visible language foundation model, such as Llama and GPT-2 models for univariate time-series forecasting by transforming data into a text format suitable for language models. `TimeGPT` [7] leverages a transformer-based architecture. `TimeGPT` also deals with missing data, irregular timestamps, uncertainty quantification, fine-tuning, and anomaly detection.

Table 1 summarizes the attributes of the above models. Zero-shot means whether models can make predictions without any fine-tuning or they are available without the need for training. We observe that most of these models cannot handle covariates or irregular (i.e., sampling rate varies over time) time-series data. To clarify, handling covariates leverages the covariant structure between external variables to predict a target variable. In contrast, multivariate forecasting predicts multiple variables simultaneously, which can assume independence between variables while also incorporating the covariance structure between them in some cases. Though covariates certainly impact building operations, due to limited model availability, we focus on six models with zero-shot abilities in making univariate predictions in this paper.

Table 1: Comparison of TSFM Attributes

Models	Zero Shot	Multi Resolution	Covariate Handling	Irregular Time-series	Task Agnostic	Takes Timestamps
`TimeLLM` (2024)	×	✓	✓	×	✓	×
`Uni2TS` (2024)	✓	✓	✓	×	×	✓
`SimMTM` (2023)	×	✓	×	×	×	×
`TimeGPT` (2024)	✓	✓	✓	✓	?	✓
`Chronos` (2024)	✓	✓	×	×	×	×
`MOMENT` (2024)	✓	✓	×	×	✓	×
`LagLlama` (2024)	✓	✓	×	×	×	✓
`TimesFM` (2024)	✓	✓	×	×	×	×
`UniTime` (2024)	×	✓	×	×	×	×

Table 2: Data Familiarity and Model Structures

Models	Level of familiarity with electricity	Level of familiarity with temperature	Obj.	Transformer architecture
`Uni2TS`	Modality, Dynamics	Modality	NLL	Encoder
`Chronos`	Modality, Dynamics, Dataset	Modality	CE	Encoder-decoder
`MOMENT`	Modality, Dynamics, Dataset	Modality	MSE	Encoder
`LagLlama`	Modality, Dynamics, Dataset	None	NLL	Decoder
`TimesFM`	Dataset	Modality	MSE	Decoder
`TimeGPT`	?	?	?	?

MSE: Mean Squared Error, NLL: Negative Log Likelihood, CE: Cross-Entropy

Table 2 summarizes the data familiarity and model structures for the studied TSFMs. We use three categories for data familiarity: (1) Modalities: the model is trained with data from the same modality as the test set; (2) Dynamics: the model's training corpus included time-series data generated by dynamical processes similar to that governing the test data; (3) Dataset: the UCI electricity dataset [20] is the only dataset that some of the TFSMs we evaluate were exposed to during training. We name it dataset-level familiarity if the model is trained with the UCI dataset.

3 METHODOLOGY

We begin by discussing the datasets gathered for our analysis of TSFMs. Our analysis aims to provide an initial understanding of how these models perform over longer horizons, encompassing seasonal variations and diverse household behaviors. Specifically, we focus on temperature and electricity consumption, and compare the results with Python implementations of univariate statistical models, including `AutoARIMA` [17], Seasonal ARIMA (`S-ARIMA`) [17], and `BestFit` (a 5th-degree polynomial function with learned coefficients). Each statistical model was trained using data from the context window for each prediction, allowing them to act as zero-shot predictors and ensuring a head-to-head comparison with TSFMs. We use the following versions of the TSFMs in our study: *timegpt-1-long-horizon*, *google/timesfm-1.0-200m*, *amazon/chronos-t5-large*, *moirai-1.0-R-large*, and *MOMENT-1-large*.

3.1 Large Public Datasets

3.1.1 ecobee DYD Dataset. To test the general ability of TSFMs in predicting indoor temperature, we utilized a large publicly available dataset from *ecobee* [13]. This dataset is a subset of ecobee's Donate Your Data program, containing data from 1,000 homes located in four U.S. states—California, Texas, Illinois, and New York—collected in 2017 at 5-minute intervals with a temperature resolution of 1°F. To ensure statistically significant yet computationally feasible tests, we selected eight houses with the least number of missing thermostat temperature values from each state, resulting in 32 houses. A starting point was sampled from each month, using the same starting points across models for a deterministic comparison, while resampling starting points for each house to ensure greater time diversity. This approach allows us to capture diverse house behaviors, climates, and seasonal variations. Sampling starting points was mainly necessary because the data duration changes based on varying context windows and prediction horizon values.

3.1.2 UCI Electricity Data. Similarly, to evaluate the general capability of these models in energy consumption predictions, we used the UCI Electricity Load Diagrams dataset [20]. This dataset, frequently used to evaluate forecasting algorithsm, provides an opportunity to reconsider the performance rankings of TSFMs using different metrics. The dataset records electricity consumption in Watts for 370 Portuguese clients from 2011 to 2014, sampled at 15-minute intervals. We sampled 30 houses in this dataset, and for each season, we sampled a starting point, resulting in 16 starting points for the each client. This method ensures a comprehensive evaluation across different seasonal contexts and house specifics.

3.1.3 Smart.* The dataset contains whole-house electricity consumption for 114 single-family apartments for the period of 2014-2016 in kW [2], sampled every 15 minutes. Similar to the previous approach, we sampled 30 houses in this dataset and, for each season, sampled 4 different starting points.

We analyzed the list of training datasets presented by each TSFM. None of them included indoor temperature data or used any electricity data from Amherst, MA, USA, ensuring no data leakage for *ecobee* and Smart*, while some had seen UCI datasets in their training.

3.2 Experiment Design

We define the following notations employed throughout this paper:

- H: Prediction Length (number of samples)
- C: Context Length (number of samples)
- f_s: Sampling Rate (minutes)
- D: Context Duration (hours), defined as $D = \frac{C \cdot f_s}{60}$
- P: Prediction Duration, also called *Horizon* (hours), defined as $P = \frac{H \cdot f_s}{60}$

While previous literature typically presents results in terms of the number of prediction steps/samples, we express these intervals in terms of hours since it is (a) more intuitive for electricity and temperature predictions, and (b) more comprehensible, particularly as we resample data to analyze performance across various durations.

In our analysis, we selected each context-prediction duration pair (D, P) based on two criteria: $C < 512$, and $H < 64$. The primary rationale behind this choice is that most models are optimized to make predictions within these limits [1, 4, 8]. During this selection, we considered the sampling rate and resampled to a lower temporal resolution when necessary. The context-prediction-sampling rate tuples (D(h), P(h), f_s(mins)) are as follows:

ecobee: {(24, 4, 5), (36, 4, 5), (36, 6, 10), (48, 6, 10), (48, 12, 15), (96, 12, 15), (168, 24, 30)}

Smart* and UCI: {(48, 12, 15), (72, 12, 15), (96, 12, 15), (48, 24, 15), (72, 24, 15), (96, 24, 30), (168, 24, 30)}

For the Smart* and UCI datasets, we started with a larger number of horizons due to the original sampling rate of UCI being 15 minutes. We maintained this rate for the Smart* dataset to ensure a fair comparison between the two datasets. Additionally, considering the common weekly patterns in household behavior, we aimed to test the prediction performance of these models when a week of data is provided to predict the next day. Hence, we introduced an additional tuple (i.e. (168, 24, 30)) to account for this pattern.

Table 3: RMSE values for forecasting performance. The best per row is bold, and the second best is underscored. Parameter values are in units of hours.

Data	Parameters		Models								
	D	P	AutoARIMA	S-ARIMA	BestFit	MOMENT*	Chronos*	TimeGPT	LagLlama*	TimesFM*	Uni2TS
(a) UCI (Watts)	48	12	276.6	409.4	140.0	127.6	**78.54**	192.6	279.4	105.2	204.4
	48	24	347.3	638.5	138.6	178.2	**93.83**	204.4	376.8	120.7	275.8
	72	12	241.1	375.4	153.6	143.3	**68.96**	190.4	273.3	87.88	147.8
	72	24	269.0	562.3	172.7	179.4	**80.31**	200.7	359.4	95.45	187.8
	96	12	221.8	271.7	175.4	149.8	**69.67**	193.2	220.0	89.17	123.5
	96	24	210.1	209.4	187.3	205.9	81.18	185.6	338.5	80.87	**79.07**
	168	24	151.4	148.1	190.7	218.4	**71.71**	182.4	208.4	75.42	72.24
(b) Smart* (Watts)	48	12	913	1,053	965	907	1,076	929	1,059	**897**	145,228
	48	24	936	1,097	970	932	1,113	956	1,148	**925**	1,699,758
	72	12	893	1,019	956	899	1,030	925	993	**886**	339,396
	72	24	921	1,034	972	926	1,071	951	1,092	**914**	305,820
	96	12	874	1,022	952	884	1,029	909	948	**869**	18,741
	96	24	775	842	840	787	879	765	944	**763**	734,520
	168	24	746	830	832	765	830	748	829	**735**	340,623
(c) *ecobee* (°F)	24	4	**1.047**	1.194	1.448	1.069	1.532	1.145	1.968	1.154	1.451
	36	4	**1.124**	1.129	2.635	1.252	1.679	1.217	2.017	1.213	1.289
	36	6	1.376	1.582	2.686	1.276	1.891	**1.246**	2.181	1.390	1.522
	48	6	1.240	1.465	1.973	1.193	1.709	**1.140**	2.033	1.204	1.400
	48	12	1.826	2.556	1.983	1.432	1.911	1.568	2.243	**1.319**	1.897
	96	12	1.824	2.243	2.154	1.512	1.732	1.625	2.033	**1.236**	1.657
	168	24	1.685	1.673	2.337	1.830	1.870	1.719	2.343	**1.379**	1.637

*Models that have seen the UCI dataset in their training phase

4 EXPERIMENTAL RESULTS

With the desiderata of understanding the long-term general performance of TSFMs in predicting electricity usage and indoor temperature, we conducted an analysis using three large datasets. Table 3 illustrates the performance of each model, measured by RMSE. Our results provide insights into model performance across a diverse range of seasons and household characteristics.

Our first dataset, UCI as shown in Table 3 (a), was used to train several TSFMs. Comparing model performance on familiar versus unfamiliar data reveals their generalization ability. Three models trained on this dataset (Chronos, TimesFM, MOMENT) perform best, especially for shorter durations. An interesting observation arises with the Uni2TS model, which approaches the performance of the others as the duration increases though it has not seen this data before. This phenomenon may occur because the data is resampled as the duration increases, resulting in a dataset that differs slightly from the original training data sampled every 15 minutes. Consequently, Uni2TS manages to close the performance gap. Comparing statistical models with foundation models, we observed that BestFit outperforms models that have not been trained on the UCI dataset (TimeGPT and Uni2TS) on $C = 48$ and 72. Besides, AutoARIMA and BestFit outperform LagLlama consistently despite LagLlama being trained on the dataset.

Switching to the Smart* electricity dataset (shown in Table 3 (b)) reveals a shift in model performance. The number of steps is not strictly different for varying durations, as data is resampled for longer periods, so performance discrepancies reflect behavioral characteristics. Evaluated on the Smart* dataset, Chronos loses its leading position to TimesFM. MOMENT and TimeGPT are comparable in terms of absolute performance, following TimesFM and AutoARIMA. Other FMs perform worse, and Uni2TS shows large errors due to outliers. Regarding statistical models, overall, AutoARIMA performs consistently close to the best TSFM TimesFM, BestFit outperforms three foundation models, and S-ARIMA outperforms two foundation models. Compared to the observations in the UCI dataset, the considerably larger errors of TSFMs suggest the challenging predictability of Smart* and the limited generalizability of TSFMs, narrowing the performance gap with statistical models.

The *ecobee* dataset presents a more diverse set of results (Table 3(c)), with statistical models outperforming others for shorter durations. As the duration increases, temperature variations tend to smooth out, where TSFMs seem to perform better. Nonetheless, AutoARIMA remains competitive, nearly matching the best-performing forecasting models. When comparing MOMENT, TimeGPT, and TimesFM, we find that TimesFM excels in long-duration predictions, whereas MOMENT and TimeGPT perform best during moderate durations where temperature changes are less variable. This nuanced performance indicates the importance of considering duration-specific characteristics when evaluating model efficacy.

In summary, models that have been trained using the UCI dataset, such as Chronos, TimesFM, and MOMENT, demonstrated superior forecasting abilities on the same dataset. Shifting to the unseen electricity dataset, the performance gap between statistical models and TSFMs is marginal, indicating limited generalizability across datasets. For indoor air temperature forecasting, TSFMs generally outperform statistical models across extended prediction horizons with the lowest errors ranging from 1 to 1.4°F for durations up to 24

hours. Conversely, shorter durations highlight the efficacy of statistical models, particularly `AutoARIMA`, which achieves comparative or superior accuracy (1°F).

5 DISCUSSION AND CONCLUSIONS

Our investigation stemmed from the belief that TSFMs, trained on diverse time-series data, could generalize to unseen settings. Despite training statistical models using only context window data for each prediction, we found marginal differences between TSFMs and statistical models on unseen datasets. In domains like indoor air temperature, where TSFMs lack familiar data, performance is comparable to statistical models. Beyond empirical insights, we identify areas needing further exploration and key features required for TSFMs.

Incorporating Context: Time-series data presents unique challenges that current models struggle to address. Models like `UniTime` [12] and `TimeLLM` [11] incorporate context but are not zero-shot predictors. In contrast, zero-shot TSFMs typically rely only on time-series data, lacking auxiliary context that could improve predictions. Unlike language models that leverage contextual cues such as syntax and semantics, time-series data often lacks this information, complicating pattern recognition. Additionally, the variability in time-series data, driven by factors like physical processes and sensor characteristics, adds complexity. To enhance robustness, TSFMs need to incorporate auxiliary information or undergo fine-tuning to account for this diversity.

Exploring New Attributes: Our study marks an initial exploration of TSFMs in predictive building analytics, but many attributes, such as handling covariates and irregular time-series, remain underexplored due to only a few TSFMs providing such features. While Section 2 outlines existing features in TSFMs, a more thorough investigation is required to assess their performance. Further exploration could unlock the potential of TSFMs in building analytics and other cyber-physical systems.

Based on our findings, we propose two key attributes for future TSFMs: (1) Task-agnostic models that handle a range of tasks beyond forecasting, such as classification and anomaly detection, without requiring task-specific fine-tuning. This can be achieved using pre-trained task-specific heads [8] or the design of natural language outputs. Replacing task-specific heads enables reasoning tailored to each task, while the integration of natural language will allow for more intuitive time-series analysis and user-friendly explanations. (2) The ability to integrate contextual metadata via natural language, enhancing model performance by incorporating factors like operational settings, weather, or occupancy patterns that influence building energy consumption and thermal dynamics.

In conclusion, the effectiveness of TSFMs can be preasumably further enhanced by integrating essential metadata and accounting for confounding variables, which are critical in the context of building physics. Our future research aims to improve the applicability of TSFMs in reliable building analytics by incorporating physics-based insights, robust covariate handling, and contextual metadata into these models.

ACKNOWLEDGMENTS

This research was sponsored in part by: Pennsylvania Infrastructure Technology Alliance (PITA); the Air Force Office of Scientific Research under Cooperative Agreement #FA95502210193; the DEVCOM ARL under Cooperative Agreement #W911NF-17-2-0196; and, the NIHm DOT Center under Award #1P41EB028242. Our code can be found in https://github.com/nesl/TSFM_Building.

REFERENCES

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815* (2024).

[2] Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, Jeannie Albrecht, et al. 2012. Smart*: An open data set and tools for enabling research in sustainable homes. *SustKDD, August* 111, 112 (2012), 108.

[3] Yongbao Chen, Mingyue Guo, Zhisen Chen, Zhe Chen, and Ying Ji. 2022. Physical energy and data-driven models in building energy prediction: A review. *Energy Reports* 8 (2022), 2656–2671.

[4] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. https://doi.org/10.48550/arXiv.2310.10688 arXiv:2310.10688 [cs].

[5] Longquan Diao, Yongjun Sun, Zejun Chen, and Jiayu Chen. 2017. Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. *Energy and Buildings* 147 (2017), 47–66. https://doi.org/10.1016/J.ENBUILD.2017.04.072

[6] Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. 2023. SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling. (Oct. 2023). http://arxiv.org/abs/2302.00861 arXiv:2302.00861.

[7] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. 2024. TimeGPT-1. (May 2024). http://arxiv.org/abs/2310.03589 arXiv:2310.03589 [cs, stat].

[8] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. MOMENT: A Family of Open Time-series Foundation Models. http://arxiv.org/abs/2402.03885 arXiv:2402.03885 [cs].

[9] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems* 36 (2024).

[10] Tyler Hoyt, Edward Arens, and Hui Zhang. 2015. Extending air temperature setpoints: Simulated energy savings and design considerations for new and retrofit buildings. *Building and Environment* 88 (2015), 89–96.

[11] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. http://arxiv.org/abs/2310.01728 arXiv:2310.01728 [cs].

[12] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. 2024. UniTime: A Language-Empowered Unified Model for Cross-Domain Time Series Forecasting. (Feb. 2024). https://doi.org/10.48550/arXiv.2310.09751 arXiv:2310.09751 [cs].

[13] Na Luo and Tianzhen Hong. 2022. Ecobee Donate Your Data 1,000 homes in 2017. (2022). https://doi.org/10.25584/ecobee/1854924

[14] Mike A Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. 2024. Language Models Still Struggle to Zero-shot Reason about Time Series. *arXiv preprint arXiv:2404.11757* (2024).

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[16] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. 2024. Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. http://arxiv.org/abs/2310.08278 arXiv:2310.08278 [cs].

[17] Seabold Skipper and Perktold Josef. 2010. statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference* (2010).

[18] Ying Sun, Fariborz Haghighat, and Benjamin CM Fung. 2020. A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and Buildings* 221 (2020), 110022.

[19] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[20] Artur Trindade. 2015. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C58C86.

[21] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified Training of Universal Time Series Forecasting Transformers. (May 2024). http://arxiv.org/abs/2402.02592 arXiv:2402.02592.