Towards Pretrained Models for Load Forecasting

Anonymous Author(s)

Abstract

The emergence of pre-trained models has transformed many fields, including computer vision, natural language processing, and speech recognition, by enabling robust, task-agnostic representations that generalize well across diverse downstream tasks. We present three pre-trained models for load forecasting. The proposed models include Energy-TTMs (adapted from IBM's Tiny Time Mixers), W-LSTMix, and MixForecast. These models are pre-trained on a large-scale energy metering dataset, with 1.26 billion readings, collected from 76,217 real buildings spanning multiple regions, building types, and usage scenarios. This extensive training enables the models to capture complex temporal patterns across diverse building types and operational scenarios. We benchmark the performance of our pre-trained models against six recent Time Series Foundation Models (TSFMs), such as Chronos, Lag-Llama, Moirai, TimesFM, TTMs, and MOMENT, as well as multiple traditional and machine learning-based forecasting models, under zero-shot and transfer learning settings, on a large-scale real-world dataset of over 1,767 residential and commercial buildings for the task of short-term load forecasting (STLF). Our results show our pre-trained models can outperform task-specific models in zero-shot settings, highlighting their generalizability and versatility. Finally, we share insights to guide future development of pre-training models for energy data analytics.

Keywords

Short-term Load Forecasting (STLF), Time Series Foundation Models (TSFM), Zero-shot Forecasting, Demand-side Load Management.

1 Introduction

Accurate load forecasting is crucial for decarbonizing the building sector, which contributes to approximately one-third of global energy consumption and greenhouse gas (GHG) emissions. Accurate forecasts enable efficient energy management by predicting future energy needs, aligning supply with demand, reducing reliance on fossil fuels, integrating renewable energy sources, and optimizing energy use within buildings.

Various techniques have been proposed for short-term load forecasting (STLF), ranging from classical statistical models to modern machine learning approaches. Traditional methods such as ARIMA are simple and easy to implement, but they often fall short in capturing the non-linear dependencies and complex temporal interactions inherent in energy consumption data, which are influenced by dynamic weather conditions and human behavior. Additionally, these models face scalability challenges, as they require separate implementations for each building, making model management increasingly impractical at scale.

In contrast, machine learning methods, particularly neural networks, are better suited to handle non-linearities and temporal dynamics, offering improved forecasting accuracy. However, these models also demand substantial computational resources and large, diverse datasets from multiple buildings to generalize effectively. To overcome such limitations, transfer learning has emerged as a promising solution by enabling the use of knowledge from data-rich source buildings to improve forecasting accuracy in target buildings with limited data, thus enhancing scalability while reducing data requirements [53].

In recent years, Time Series Foundation Models (TSFMs) have emerged as a powerful paradigm for universal time series analysis [29]. These models are pre-trained on large corpora of time series data across various domains—including finance, healthcare, and energy—and across different time granularities, enabling them to learn task-agnostic representations that generalize well across a range of downstream tasks. Unlike traditional transfer learning, TSFMs support zero-shot or few-shot inference on previously unseen series without requiring fine-tuning, making them highly attractive for large-scale, heterogeneous forecasting problems. Despite their growing success in other domains, the application of TSFMs to energy forecasting remains underexplored. This is a critical gap given the urgent need for generalizable and scalable forecasting solutions that can operate effectively across diverse building types, geographies, and usage contexts. Pre-trained models tailored to energy time series can significantly reduce the need for custom model development per building, enabling more sustainable and data-efficient analytics pipelines.

In this work, we attempt to address this gap by pre-training existing models for load forecasting task. We introduce three pre-trained models, based on IBM's Tiny Time Mixers (TTMs), W-LSTMix [13], and MixForecast [27], trained on a large-scale dataset of hourly electricity consumption from 76,217 buildings across multiple regions, seasons, and usage scenarios. We further evaluate four recent open-source TSFMs, Chronos [6], Lag-Llama [50], Moirai [61], and TimesFM [12], on a new large-scale STLF benchmark, comparing them to traditional and neural forecasting baselines under zero-shot and transfer learning.

The key contributions of this paper are:

- We present a large-scale dataset comprising hourly electricity consumption data from over 76,217 buildings across the world.
- We present three pre-trained models, IBM's Tiny Time Mixers (TTMs) [15] adopted to as Energy-TTMs, W-LSTMix [13] and MixForecast [27] architectures. All models are pre-trained on our dataset containing 1.26 billion meter readings.
- We present a comprehensive zero-shot and transfer learning evaluation of the above models along with six recent TSFMs, Chronos [6], Lag-Llama [50], Moirai [61], MOMENT, and TimesFM [12] on a large-scale real-world dataset covering over 1,767 residential and commercial buildings.
- We also compare the zero-shot performance of TSFMs against traditional machine learning-based forecasting models, including LightGBM, Linear Regression, Auto-ARIMA, and transformers trained from scratch (TFS).

Table 1: Comparison of pre-trained models and their key attributes. (MCL-Maximum Context Length)

Model	Architecture	Parameters	MCL	Downstream Tasks	Pre-training Data	Energy Data Used
Chronos [6]	Encoder-Decoder	46M	512	Forecasting	893K Series	UCIE (370)
Moirai [61]	Encoder-only	14M	5000	Forecasting	27B Obs	Buildings-900K
TimesFM [12]	Decoder-only	200M	512	Forecasting	100B Obs	UCIE (370)
Lag-Llama [50]	Decoder-only	2.45M	1024	Forecasting	1B Obs	LCL, UCIE (370)
TTMs [14]	MLP	1M	1536	Forecasting	282K Series	LCL, AUE
TSPulse [16]	MLP	1M	512	Imputation, Anomaly Detection, Classification, Similarity Search	1B Obs	-
MOMENT [20]	Encoder-only	125M	512	Forecasting, Imputation, Anomaly Detection, Classification	1.13B Obs	UCIE (370)
Energy domain						
Transformer-L (G) [17]	Encoder-Decoder	160M	168	Forecasting	7.9B points	Buildings-900K
MixForecast [27]	**MLP**	**0.69M**	**168**	**Load forecasting**	**1.26B points**	**Our dataset**
W-LSTMix [13]	**MLP, LSTM**	**0.13M**	**168**	**Load forecasting**	**1.26B points**	**Our dataset**
Energy-TTMs	**MLP**	**1M**	**168**	**Load forecasting**	**1.26B points**	**Our dataset**

2 Methodology

2.1 Time Series Foundation Models (TSFM)

In recent years, there has been growing interest within the AI research community in building foundation models for universal time series analysis. These TSFMs differ significantly in size, architectural complexity, pretraining strategies, underlying datasets, and their ability to support multiple downstream tasks. Table 1 summarizes the key attributes of several widely used TSFMs, including their base architecture, number of parameters, maximum supported context length, supported downstream tasks, and the datasets used for pretraining. Nevertheless, developing such models remains highly challenging due to the complexity of modeling time series patterns from heterogeneous domains, capturing long-term temporal dependencies, and the substantial resource requirements, particularly the need for large volume of quality datasets and large-scale compute infrastructure (e.g., GPU cluster). As shown in Table 1, most existing TSFMs are primarily designed for the forecasting task and support a range of context lengths, between 500 and 5000 time steps, except MOMENT and TSPulse which can support time series data imputation, anomaly detection, and classification. Notably, as seen from Table 1, some of these models have already been pretrained on time series data from various domains, including energy, and have demonstrated strong generalization across diverse forecasting scenarios.

A few recent studies [43, 55] have examined the performance of TSFMs under zero-shot settings; however, their evaluations have been limited to a small number of buildings. In this work, we evaluate five open-source TSFMs for the load forecasting task under both zero-shot and transfer learning settings, using a significantly larger and more diverse set of buildings. These models were selected based on the availability of publicly accessible code and documentation.

Lag-Llama [50] is based on the decoder-only transformer-based architecture of LLaMA. It generates lagged features from time series using specified indices and includes date-time covariates for enhanced context. It predicts future values through a distribution head that outputs parameters for a chosen probability distribution, such as the Student's t-distribution. The model employs robust standardization to ensure resilience to outliers across diverse datasets.

TimesFM [12] is a decoder-only foundation model for time-series forecasting, pre-trained on a large volume of temporal data that displays good zero-shot performance on a variety of public benchmarks from different domains and granularities. The architecture is based on three key principles: 1) breaking time-series data into patches for efficient training, 2) using a decoder-only approach to predict future patches from past ones, and 3) supporting longer output patches to improve forecasting.

Chronos [6] is a foundation model for probabilistic time series forecasting, based on the encoder-decoder T5 (Text-to-Text Transfer Transformer) family. It follows a minimalist language model approach by first tokenizing time series values through scaling and quantization, mapping the continuous values to a fixed vocabulary. The model then trains existing T5 architectures on these sequences of tokens without incorporating any time-series-specific design modifications.

Moirai [61] is a model built on a masked encoder-only Transformer architecture. It employs a patch-based approach, projecting time series patches into vector representations using a multi-patch size input projection layer. This design, which learns distinct input and output embeddings for various patch sizes, allows the model to effectively handle time series with diverse frequencies. While powerful in zero-shot scenarios, its performance can be significantly enhanced through fine-tuning, where it has demonstrated state-of-the-art results in tasks like residential load forecasting.

MOMENT [20] is a family of encoder-only Transformer-based foundation models designed for general-purpose time-series analysis. MOMENT employs a simple "mask and reconstruct" pretraining objective inspired by BERT in natural language processing. This architecture tokenizes time series into non-overlapping patches, which are then linearly projected into embeddings. Furthermore, this approach allows MOMENT to deliver strong zero-shot and few-shot performance across a wide range of downstream tasks, establishing it as a robust general-purpose tool for time-series analysis.

Tiny Time Mixers (TTMs) [15] is an MLP-based pre-trained model for multivariate time-series forecasting. It builds on the lightweight TSMixer backbone to jointly mix information along temporal and channel dimensions, replacing costly self-attention blocks with efficient token-wise mixers. TTMs introduces adaptive patching—variable-length tokenization of time windows—and a resolution prefix conditioning mechanism that lets a single model generalize across diverse sampling rates. It also uses a shared trunk to learn global temporal embeddings and multiple lightweight heads for decoder-channel mixing and for incorporating exogenous or categorical covariates during fine-tuning, enhancing robustness to outliers and dataset shifts. For the experiments, TTMs used a context length of 512 hours with a forecasting length of 96 hours.

2.2 Energy Pre-trained Models

Recently, there has been a growing interest in developing pre-trained models tailored for energy data analysis tasks. These models are typically built by adapting existing Transformer-based or lightweight MLP-based architectures and are pre-trained on large-scale energy datasets mainly for load forecasting. By leveraging domain-specific temporal patterns, such as daily load cycles and seasonal consumption trends, these models offer strong initialization for downstream tasks. Their pre-training enables effective performance in both zero-shot and fine-tuning settings, making them particularly well-suited for short-term load forecasting at the building level.

BuildingsBench's Transformer-L (Gaussian)[17] is a large, pre-trained model that is part of the BuildingsBench evaluation platform [17] Its backbone is based on the Transformer architecture presented in [62]. The model is pre-trained on hourly electricity time series data generated from EnergyPlus simulations, covering over 900,000 commercial and residential buildings in the USA, using a fixed context length (168) tailored for short-term load forecasting. BuildingsBench has two variants of pre-trained models: Transformer-L (Gaussian) and Transformer-L (Tokens). In our study, we selected Transformer-L (Gaussian), as it has demonstrated best performance in short-term load forecasting tasks on various real-world building datasets. The "Gaussian" designation indicates that the model is designed for probabilistic forecasting, where its prediction head outputs the parameters of a Gaussian distribution instead of a single point estimate.

MixForecast [27] is a Mixer-enhanced model designed for efficient short-term load forecasting. It employs a hybrid architecture that enhances the NBEATS framework by replacing its fully connected layers with TSMixer blocks. This structural modification improves the model's ability to capture complex temporal dependencies and multivariate feature interactions. The resulting lightweight design has demonstrated superior accuracy and adaptability compared to both foundation and traditional models in zero-shot and fine-tuned settings. In this work, we pre-train MixForecast on 1.26 billion energy meter readings from 76,217 real buildings and compare its performance with other baseline models.

W-LSTMix [13] is a modular, lightweight hybrid forecasting architecture for building-level short-term load forecasting. It begins with a decomposition step that separates the time series into trend and seasonal-residual components, which are then processed individually using a specialized dual-stack architecture built on the N-BEATS framework. A stack of Long Short-Term Memory (LSTM)-enhanced blocks models the long-term trend, while a second stack augmented with an MLP-Mixer, forecasts the seasonal and residual patterns through efficient patch-wise temporal mixing. This hybrid decompositional design enables robust generalization, allowing the model to consistently outperform other prominent foundation models in both zero-shot and fine-tuned forecasting scenarios. Similar to MixForecast, we pre-train W-LSTMix architecture on 1.26 billion energy meter readings from 76,217 real buildings and compare its performance with other baseline models.

Energy-TTMs is our adaptation of the Tiny Time Mixers (TTMs) architecture [14], specialized for univariate time-series forecasting. Unlike many recent TSFMs that rely on transformer-based architectures, TTMs adopt a non-transformer design, using a lightweight, MLP-based architecture, whcih significantly reducing computational overhead while maintaining strong forecasting performance. TTMs excel in both zero-shot and few-shot scenarios, achieving accurate predictions even with limited or no task-specific training data. TTMs also support multivariate time series forecasting and can effectively incorporate exogenous variables, such as weather data, through multi-level modeling, enabling them to capture complex temporal and contextual dependencies. Additional features such as adaptive patching allow TTMs to handle sequences of varying resolutions, while resolution prefix tuning enables them to generalize across different time series frequencies and forecast horizons. Due to their MLP-based architecture, TTMs are relatively easy to train with modest computational resources. For example, the original TTM model was pre-trained on 282,000 time series using 8 A100 GPUs over the course of two days. Due to its efficient architecture and the availability of pre-training code, we selected TTMs for adaptation and pre-training on energy domain data, referring to this version as **Energy-TTMs**.

It is important to note that, despite being specialized for the energy domain, all these models share a common limitation with most generic TSFMs − they are primarily designed to handle only the forecasting task with a context length of up to 168. (See Table 1).

2.3 Trained from Scratch Transformer (TFS)

Transformers [60] are a type of deep learning model that utilizes self-attention mechanisms to efficiently process sequential data, enabling state-of-the-art performance in tasks such as time series forecasting. These models are trained from scratch for the forecasting task by learning directly from context windows, e.g., 1 week, derived from real building data. Once trained, these models generate day-ahead forecasts for the test windows, making predictions based on the patterns learned during training. In this, we have included two models, PatchTST [44] and Temporal Fusion Transformer (TFT) [31].

2.4 Traditional Machine Learning (ML) Models

Several traditional ML models have been widely used for load forecasting tasks in the literature. In this study, we have selected three baseline models: Auto-ARIMA [39], DLinear [64], and LightGBM [46].

(1) **Naive model** The Naive forecasting method assumes that the most recent observed value will persist into the future. For day-ahead STLF, this typically involves repeating the load values from the same hour on the previous day, i.e., $\hat{y}_{t+h} = x_{t+h-24}$ for $h = 1, 2, \ldots, 24$. Despite its simplicity, the naive model often performs surprisingly well in stable load environments with strong daily or weekly seasonal patterns, making it a strong baseline for comparison [59].

Table 2: Dataset Summary

Type	# Buildings	# data points
Pre-training		
Commercial Buildings	2,792	59M
Residential Buildings	73,425	1.2B
Total	76,217	1.26B
Evaluation - Out-of-Distribution (OOD)		
Commercial Buildings	253	2.32M
Residential Buildings	245	2.37M
Total	498	4.69M
Evaluation - In-Distribution (ID)		
Commercial Buildings	98	1.56M
Residential Buildings	1,171	16M
Total	1,269	17.56M

(2) **Auto-ARIMA** is an automatically tuned ARIMA model, a popular time series forecasting method based on the autoregressive integrated moving average model.
(3) **Linear Regression** is a simple yet effective technique for predicting continuous outcomes.
(4) **LightGBM** is a gradient boosting framework known for its efficiency and scalability, especially in handling large datasets. In this study, we use the multi-step forecasting implementation from skforecast[1] with 100 estimators and no max depth.

2.5 Datasets

Pre-training Data: For the pre-training stage, we collected a large-scale energy consumption dataset consisting of 1.26 billion hourly observations collected from 76,217 real-world buildings, encompassing both commercial and residential types across diverse countries and temporal spans. Table 8 and Table 9 present the details of commercial and residential datasets used for pre-training. Table 2 provides an overview of the complete dataset composition used for pre-training and evaluation across commercial and residential buildings.

Out-of-Distribution (OOD): This evaluation set comprises datasets entirely excluded from the pre-training corpus, enabling assessment of model generalization to unseen buildings across distinct regions. For this purpose, we randomly selected 498 buildings from commercial and residential categories.

In-Distribution (ID): This evaluation set is derived from datasets where a subset of buildings is used during pre-training, while the remaining buildings are held out for evaluation. This split encompasses 1,269 buildings, covering both commercial and residential types.

Sliding Window Extraction: We used sliding windows for each building. Specifically, we employed an 8-day sliding window comprising a 192-hour load sub-sequence, with a stride of one day. The initial 7 days (168 hourly energy meter readings) served as context to forecast the subsequent 24-hour readings of the 8$^{\text{th}}$ day.

3 Experiments and Results

3.1 Experimental Setup

All experiments were carried out using `JupyterLab` and `VS Code`, leveraging a suite of open-source libraries for time series forecasting. Specifically, we utilized `GluonTS`[2] and `AutoGluon`[3] for implementing TSFM, Naive, and Auto-ARIMA baselines; `skforecast` for LightGBM and Linear Regression models; and `neuralforecast`[4] for trained from scratch transformers.

Energy-TTMs, Mix-Forecast, and W-LSTMix are pre-trained on energy consumption data from 76,217 buildings using an 8-day (192-hour) sliding window with a 1-day stride. Each sample contains a 168-hour context window used to forecast the next 24 hours.

- The Energy-TTMs model's encoder is a stack of `num_layers=3` TSMixer blocks, each with hidden dimension `d_model = 16` and expansion factor 3, and employs GELU activations, a channel-mixing mode that alternates between "common_channel" and "mix_channel," gated attention (no self-attention), with dropout = 0.30 (and head_dropout = 0.20). We disable positional encoding and use LayerNorm in the MLPs.
  On the decoder side, we attach `decoder_num_layers = 8` TSMixer blocks of dimension `decoder_d_model = 8`, without adaptive patching or raw residuals, operating in `common_channel` mode. We train for up to 100 epochs with early stopping (patience = 10, threshold = 1e-4). This experiment was conducted on 4 × H100 GPUs and completed in approximately 40 hours.
- The MixForecast model features a hierarchical architecture composed of 6 stacks, where each stack contains 3 TSMixer-enhanced blocks. These blocks operate with a hidden dimension of `512`, utilize internal expansion factors of `[4,2,1]` for different size patch lengths, and employ `GELU` activations.The model was optimized using the `Huber` loss function and trained for up to `35` epochs on a server equipped with 4 × V100 GPUs and completed in 4 days.
- The W-LSTMix model first processes the input time series using a `seasonal_decompose` method into trend and seasonal components. The architecture is constructed from trend and seasonal stacks, each containing `3` hybrid blocks. These blocks operate with a hidden dimension of `256` and a patch size of `8`. Key internal components include an embedding dimension of `8` and a basis expansion dimension (`thetas_dim`) of `8`. The model is optimized using a `Huber` loss function and trained for up to `35` epochs on a server equipped with `4` x AMD Instinct MI300X GPUs, also completed in 4 days.

It is important to note that all our pre-trained models do not require GPUs for inference and fine-tuning, as they are based on non-transformer architectures. In contrast, most existing general-purpose TSFMs require GPUs, at least for fine-tuning, due to their large model size and training complexity.

3.2 Evaluation Metrics

We evaluate the performance of the forecasting models using the normalized root mean square error (NRMSE) metric, similar to the BuildingsBench platform [17]. The NRMSE (also known as the

[1] https://skforecast.org/

[2] https://ts.gluon.ai/stable/index.html

[3] https://auto.gluon.ai/

[4] https://nixtlaverse.nixtla.io/neuralforecast/docs/getting-started/introduction.html

Table 3: Comparison of model performance using mean NRMSE of Out-of-distribution datasets.

Models	Commercial	Residential
Not Pre-trained + Not Fine-tuned		
Naive (Mean)	47.72	96.86
Not Pre-trained + Fine-tuned		
Auto-ARIMA	48.61	85.00
LR	37.51	103.47
LightGBM	**35.26**	110.37
PatchTST	36.54	**71.73**
TFT	39.81	76.76
Pre-trained + Not Fine-tuned		
Moirai	53.10	76.14
Lag-Llama	57.45	82.49
Chronos	33.87	73.19
TimesFM	48.67	**66.98**
MOMENT	44.91	76.37
TTMs	**30.90**	73.16
Pre-trained (Energy) + Not Fine-tuned		
Transformer-L (G)	-	85.67
Energy-TTMs (Ours)	17.41	76.75
Mix-Forecast (Ours)	20.57	83.64
W-LSTMix (Ours)	**12.74**	**49.94**
Pre-trained + Fine-tuned		
Moirai	**27.24**	**57.18**
Lag-Llama	30.92	66.63
MOMENT	42.95	73.20

coefficient of variation of the RMSE) is widely used, as it captures the ability to predict the correct load shape. For a target building with M days of load time series:

$$NRMSE := 100 \times \frac{1}{\bar{y}} \sqrt{\frac{1}{24M} \sum_{j=1,i=1}^{M,24} (y_{i,j} - \hat{y}_{i,j})^2}, \tag{1}$$

where $\hat{y}$ is the predicted load, y is the actual load, and $\bar{y}$ is the average actual load over all M days. Additional details are made available in the reproducible code repository[5].

3.3 Results and Analysis

We follow the experimental protocol proposed in the BuildingsBench platform [17] and consider two settings for evaluating the models: *Zero-shot* and *Transfer Learning*.

Zero-shot: In the zero-shot STLF setting, models are tasked with generating day-ahead hourly forecasts for 1,767 unseen commercial and residential buildings, given hourly data (168 hours of context and 24 hours of forecast). This task evaluates each model's ability to forecast energy consumption for new buildings with minimal historical data.

Transfer Learning: In this setup, a pre-trained model is fine-tuned on each target building. To evaluate this process, each building's historical data is chronologically split into training (first 40%), validation (next 10%), and test (final 50%) sets. The general-purpose TSFMs (Moirai, Lag-Llama, and MOMENT), as well as our pre-trained models (Energy-TTMs, MixForecast, and W-LSTMix), are fine-tuned on the training set, with early stopping based on the validation set to prevent overfitting. Final performance is reported on the held-out test set.

Table 3 compares the average NRMSE for all evaluated models across commercial and residential out-of-distribution (OOD) datasets. We adopt the terms *pre-trained* and *fine-tuned* from the BuildingsBench benchmark, as we follow their experimental protocol.

In the **Not Pre-trained + Fine-tuned** category, *LightGBM* achieves the best performance on commercial buildings with an NRMSE of 35.26, followed by *LR* at 37.51. The worst performer is *Auto-ARIMA* at 48.61. On residential buildings, *PatchTST* performs best with 71.73, followed by *TFT* at 76.76, while *LightGBM* performs the worst at 110.37.

In the **Pre-trained + Not Fine-tuned** category, the best-performing model for commercial buildings is *TTMs* with an NRMSE of 30.90, followed by *Chronos* at 33.87. The worst is *Lag-Llama* at 57.45. For residential buildings, *TimesFM* performs best with 66.98, followed by *TTMs* at 73.16, while *Lag-Llama* again performs worst with 82.49. These models outperform traditional machine learning-based forecasting methods, owing to their strong generalization capabilities, enabled by pre-training on large-scale time series datasets from diverse domains.

In the **Pre-trained + Fine-tuned** category, the best model for commercial buildings is *Moirai* with an NRMSE of 27.24, followed by *Lag-Llama* at 30.92, while *MOMENT* is worst with 42.95. For residential buildings, *Moirai* again performs best with 57.18, followed by *Lag-Llama* at 66.63, and *MOMENT* again performs worst with 73.20. Both *Moirai* and *Lag-Llama* outperform traditional machine learning models and show significant improvements after fine-tuning compared to their zero-shot counterparts. Fine-tuning results for other foundation models are not reported due to the lack of publicly available code and documentation necessary for fine-tuning them.

In the **Pre-trained + Not Fine-tuned** category for models trained exclusively on energy meter data, *W-LSTMix* achieves the best performance on both commercial (12.74) and residential (49.94) buildings. The second-best is *Energy-TTMs*, with NRMSEs of 17.41 for commercial and 76.75 for residential. The worst performer is *Transformer-L (G)* with an NRMSE of 85.67 on residential buildings. This high error may be due to the model being pre-trained only on simulated energy time series, which limits its ability to generalize to real-world data. Results for commercial buildings are not reported, as the model produced unusually high errors that are still under investigation. Overall, *W-LSTMix* demonstrates the best performance across both building types, outperforming baseline methods and existing foundation models.

Note: Existing TSFMs have been pre-trained on various datasets, including some containing energy domain data (see Table 1). Retraining these models—either exclusively on energy datasets or excluding energy-related data—is extremely challenging due to the

[5] https://drive.google.com/drive/folders/1JZGXBx2XrDXVJidS0TOUbVW5kyKKLvKe?usp=sharing

Table 4: Comparison of models using median NRMSE for in-distribution datasets (TFT: Temporal Fusion Transformer)

Dataset	Traditional ML Models				TFS		Pre-trained		
	Naive	Auto-ARIMA	LR	LightGBM	PatchTST	TFT	Energy-TTMs	MixForecast	W-LSTMix
Commercial Buildings									
BDG-2	28.95	29.05	21.65	21.71	21.24	26.06	14.08	14.49	10.16
Enernoc	43.56	40.80	26.29	26.05	28.41	33.60	24.35	19.15	18.98
IBlend	36.65	37.85	21.44	22.15	22.06	21.05	32.45	33.49	21.57
PSS	107.90	113.82	76.13	72.34	79.73	94.24	69.34	65.77	60.44
SKC	41.15	40.47	45.68	39.37	48.49	43.76	35.12	33.46	24.50
UNICON	28.10	29.65	33.84	29.95	19.31	20.15	17.24	21.46	14.11
Residential Buildings									
DESM	82.33	84.48	78.68	70.15	58.39	64.52	65.37	65.90	59.64
DTH	104.20	95.11	34.44	34.05	30.46	81.04	29.32	27.11	22.59
ECCC	104.10	86.67	87.07	83.76	72.46	83.27	62.37	60.43	59.25
GoiEner	143.39	126.59	131.40	129.35	140.30	119.48	84.81	169.08	50.92
HES	105.09	91.63	111.20	128.14	90.68	96.35	57.10	222.48	42.65
HSG	0.94	0.94	32.44	54.59	4.38	1.47	6.28	6.54	2.01
HUE	91.28	90.77	249.94	289.03	75.46	75.80	45.91	119.15	34.54
IRH	100.49	98.17	95.09	92.15	81.52	87.62	80.98	80.78	74.39
NEEA	110.49	87.80	93.59	91.42	76.12	80.39	93.27	97.05	60.70
NESEMP	158.39	132.68	123.04	123.56	76.23	89.24	74.78	121.76	54.37
Norwegian	63.78	59.06	62.28	61.15	54.36	55.40	45.44	45.41	43.01
PES	109.11	87.70	196.68	204.44	62.55	87.45	38.75	90.51	39.75
RSL	99.11	84.60	111.19	138.36	79.20	83.20	46.85	113.03	32.24
SAVE	1.01	1.06	8.13	40.67	4.13	0.93	1.53	6.35	1.95
SGSC	129.07	112.43	115.05	112.57	97.19	104.68	86.03	113.02	50.80
UKST	110.87	98.13	99.91	102.13	89.58	94.45	64.40	93.70	51.15
iFlex	41.92	39.61	41.91	40.32	34.38	34.78	16.87	58.95	14.69

Table 5: Comparison of models using median NRMSE for in-distribution datasets – TSFM and Pre-trained Models.

Dataset	TSFM-Zeroshot							TSFM-Fine-tuned			**Pre-trained**
	Moirai	Lag-Llama	Chronos	TimesFM	Moment	TTMs	TF-L (G)	Moirai	Lag-Llama	Moment	W-LSTMix
Commercial Buildings											
BDG-2	29.49	46.37	14.56	29.05	28.31	15.45	103.55	15.18	18.57	29.08	10.16
Enernoc	28.97	43.61	21.92	40.80	41.01	21.17	109.56	22.87	30.36	35.95	18.98
IBlend	75.10	55.13	30.10	37.85	35.85	30.31	111.03	24.90	21.44	22.44	21.57
PSS	83.09	95.98	84.30	113.82	96.44	70.07	133.89	47.59	63.52	97.12	60.44
SKC	37.65	46.65	36.17	40.47	42.97	31.12	110.95	32.63	31.39	49.25	24.50
UNICON	64.32	56.98	16.17	29.65	24.86	17.25	101.47	20.25	20.26	23.85	14.11
Residential Buildings											
DESM	80.75	80.75	71.60	66.98	74.33	63.70	130.33	50.39	62.91	63.56	59.64
DTH	34.93	53.28	28.66	27.68	82.49	30.17	108.25	33.22	29.37	61.16	22.59
ECCC	69.05	78.11	76.68	64.02	79.78	61.57	98.05	64.82	72.45	83.75	59.25
GoiEner	121.39	131.44	121.14	111.12	119.02	117.69	15.65	101.45	110.13	117.85	50.92
HES	85.64	105.07	92.00	89.23	90.99	74.88	98.09	79.66	91.25	99.47	42.65
HSG	1.04	2.60	1.78	2.75	2.89	-	152.99	2.38	3.27	0.87	2.01
HUE	87.45	92.74	77.05	71.76	78.24	-	17.64	57.78	62.26	82.05	34.54
IRH	85.49	100.60	92.36	80.68	94.47	76.66	137.60	67.93	85.37	85.01	74.39
NEEA	96.10	90.14	87.56	75.49	83.80	71.73	155.39	68.33	87.45	79.48	60.70
NESEMP	123.64	111.69	95.87	79.26	95.04	-	135.66	66.13	74.74	82.41	54.37
Norwegian	49.36	62.24	52.44	52.68	51.84	46.83	28.72	44.13	53.94	54.85	43.01
PES	68.21	81.60	75.70	62.66	79.20	-	22.87	47.04	56.64	67.61	39.75
Plegma	139.77	142.43	141.66	140.31	138.57	186.82	224.11	110.99	133.61	137.34	87.23
RSL	99.39	87.34	84.50	62.76	75.48	94.76	13.19	61.62	71.33	74.10	32.24
SAVE	1.12	2.91	1.81	2.75	3.37	10.86	118.38	1.03	2.36	1.73	1.95
SGSC	97.46	109.90	101.12	91.06	99.85	78.18	33.97	78.46	90.52	100.13	50.80
UKST	96.69	98.72	82.79	83.05	87.24	70.03	20.86	69.61	76.42	88.11	51.15
iFlex	32.96	53.19	32.76	41.38	38.11	40.31	29.45	24.24	35.24	38.10	14.69

Table 6: Comparison of median NRMSE for tradition ML models and our pre-trained models on out-of-distribution datasets.

Dataset	ML Models - Not Pre-trained + Fine-tuned						Pre-trained + Not fine-tuned		
	Naive	Auto-ARIMA	LR	LightGBM	PatchTST	TFT	Energy-TTMs	MixForecast	W-LSTMix
Commercial Buildings									
IPC-Commercial	84.53	59.13	44.2	44.83	55.61	54.68	29.49	28.94	13.41
NREL	39.17	43.10	12.05	12.50	21.01	30.10	16.31	12.19	12.06
Residential Buildings									
CEEW	100.74	84.04	129.24	132.07	89.82	90.83	62.12	89.06	49.81
ECWM	48.69	52.03	53.52	50.39	38.00	38.34	37.50	38.30	32.42
HONDA-SH	33.01	37.28	15.33	14.96	26.22	25.79	12.64	12.83	11.30
RHC	76.08	71.02	66.23	64.33	56.94	57.79	51.54	51.13	50.21
NREL	84.68	78.65	49.83	49.62	52.92	76.32	42.15	40.53	38.97
fIEECe	85.84	63.18	67.37	64.13	53.92	67.26	41.14	49.73	44.47

Table 7: Comparison of median NRMSE for TSFM models (zero-shot and fine-tuned) on out-of-distribution datasets.

Dataset	**TSFM-Zeroshot**							**TSFM-Fine-tuned**			**Pre-Trained**
	Moirai	Lag-Llama	Chronos	TimesFM	Moment	TTMs	TF-L (G)	Moirai	Lag-Llama	Moment	W-LSTMix
Commercial Buildings											
IPC-Commercial	102.07	53.94	24.69	27.48	68.48	32.38	122.50	42.05	39.46	41.48	13.41
NREL	21.25	42.57	16.90	20.31	35.61	12.30	101.75	13.62	19.62	35.10	12.06
Residential Buildings											
CEEW	266.44	87.57	95.87	60.06	76.67	88.73	16.12	96.20	75.39	81.44	49.81
ECWM	38.85	64.63	39.48	37.70	43.08	37.41	105.81	30.80	39.30	43.06	32.42
HONDA-SH	14.00	39.28	12.80	13.66	26.30	13.02	102.90	11.29	16.97	24.50	11.30
RHC	56.52	64.55	59.83	82.84	60.91	52.05	116.68	52.14	54.92	55.10	50.21
NREL	44.57	65.15	43.15	54.01	65.83	40.70	19.54	33.80	44.34	62.08	38.97
fIEECe	53.16	64.09	55.91	63.58	57.02	49.61	13.20	40.73	54.44	54.81	44.47

lack of publicly available code and documentation, as well as the high computational costs involved. As a result, our comparison with these pre-trained models is not entirely bias-free. Since TSFMs are pre-trained independently under different settings, establishing a common, unbiased benchmark for energy analytics remains an open challenge. Nevertheless, we include them in our evaluation to provide initial baselines and assess their relative performance.

In addition, Tables 6 and 7 present the median NRMSE for each model across individual out-of-distribution (OOD) datasets, while Tables 4 and 5 report performance on in-distribution datasets. In the following section, we summarize our key observations.

3.3.1 In-Distribution (ID) Evaluation: The in-distribution evaluation setup comprises datasets where a small subset of buildings (approximately 2%) from each dataset, covering both commercial and residential categories, were held out from our pre-training was reserved exclusively for evaluation. This controlled split enables a realistic assessment of model performance when exposed to partially known geographical and domain settings. The ID evaluation dataset contains 1,269 buildings distributed across a wide range of countries, including the USA, UK, India, France, Canada, Norway, Ireland, Germany, and others, as detailed in Table 11. This large and geographically diverse dataset allows for robust benchmarking to assess a model's ability to generalize to previously seen regions and building types, while still evaluating its robustness to unseen buildings within those regions.

Key Observations: Across commercial building datasets, traditional ML models such as *LightGBM* and *LR* perform reasonably well on datasets like *BDG-2* and *UNICON*, but struggle in more complex scenarios such as *PSS*, where NRMSE exceeds 70. Transformer-based supervised models such as *PatchTST* and *TFT* show marginal improvements on a few datasets but remain inconsistent overall. In contrast, *W-LSTMix* and *Energy-TTMs* consistently outperform other models, with *W-LSTMix* achieving the lowest NRMSE across all commercial datasets (e.g., 10.16 on *BDG-2*, 14.11 on *UNICON*). Even when compared against fine-tuned TSFMs like *Moirai*, *Lag-Llama*, and *Chronos*, *W-LSTMix* remains the top performer in most commercial test cases, underscoring the benefits of energy-specific pre-training.

In residential datasets, performance varies significantly due to greater consumption heterogeneity and dataset-specific complexities. Traditional models often perform poorly, recording extremely high NRMSEs on several datasets (e.g., 289 on *HUE*, 204 on *PES*). While TSFMs in zero-shot and fine-tuned settings show some improvements, their results remain variable. *W-LSTMix* demonstrates strong generalization and robustness, achieving the best performance on 13 out of 17 residential datasets, including highly diverse ones such as *RSL* (32.24), *GoiEner* (50.92), and *UKST* (51.15). Although some datasets (e.g., *HSG*, *SAVE*) are relatively easy for most models, *W-LSTMix* still delivers consistently competitive or superior results. These findings reinforce its adaptability across different regions, building types, and consumption patterns when evaluated in semi-familiar (ID) settings.

3.3.2 Out-of-Distribution (OOD) Evaluation: This evaluation setting assesses the true generalization ability of models to completely unseen datasets, i.e., buildings and regions that were entirely excluded during pre-training. The OOD test set includes 498 buildings drawn from six geographically and temporally diverse datasets spanning countries such as the USA, China, India, Mexico, Canada, and South Africa. Unlike in-distribution evaluation, OOD performance reflects a model's robustness to regional and consumer behavior heterogeneity, and data distribution shifts. As such, it serves as a rigorous set for evaluating the transferability of learned representations.

For commercial buildings, traditional ML models (e.g., *LR*, *LightGBM*) perform moderately well on the *NREL* dataset (NRMSE ~12), but their performance deteriorates significantly on more complex data such as *IPC-Commercial* (NRMSE > 44). Transformer-based models like *PatchTST* and *TFT* also exhibit inconsistent results across these datasets. In contrast, pre-trained models—especially those adapted for energy data—offer significant performance advantages. Notably, *W-LSTMix* achieves the best results on both commercial datasets (NRMSE: 13.41 on *IPC-Commercial*, 12.06 on *NREL*), outperforming all traditional and TSFM baselines, including fine-tuned versions of *Moirai* and *Lag-Llama*. These results confirm that energy-domain-specific pretraining confers robust generalization capabilities even in zero-shot settings.

Residential building datasets present greater variability due to region-specific consumption patterns and appliance diversity. Most traditional and TSFM models perform poorly on datasets such as *CEEW* and *fIEECe*, with NRMSEs exceeding 85 in many cases. Fine-tuned TSFMs show mixed results, often failing to generalize effectively to these unseen domains. In contrast, *W-LSTMix* consistently delivers strong performance, achieving the lowest NRMSE on all six residential OOD datasets (e.g., 11.30 on *HONDA-SH*, 32.42 on *ECWM*, 38.97 on *NREL-Residential*). Even when TSFMs are fine-tuned, *W-LSTMix* either matches or outperforms them, reinforcing the importance of energy-specific pretraining for robust, cross-regional generalization in residential energy forecasting.

4 Conclusion and Future Work

In this study, we demonstrate that domain-specific pre-training substantially enhances the performance of pre-trained models for short-term load forecasting. By pre-training on a large corpus comprising 1.26 billion meter readings from over 76,217 buildings, our proposed architectures Energy-TTMs, MixForecast, and W-LSTMix consistently surpass general-purpose TSFMs.

Our experiments reveal complementary strengths across these models. In zero-shot forecasting on both in-distribution (ID) and out-of-distribution (OOD) datasets, W-LSTMix achieves the lowest normalized NRMSE, reducing error by nearly 40 % compared to the best general TSFM (Chronos). This performance underscores the efficacy of its decomposition-based design in capturing fundamental consumption patterns. Conversely, following supervised fine-tuning, MixForecast attains superior accuracy on both commercial (18.67 NRMSE) and residential (51.27 NRMSE) OOD tasks. This attests to the adaptability of its MLP-Mixer backbone when trained on limited, building-specific data. The robust OOD performance of our models confirms their readiness for deployment in realistic settings, where forecasting for novel buildings is essential.

We acknowledge that our models were pre-trained only on hourly electricity meter readings derived from consumption domain, without incorporating any contextual information such as weather conditions or occupancy patterns. Currently, our dataset includes such contextual variables for only a limited number of buildings (less than 50%). We are actively working to collect historical weather data for all building locations and aim to expand our dataset to include generation-side data such as solar PV and wind energy generation data the the associated contextual data. Furthermore, we plan to pre-train the TTM model, which inherently supports exogenous variables, and extend W-LSTMix to effectively incorporate such contextual inputs. Integrating these additional features is critical for improving the accuracy and robustness of load forecasting models.

Similarly, the Transformer-L (Gaussian) model from BuildingsBench shares this limitation, having been pre-trained exclusively on univariate electricity consumption data – all from simulations. This limitation primarily stems from the limited availability of large-scale, high-quality metadata, which is often not collected or made publicly available by dataset providers. Importantly, this challenge is not unique to the energy domain. Most generic TSFM models are also pre-trained on raw time series data with minimal or no contextual information, primarily due to the scarcity of well-annotated contextual datasets. For example, in environmental monitoring applications, while time series data such as temperature and humidity are widely accessible, critical contextual variables—like land use patterns, vegetation density, or proximity to pollution sources—are rarely included. Yet, such information can significantly enhance the representational capacity and downstream performance of pre-trained models across domains.

Looking ahead, we identify four promising directions:

(1) **Unified Architectures for Generalization and Adaptation.** Investigate hybrid models that combine W-LSTMix's signal-decomposition modules with MixForecast's adaptive mixer layers, aiming to deliver both strong zero-shot generalization and rapid few-shot adaptation.
(2) **Incorporation of Exogenous Variables.** Extend our univariate framework to a multivariate setting by integrating weather, occupancy, and calendar features, and quantify the resulting gains in forecasting accuracy.
(3) **Cross-Domain Transferability.** Evaluate the transfer-learning capabilities of these energy-specific foundation models on related tasks—such as electric-vehicle charging demand and solar-generation forecasting—to assess their broader applicability within the energy ecosystem.
(4) **Edge-Deployment Optimization.** Characterize and optimize inference latency, memory footprint, and power consumption of W-LSTMix and MixForecast for deployment on resource-constrained devices typical of smart-building controllers.

References

[1] S Agrawal, S Mani, K Ganesan, and A Jain. [n. d.]. High frequency smart meter data from two districts in India (Mathura and Bareilly)(2021).

[2] Baldemar Aguirre-Fraire, Jessica Beltrán, and Valeria Soto-Mendoza. 2024. A comprehensive dataset integrating household energy consumption and weather conditions in a north-eastern Mexican urban city. *Data in Brief* 54 (2024), 110452.

[3] Omar Al-Khadher, Azharudin Mukhtaruddin, Fakroul Ridzuan Hashim, Muhammad Mokhzaini Azizan, Hussin Mamat, and Ahmed Aqlan. 2024. CLEMD, a circuit-level electrical measurements dataset for electrical energy management. *Scientific Data* 11, 1 (2024), 594.

[4] Chanuka Algama, Isuruni Fernando, Merl Chandana, Pasindu Ranage, Dinithi Dissanayake, Jesus Ramos, and Kasun Amarasinghe. 2025. Residential Electricity Consumption: Dataset combining Multi-round Longitudinal Surveys and Energy Provider Data. https://doi.org/10.21227/n1dk-q860

[5] Abdullah Alsalemi, Abbes Amira, Hossein Malekmohamadi, and Kegong Diao. 2023. Novel domestic building energy consumption dataset: 1D timeseries and 2D Gramian Angular Fields representation. *Data in Brief* 47 (2023), 108985.

[6] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024. Chronos: Learning the Language of Time Series.

[7] Sotirios Athanasoulias, Fernanda Guasselli, Nikolaos Doulamis, Anastasios Doulamis, Nikolaos Ipiotis, Athina Katsari, Lina Stankovic, and Vladimir Stankovic. 2024. the Plegma dataset: Domestic appliance-level and aggregate electricity demand with metadata from Greece. *Scientific Data* 11, 1 (2024), 376.

[8] Didier Calogine, Johann Francou, Cedric Abbezzot, and Tovondahiniriko Fanjirindratovo. 2023. Data in experimental stand-alone microgrid: Solar production, domestic loads, battery storage and meteorological series. *Data in Brief* 51 (2023), 109643.

[9] Matej Cenkỳ, Jozef Bendík, Boris Cintula, Peter Janiga, Žaneta Eleschová, and Anton Beláň. 2023. Dataset of 15-minute values of active and reactive power consumption of 1000 households during single year. *Data in Brief* 50 (2023), 109588.

[10] Tony Craig and Ian Dent. 2016. North East Scotland Energy Monitoring Project, 2010-2012, Study Level Documentation. *UK Data Service* (2016). https://doi.org/10.5255/UKDA-SN-8122-1

[11] Borui Cui, Sangkeun Lee, Piljae Im, Michael Koenig, and Mahabir Bhandari. 2022. High resolution dataset from a net-zero home that demonstrates zero-carbon living and transportation capacity. *Data in brief* 45 (2022), 108703.

[12] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2023. A decoder-only foundation model for time-series forecasting.

[13] SHIVAM DWIVEDI, Anuj Kumar, Harish Kumar Saravanan, and Pandarasamy Arjunan. 2025. W-LSTMix: A Hybrid Modular Forecasting Framework for Trend and Pattern Learning in Short-Term Load Forecasting. https://openreview.net/forum?id=bG04Z3Jioc

[14] Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam Nguyen, Wesley M Gifford, Chandra Reddy, and Jayant Kalagnanam. 2024. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *Advances in Neural Information Processing Systems* 37 (2024), 74147–74181.

[15] Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. 2024. Tiny Time Mixers (TTMs): Fast Pre-trained Models for Enhanced Zero/Few-Shot Forecasting of Multivariate Time Series. arXiv:2401.03955 [cs.LG] https://arxiv.org/abs/2401.03955

[16] Vijay Ekambaram, Subodh Kumar, Arindam Jati, Sumanta Mukherjee, Tomoya Sakai, Pankaj Dayama, Wesley M. Gifford, and Jayant Kalagnanam. 2025. TSPulse: Dual Space Tiny Pre-Trained Models for Rapid Time-Series Analysis. arXiv:2505.13033 [cs.LG] https://arxiv.org/abs/2505.13033

[17] Patrick Emami, Abhijeet Sahu, and Peter Graf. 2023. Buildingsbench: A large-scale dataset of 900k buildings and benchmark for short-term load forecasting. *Advances in Neural Information Processing Systems* 36 (2023), 19823–19857.

[18] Patrick Emami, Abhijeet Sahu, and Peter Graf. 2024. BuildingsBench: A Large-Scale Dataset of 900K Buildings and Benchmark for Short-Term Load Forecasting. arXiv:2307.00142 [cs.LG] https://arxiv.org/abs/2307.00142

[19] Calvin Goncalves, Ruben Barreto, Pedro Faria, Luis Gomes, and Zita Vale. 2022. Dataset of an energy community's generation and consumption with appliance allocation. *Data in Brief* 45 (2022), 108590.

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

[21] Jung Min Han, Ali Malkawi, Xu Han, Sunghwan Lim, Elence Xinzhu Chen, Sang Won Kang, Yiwei Lyu, and Peter Howard. 2024. A two-year dataset of energy, environment, and system operations for an ultra-low energy office building. *Scientific Data* 11, 1 (2024), 938.

[22] Georges Hebrail and Alice Berard. 2006. Individual Household Electric Power Consumption. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C58K54.

[23] Philipp Heer, Curdin Derungs, Benjamin Huber, Felix Bünning, Reto Fricker, Sascha Stoller, and Björn Niesen. 2024. Comprehensive energy demand and usage data for building automation. *Scientific Data* 11, 1 (2024), 469.

[24] Matthias Hofmann, Sigurd Bjarghov, and Stian Nessa. 2023. Norwegian hourly residential electricity demand data with consumer characteristics during the European energy crisis. *Data in Brief* 51 (2023), 109687.

[25] Matthias Hofmann and Turid Siebenbrunner. 2023. A rich dataset of hourly residential electricity consumption data and survey answers from the iFlex dynamic pricing experiment. *Data in Brief* 50 (2023), 109571.

[26] Mohammed Saeed Jawad, Chitra Dhawale, Abdel Rahman Al Ali, and Azizul Azhar Bin Ramli. 2024. Power spectrum: A detailed dataset on electric demand and environmental interplays. *Data in Brief* 52 (2024), 109788.

[27] Anuj Kumar, Harish Kumar Saravanan, Shivam Dwivedi, and Pandarasamy Arjunan. 2025. MixForecast: Mixer-Enhanced Foundation Model for Load Forecasting. In *Proceedings of the 2nd International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things* (Irvine, CA, USA) *(FMSys)*. Association for Computing Machinery, New York, NY, USA, 25–30. https://doi.org/10.1145/3722565.3727193

[28] Eunjung Lee, Keon Baek, and Jinho Kim. 2022. Datasets on South Korean manufacturing factories' electricity consumption and demand response participation. *Scientific Data* 9, 1 (2022), 227.

[29] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. 2024. Foundation models for time series analysis: A tutorial and survey.

[30] Wei Liao, Xiaoyu Jin, Yi Ran, Fu Xiao, Weijun Gao, and Yanxue Li. 2024. A twenty-year dataset of hourly energy generation and consumption from district campus building energy systems. *Scientific Data* 11, 1 (2024), 1400.

[31] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. 2020. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. arXiv:1912.09363 [cs, stat] http://arxiv.org/abs/1912.09363

[32] Guolong Liu, Jinjie Liu, Yan Bai, Chengwei Wang, Haosheng Wang, Huan Zhao, Gaoqi Liang, Junhua Zhao, and Jing Qiu. 2023. Eweld: A large-scale industrial and commercial load dataset in extreme weather events. *Scientific data* 10, 1 (2023), 615.

[33] Stephen Makonin. 2019. HUE: The hourly usage of energy dataset for buildings in British Columbia. *Data in brief* 23 (2019), 103744.

[34] Stephen Makonin, Bradley Ellert, Ivan V Bajić, and Fred Popowich. 2016. Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. *Scientific data* 3, 1 (2016), 1–12.

[35] Stephen Makonin, Fred Popowich, Lyn Bartram, Bob Gill, and Ivan V. Bajić. 2013. AMPds: A public dataset for load disaggregation and eco-feedback research. In *2013 IEEE Electrical Power & Energy Conference*. 1–6. https://doi.org/10.1109/EPEC.2013.6802949

[36] RO Masebinu and N Kambule. 2022. Electricity consumption data of a middle-income household in Gauteng, South Africa: Pre and Post COVID-19 lockdown (2019-2021). *Data in Brief* 43 (2022), 108341.

[37] SO Masebinu, JB Holm-Nielsen, C Mbohwa, S Padmanaban, and N Nwulu. 2020. Electricity consumption data of a student residence in Southern Africa. *Data in Brief* 32 (2020), 106150.

[38] Christoph J Meinrenken, Noah Rauschkolb, Sanjmeet Abrol, Tuhin Chakrabarty, Victor C Decalf, Christopher Hidey, Kathleen McKeown, Ali Mehmani, Vijay Modi, and Patricia J Culligan. 2020. MFRED, 10 second interval real and reactive power for groups of 390 US apartments of varying size and vintage. *Scientific Data* 7, 1 (2020), 375.

[39] Guy Mélard and J-M Pasteels. 2000. Automatic ARIMA modeling including interventions, using time series expert software. *International Journal of Forecasting* 16, 4 (2000), 497–508.

[40] Clayton Miller, Pandarasamy Arjunan, Anjukan Kathirgamanathan, Chun Fu, Jonathan Roth, June Young Park, Chris Balbach, Krishnan Gowri, Zoltan Nagy, Anthony D Fontanini, et al. 2020. The ASHRAE great energy predictor III competition: Overview and results. *Science and Technology for the Built Environment* 26, 10 (2020), 1427–1447.

[41] Francisco Monteiro, Rafael Oliveira, João Almeida, Pedro Gonçalves, Paulo Bartolomeu, Jorge Neto, and Ricardo Deus. 2024. Electricity consumption dataset of a local energy cooperative. *Data in Brief* 54 (2024), 110373.

[42] Harsha Moraliyage, Nishan Mills, Prabod Rathnayake, Daswin De Silva, and Andrew Jennings. 2022. UNICON: An Open Dataset of Electricity, Gas and Water Consumption in a Large Multi-Campus University Setting. In *2022 15th International Conference on Human System Interaction (HSI)*. 1–8. https://doi.org/10.1109/HSI55341.2022.9869498

[43] 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 *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 169–173.

[44] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers.

[45] Frederick Paige, Philip Agee, and Farrokh Jazizadeh. 2019. flEECe, an energy use and occupant behavior dataset for net-zero energy affordable senior residential buildings. Scientific Data 6.

[46] Sungwoo Park, Seungmin Jung, Seungwon Jung, Seungmin Rho, and Eenjun Hwang. 2021. Sliding window-based LightGBM model for electric load forecasting using anomaly repair. *J. Supercomput.* 77, 11 (Nov. 2021), 12857–12878. https://doi.org/10.1007/s11227-021-03787-4

[47] Energy Group Prayas. 2021. Processed data. https://doi.org/10.7910/DVN/YJ5SP1

[48] Carlos Quesada, Leire Astigarraga, Chris Merveille, and Cruz E Borges. 2024. An electricity smart meter dataset of Spanish households: insights into consumption patterns. *Scientific Data* 11, 1 (2024), 59.

[49] Haroon Rashid, Pushpendra Singh, and Amarjeet Singh. 2019. I-BLEND, a campus-scale commercial and residential buildings electrical energy dataset. *Scientific data* 6, 1 (2019), 1–12.

[50] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, et al. 2024. Lag-llama: Towards foundation models for probabilistic time series forecasting.

[51] Tom Rushby, Benjamin Anderson, Patrick James, and Abubakr Bahaj. 2020. Solent Achieving Value from Efficiency (SAVE) Data, 2017-2018.

[52] Jason Avron Samuels, Terhemba Michael-Ahile, and MJ Thinus Booysen. 2025. Dataset on electricity usage measurement for lower-to-middle-income primary and secondary schools in Western Cape, South Africa. *Data in Brief* (2025), 111321.

[53] Miguel López Santos, Saúl Díaz García, Xela García-Santiago, Ana Ogando-Martínez, Fernando Echevarría Camarero, Gonzalo Blázquez Gil, and Pablo Carrasco Ortega. 2023. Deep learning and transfer learning techniques applied to short-term load forecasting of data-poor buildings in local energy communities. *Energy and Buildings* 292 (2023), 113164.

[54] Milagros Santos-Moreno, Jorge Ángel González-Ordiano, J Emilio Quiroz-Ibarra, DA Perez-DeLaMora, Jaime Mizrahi-Cojab, Emilio Román-Sánchez, José Pablo Montero-Cantú, and Lazaro Bustio-Martinez. 2024. Weather and electrical demand and consumption data of a small Mexican community. *Data in Brief* 52 (2024), 109977.

[55] Harish Kumar Saravanan, Shivam Dwivedi, P Praveen, 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*. 346–349.

[56] Marlon Schlemminger, Tobias Ohrdes, Elisabeth Schneider, and Michael Knoop. 2022. Dataset on electrical single-family house and heat pump load profiles in Germany. *Scientific data* 9, 1 (2022), 56.

[57] Changho Shin, Eunjung Lee, Jeongyun Han, Jaeryun Yim, Wonjong Rhee, and Hyoseop Lee. 2019. The ENERTALK dataset, 15 Hz electricity consumption data from 22 houses in Korea. *Scientific data* 6, 1 (2019), 193.

[58] Rohit Trivedi, Mohamed Bahloul, Aziz Saif, Sandipan Patra, and Shafi Khadem. 2024. Comprehensive dataset on electrical load profiles for energy community in ireland. *Scientific Data* 11, 1 (2024), 621.

[59] AC Tsakoumis, SS Vladov, and VM Mladenov. 2002. Daily load forecasting based on previous day load. In *6th Seminar on Neural Network Applications in Electrical Engineering*. IEEE, 83–86.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[61] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers.

[62] Neo Wu, Bradley Green, Xue Ben, and Shawn O'Banion. 2020. Deep transformer models for time series forecasting: The influenza prevalence case. *arXiv preprint arXiv:2001.08317* (2020).

[63] Wutao Xiong. 2024. UCI dataset. https://doi.org/10.21227/e40r-rf81

[64] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2022. Are Transformers Effective for Time Series Forecasting? arXiv:2205.13504 [cs.AI] https://arxiv.org/abs/2205.13504

[65] Kaile Zhou, Dingding Hu, Rong Hu, and Jiong Zhou. 2023. High-resolution electric power load data of an industrial park with multiple types of buildings in China. *Scientific Data* 10, 1 (2023), 870.

Table 9: Pre-training Data - Commercial Buildings

Datasets	# Buildings	Location	Year	# Obs
Commercial Buildings				
IBlend [49]	8	India	2013-17	261895
CLEMD [3]	1	Malaysia	2023-23	706
EWELD [32]	386	China	2016-22	13730258
SKC [28]	8	South Korea	2019	40452
SEWA [26]	1	UAE	2020-21	17489
PSS [52]	50	South Africa	2022-23	428570
Enernoc[14]	90	USA	2012-13	790206
DGS[15]	316	USA	2016-18	5227713
BDG-2 [40]	1500	USA	2016-17	23926381
ULE [21]	1	Cambridge	2023-24	8783
UCIE [63]	370	Portugal	2011-15	12974050
NEST [23]	1	Switzerland	2019-23	34798
UNICON [42]	60	Australia	2018-22	1950238
Total	2,792			59M

Table 10: Evaluation Data - Out-of-Distribution

Dataset	# Buildings	Location	Year	# Obs
Commercial Buildings				
IPC-Commercial [65]	3	China	2016-21	132519
NREL [18]	250	USA	2018	2190000
Residential Buildings				
CEEW [1]	84	India	2019-21	923897
MIHEC [36]	1	South Africa	2019-21	19358
RHC [34]	1	Canada	2012-14	17520
fIEECe [45]	6	USA	2017-18	74492
Honda-SH [11]	1	USA	2020	8760
ECWM [2]	1	Mexico	2022-24	10184
NDB [5]	1	UK	2022	3282
NREL [18]	150	USA	2018	1314000

Table 11: Evaluation Data - In-Distribution

Test	# Buildings	Location	Year	# Obs
Commercial Buildings				
IBlend	1	India	2013-17	34462
SKC	2	South Korea	2019	10256
PSS	3	South Africa	2022-23	26204
Enernoc	10	USA	2012-13	87522
BDG-2	78	USA	2016-17	1286198
UNICON	4	Australia	2018-22	116063
Residential Buildings				
RSL	100	Sri Lanka	2023-24	394785
HUE	8	Canada	2012-20	188898
PES	2	USA	2011-14	10707
SAVE	91	UK	2016-18	1074267
HES	2	UK	2010-11	9432
UKST	219	UK	2008-10	3205499
IRH	4	Ireland	2020-21	35083
DESM	1	France	2020-21	8621
NEEA	6	Portland	2018-20	82883
GoiEner	300	Spain	2014-22	6766058
Plegma	3	Greece	2022-23	21211
HSG	2	Germany	2014-19	47238
iFlex	29	Norway	2020-21	87936
Norwegian	36	Norway	2020-22	472608
ECCC	3	Portugal	2019-20	26352
NESEMP	15	Scotland	2010-12	129189
DTH	100	Slovak Republic	2016-17	944379
SGSC	250	Australia	2011-14	3117482

A Dataset Details

Tables 8, 9, 10, and 11 show the summary of datasets used for pre-training and evaluation of models.

Table 8: Pre-training Data - Residential Buildings

Datasets	# Buildings	Location	Year	# Obs
Residential Buildings				
Prayas [47]	116	India	2018-20	1536409
RSL [4]	2803	Srilanka	2023-24	10978185
IPC-Residential [65]	1	China	2016-21	45471
ENERTALK [57]	18	South Korea	2016-17	32634
DCB [30]	1	Japan	2002-22	187008
SRSA [37]	1	South Africa	2016-18	13966
AMPD [35]	1	Canada	2012-14	17520
HUE [33]	20	Canada	2012-20	422051
SMART-Star[6]	114	USA	2014-16	2166462
MFRED [38]	26	USA	2019-20	227622
PES[7]	8	USA	2011-14	2801
WED [54]	5	Mexico	2022-23	45096
SAVE [51]	4600	UK	2016-18	58844644
HES[8]	14	UK	2010-11	106392
LCL[9]	5561	UK	2011-14	83925571
UKST[10]	14100	UK	2008-10	204373784
IRH [58]	16	Ireland	2020-21	139315
IHEPC [22]	1	France	2006-10	34168
DESM [8]	2	France	2020-21	15397
NEEA[11]	200	Portland	2018-20	2839406
GoiEner [48]	25259	Spain	2014-22	625547875
Plegma [7]	10	Greece	2022-23	62241
SFHG [56]	38	Germany	2019-20	769247
HSG[12]	9	Germany	2014-19	173537
iFlex [25]	4400	Norway	2020-21	13427664
Norwegian [24]	1100	Norway	2020-22	14440800
LEC [41]	166	Portugal	2022-23	1457772
ECCC [19]	48	Portugal	2019-20	421632
NESEMP [10]	200	Scotland	2010-12	1446493
NEST-Res [23]	2	Switzerland	2019-23	69465
DTH [9]	900	Slovak Republic	2016-17	8508672
SGSC[13]	13485	Australia	2011-14	169159731
Total	73,425			1.2B