TempusBench: An Evaluation Framework for Time-Series Forecasting

Denizalp Goktas* **Gerardo Riaño-Briceño*** **Alif Abdullah** **Aryan Nair**

Chenkai Shen **Beatriz de Lucio** **Alexandra Magnusson** **Farhan Mashrur**

Ahmed Abdulla **Shawrna Sen** **Mahitha Thippireddy** **Gregory Schwartz**

Amy Greenwald

Simulacrum
New York City, NY, USA
{deni, gerardo, amy}@smlcrm.com

Abstract

Foundation models have transformed natural language processing and computer vision, and a rapidly growing literature on time-series foundation models (TSFMs) seeks to replicate this success in forecasting. While recent open-source models demonstrate the promise of TSFMs, the field lacks a comprehensive and community-accepted model evaluation framework. We see at least four major issues impeding progress on the development of such a framework. First, current evaluation frameworks consist of benchmark forecasting tasks derived from often outdated datasets (e.g., M3), many of which lack clear metadata and overlap with the corpora used to pre-train TSFMs.Second, existing frameworks evaluate models along a narrowly defined set of benchmark forecasting tasks such as forecast horizon length or domain, but overlook core statistical properties such as non-stationarity and seasonality. Third, domain-specific models (e.g., XGBoost) are often compared unfairly, as existing frameworks neglect a systematic and consistent hyperparameter tuning convention for all models. Fourth, visualization tools for interpreting comparative performance are lacking. To address these issues, we introduce TempusBench, an open-source evaluation framework. TempusBench consists of 1) new datasets which are not included in existing TSFM pretraining corpora, 2) a set of novel benchmark tasks that go beyond existing ones, and 3) a model evaluation pipeline with a standardized hyperparameter tuning protocol, and 4) a tensorboard-based visualization interface. We provide access to our code on GitHub: https://github.com/Smlcrm/TempusBench.

1 Introduction

The success of foundation models (i.e., models trained on large and diverse datasets that can be used to solve downstream tasks) in natural language processing (NLP) and computer vision has inspired an emerging literature on time-series foundation models. *Time-series foundation models (TSFMs)* are models that take past time-series data (and possibly covariate time-series data) as input

*Equal contribution.

NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models (BERT^{2}S).

and output future values (or distributions over them), typically formulated as neural networks trained via supervised learning. While about a dozen open-source TSFMs are now available, comparing their performance to one another and to traditional domain-specific models (e.g., ARIMA [1], SVR [2, 3]) remains difficult. A handful of evaluation frameworks have been released, but the field still lacks comprehensive, community-accepted standards for model evaluation [4], creating an impediment for the replication of the success of foundation models in NLP and computer vision [5].

We see four major challenges facing existing evaluation frameworks. First, the evaluation ecosystem relies on outdated datasets such as M3 [6] and M4 [7], many lacking metadata (e.g., variable names). More importantly, the existing evaluation datasets overlap with the pretraining corpora of TSFMs, leading to inflated estimates of zero-shot generalization [8]. For example, except for Moirai2, all TSFMs assessed by GIFT-Eval include test data in their training corpus [9, 10]. Second, current frameworks define benchmark forecasting tasks only along narrow axes (i.e., forecast horizon, variate type, frequency, and domain). While useful, these miss key statistical properties long studied in time-series analysis such as (non-)stationarity, and seasonality. Without evaluation across such properties, it seems unlikely that frameworks can yield generalizable conclusions about model capabilities. Third, existing frameworks have not yet developed standardized hyperparameter tuning routines, leading to comparisons made between TSFMs and domain-specific models to be unfair as the performance of domain-specific models depend heavily on hyperparameter choice.[2] Indeed, as noted by practitioners [11], simple statistical models with well-chosen hyperparameters can outperform more complex ones, highlighting the need for consistent tuning routines. Fourth, currently, evaluation typically reduces to numerical metrics such as mean squared error, which practitioners remark [12] provide limited interpretability. For instance, under GIFT-Eval, seasonal naive outperforms five open-source TSFMs, but this offers no insight into the strength and weaknesses of TSFMs, since seasonal naive fails when seasonality is weak. Beyond quantitative scores, qualitative analyses—especially forecast visualizations—are essential.

To address these issues, we introduce TempusBench, an open-source evaluation framework. TempusBench consists of 1) new datasets which are not included in existing TSFM pretraining corpora, 2) a set of novel benchmark tasks that goes beyond existing ones, , and 3) a model evaluation pipeline with a standardized hyperparameter tuning protocol, and 4) tensorboard-based visualization interface.

1.1 Contributions

TempusBench, going beyond TSFMs, includes 20 forecasting models , a number of which such as XGBoost, have previously not been considered by evaluation frameworks, and overcomes the aforementioned four issues by improving along the following dimensions. First, we introduce new time-series datasets which do not come from existing time-series evaluation datasets, and which are not contained in the training corpus of open-source TSFMs released to date. Second, we propose new benchmark task types that extend beyond horizon length, variate type, frequency, and domain. These include categories based on stationarity, seasonality, variable type (continuous, count, binary, categorical), sparsity (sparse vs. dense), dataset size (small vs. large), and quality (noisy vs. measurement error). Third, we introduce a model evaluation pipeline which runs a standardized and automated hyperparameter selection procedure for all forecasting models with hyperparameters, allowing a fair comparison of all forecasting methods. Fourth, TempusBench comes packages with a tensorboard-based visualization application which easily allows researchers and practitioners to visualize and interpret the performance of various models on different task types.

2 Background

We refer the reader to Appendix A for the notational convention we adopt, as well as for additional mathematical preliminaries and evaluation metric definitions.

Forecasters A *(time-series) forecasting task* $\mathcal{T} \doteq (l, h, n, m, \mathcal{X}, \mathcal{Y}, \boldsymbol{X}, \boldsymbol{Y})$ consists of a *context length* $l \in \mathbb{N}$, a *forecast horizon* $h \in \mathbb{N}$, $m \in \mathbb{N}$ *target time-series* $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)^T$ where for each variate $i \in [m]$, entries of $\boldsymbol{y}_i \in \mathcal{Y}_i^l$ take values from a *set of target values* $\mathcal{Y} \subseteq \mathbb{R}$, and $n \in \mathbb{N}$ *covariate time-series* $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ where for each covariate $j \in [n]$, $\boldsymbol{x}_j \in \mathcal{X}^{l+h}$ takes values from a *set of covariate values* $\mathcal{X}_j \subseteq \mathbb{R}$. For convenience, we denote the *joint set of*

[2] TSFMs require hyperparameter searches during pretraining, but not during evaluation.

Table 1: Property comparisons of various forecasting benchmarks.

Property	Monash [13]	TFB [14]	LTSF [15]	BasicTS+ [16]	ProbTS [17]	GIFT-Eval [9]	TempusBench
Frequency Range	Second to Year	Minute to Year	Minute to Week	Minute to Day	Minute to Week	Second to Year	Second to Year
Num. Domains	7	6	5	3	5	7	10
Train/Test data leak	✓	✓	✓	✓	✓	✓	✗
Variate Types	Uni	Uni/Multi	Multi	Multi	Multi	Uni/Multi	Uni/Multi
Prediction Length	Short	Short	Long	Short/Long	Short/Long	Short/Long	Short/Long
Stat. Benchmarks	✗	✗	✗	✗	✗	✗	✓
Forecaster types	Stat./DL	Stat./DL	Stat./DL	Stat./DL	Stat./DL/FM	Stat./DL/FM	Stat./ML/DL/FM
Hyperparam. autotuning	✗	✗	✗	✗	✗	✗	✓

target variate values by $\mathcal{Y} \doteq \bigtimes_{i\in[m]} \mathcal{Y}_i$ and the *joint set of covariate values* $\mathcal{X} \doteq \bigtimes_{j\in[n]} \mathcal{X}_j$. A forecasting task $\mathcal{T}$ is said to be *univariate* (resp. *multivariate*) iff $m = 1$ ($m > 1$). A forecasting task $\mathcal{T}$ is said to be *unconditional* (resp. *conditional*) iff $n = 0$ (resp. $n > 0$). A forecasting task $\mathcal{T}$ is said to be a continuous (resp. count | categorical | binary) forecasting task iff for all $i \in [m]$ $\mathcal{Y}_i \subseteq \mathbb{R}$ is a continuous set (resp. $\mathcal{Y}_i = \mathbb{N} \mid \mathcal{Y}_i \subsetneq \mathbb{N} \mid \mathcal{Y}_i = \{0, 1\}$). A *(point) forecast* for a forecasting task $\mathcal{T}$ is a matrix $\widehat{\boldsymbol{Y}} \doteq (\widehat{\boldsymbol{y}}_1, \ldots, \widehat{\boldsymbol{y}}_m)$ s.t. for all target variates $i \in [m]$, $\widehat{\boldsymbol{y}}_i \in \mathcal{Y}_i^h$ corresponds to forecasted values of variate i for h steps. A *(point) forecasting model* (or, colloquially, a *forecaster)* is a mapping $\boldsymbol{F} : \mathcal{X}^{l+h} \times \mathcal{Y}^l \to \mathcal{Y}^h$ s.t. $\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{Y}) \doteq (\boldsymbol{f}_1(\boldsymbol{X}, \boldsymbol{Y}), \ldots, \boldsymbol{f}_m(\boldsymbol{X}, \boldsymbol{Y}))^T$ is a forecast for $\mathcal{T}$.[3] A *probabilistic forecasting model* is a mapping $\pi : \mathcal{X}^{l+h} \times \mathcal{Y}^l \to \Delta(\mathcal{Y}^h)$ s.t. $\pi(\boldsymbol{X}, \boldsymbol{Y})[\widehat{\boldsymbol{Y}}] \geq 0$ denotes the probability of $\widehat{\boldsymbol{Y}} \in \mathcal{Y}^l$ being realized.

Forecasting Evaluation Frameworks In reality, many forecasters $\boldsymbol{F}^{\boldsymbol{\theta}} : \mathcal{X}^{l+h} \times \mathcal{Y}^l \to \mathcal{Y}^h$ are dependent on some hyperparameters $\boldsymbol{\theta} \in \Theta$, and it is more appropriate to talk about a *family of forecasters* $\mathcal{F}^{\Theta} \doteq \{\boldsymbol{F}^{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$, and choose the forecaster with parameters which is the most appropriate for a forecasting task.[4] A forecaster evaluation framework $\mathcal{B} \doteq (p, q, \mathcal{E}, \{\Theta_i\}_{i=1}^p, \{\mathcal{F}^{\Theta_i}\}_{i=1}^p, \{\mathcal{T}_j\}_{j\in[q]})$ consists of $p \in \mathbb{N}$ familes of forecasters, with for each $i \in [p]$, $\mathcal{F}^{\Theta_i}$ being defined by a set of hyperparameters Θ_i; $q \in \mathbb{N}$ forecasting tasks (or, colloquially, *benchmarks*, or *benchmark tasks*) $\{\mathcal{T}_j\}_{j\in[q]}$; and a hyperparameter tuner $\mathcal{E}$, which takes as input a benchmark and outputs some hyperparameters.

3 TempusBench

We describe in Section 4, additional extensions of TempusBench which will be released in the full-paper version. TempusBench, denoted $\mathcal{B}^{\mathrm{TB}}$, is a forecasting evaluation framework where the hyperparameter tuner $\mathcal{E}^{\mathrm{TB}}$ is given by three-step procedure: given a benchmark, a (sub)set of hyperparameters, and a family of forecasters, 1) a validation dataset of subsets of the target and covariate time-series are created, 2) for each hyperparameter in the (sub)set of hyperparameters, the average MSE is computed across all samples in the validation dataset, 3) the hyperparameter with the lowest MSE is output. We summarize the set of families of forecasters, and the set of benchmarks included in TempusBench in Table 12 (Appendix D) and Table 14 (Appendix D.1) respectively. See Appendix B for additional details on computation.

4 Next Directions and Conclusion

We omit for the workshop version of TempusBench two directions in which we have been developing TempusBench, namely the inclusion of conditional forecasting problems. We plan to release this more general version of TempusBench in the coming months as part of the full-version of our paper.

[3] While our definition is in line with the literatur [9], more generally, a forecaster can be defined as a mapping from forecasting tasks to forecasts, i.e., $\mathcal{T} \mapsto \boldsymbol{F}(\mathcal{T}) = \widehat{\boldsymbol{Y}}$.

[4] For instance, the forecast of an ARIMA model is dependent on choices of hyperparameters given by the order of number of time lags, the degree of differencing, and the order of the moving-average model, and it is more appropriate to talk of the family of ARIMA models.

Table 2: Taxonomy of all univariate and multivariate benchmark tasks included in TempusBench.

Category	**Benchmark Tasks**
Movement	Stationary, Non-Stationary
Data Quality	Noisy data, Data with measurement error
Frequency	Seconds, Minutes, Hours, Days, Weeks, Months, Quarterly, Years
Context Length	30, 100, 500, 1000
Forecast Horizon	1, 20, 100, 500, 1000
Seasonality	Cyclical, Non-Stationary cyclical, Regressive, Irregular, Additive, Multiplicative
Domain	Energy, Transport, Climate, Software, Web, Sales, Nature, Econ., Healthcare, Manufacturing
Dataset Coverage	sparse, dense
Target Type	continuous, count, binary, categorical

Table 3: Average Win Rates for deterministic and probabilistic forecasting models.

(a) Average Win Rate for MAPE Metric.

Model Name	**Average Win Rate**
Lafn	0.7931
Timesfm	0.6730
Croston Classic	0.6164
Seasonal Naive	0.5849
Toto	0.5789
Varmax	0.5714
Arima	0.5346
Moment	0.5220
Lagllama	0.5031
Lstm	0.4969
Moirai	0.4966
Svr	0.4874
Tabpfn	0.4828
Random Forest	0.4748
Tiny Time Mixer	0.4151
Chronos	0.4025
Exponential Smoothing	0.3333
Prophet	0.3333
Theta	0.3300

(b) Average Win Rate for CRPS Metric

Model Name	**Average Win Rate**
Toto	1.0000
Moirai	0.7857
Lafn	0.5714
Chronos	0.4000
Lagllama	0.2667

We expect that the datasets used to define our benchmarks will eventually get included in the pretaining corpus of TSFMs, as has been the case often with NLP benchmarks. To this end, we are developing dynamic benchmarks where test data is continuously refreshed. While dynamic benchmarks can easily be defined benchmarks making use of synthetic data (e.g., our seasonality benchmarks) by continuously generating new datasets, for other benchmarks (e.g., our domain benchmarks) we are building a rotating set of datasets which are pulled from live data APIs.

Finally, for the workshop version of TempusBench, in line with existing forecasting evaluation frameworks, we consider benchmark categories such as target variate type, context length, forecast length as defining individual forecasting tasks. However, a more comprehensive way to see these benchmark categories would be as hyperparameters for other benchmark categories such as domains. That is, for instance, a more comprehensive list of benchmarks would test the performance of forecasting models for each domain (e.g., economics) for different choices of target variate types, context lengths, and forecast lengths) We are planning to release these more comprehensive benchmark types in the coming months as part of the full-version of our paper.

References

[1] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 1970. 2

[2] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995. 2

[3] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alexander Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, 1997. 2

[4] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6555–6565, 2024. 2, 34

[5] Prajakta S Kalekar et al. Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi school of information Technology*, 4329008(13):1–13, 2004. 2, 25, 43, 44, 45

[6] Spyros Makridakis and Michele Hibon. The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476, 2000. 2, 42

[7] Spyros Makridakis, Evangelos Spiliotis, and Vassilis Assimakopoulos. The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4): 802–808, 2018. 2, 42

[8] Pooja Anand, Mayank Sharma, and Anil Saroliya. A comparative analysis of artificial neural networks in time series forecasting using arima vs prophet. In *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*, pages 527–533. IEEE, 2024. 2, 26, 42, 43, 44, 45

[9] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024. 2, 3

[10] Salesforce. Gift-eval. Hugging Face Space, 2024. URL https://huggingface.co/spaces/Salesforce/GIFT-Eval. Accessed: 2025-08-29. 2

[11] u/nkafr. The rise of foundation time-series forecasting models. https://www.reddit.com/r/datascience/comments/1e865bt/the_rise_of_foundation_timeseries_forecasting/, 2024. URL https://www.reddit.com/r/datascience/comments/1e865bt/the_rise_of_foundation_timeseries_forecasting/. Reddit post on r/datascience. 2

[12] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006. 2

[13] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, et al. Monash time series forecasting archive. In *NeurIPS Datasets and Benchmarks Track*, 2021. 3, 42

[14] Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *arXiv preprint arXiv:2403.20150*, 2024. 3, 42

[15] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023. 3, 42

[16] Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, et al. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 3

[17] Jiawen Zhang, Xumeng Wen, Zhenwei Zhang, Shun Zheng, Jia Li, and Jiang Bian. Probts: Benchmarking point and distributional forecasting across diverse prediction horizons. *Advances in Neural Information Processing Systems*, 37:48045–48082, 2024. 3

[18] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024. 20, 34

[19] Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024. 20

[20] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024. 20, 21, 34

[21] 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. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024. 21, 34, 42

[22] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: How tabpfn-v2 outperforms specialized time series forecasting models. *arXiv preprint arXiv:2501.02945*, 2025. 22

[23] Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam Nguyen, Wesley M Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *Advances in Neural Information Processing Systems*, 37:74147–74181, 2024. 22

[24] 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. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023. 23

[25] Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, et al. This time is different: An observability perspective on time series foundation models. *arXiv preprint arXiv:2505.14766*, 2025. 23, 45

[26] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024. 24, 42, 45

[27] Sima Siami-Namini and Akbar Siami Namin. Forecasting economics and financial time series: Arima vs. lstm. *arXiv preprint arXiv:1803.06386*, 2018. 24, 26, 42, 43, 44, 45

[28] Thomas R Willemain, Charles N Smart, Joseph H Shockor, and Philip A DeSautels. Forecasting intermittent demand in manufacturing: a comparative evaluation of croston's method. *International Journal of forecasting*, 10(4):529–538, 1994. 25

[29] Lin Lin, Fang Wang, Xiaolong Xie, and Shisheng Zhong. Random forests-based extreme learning machine ensemble for multi-regime time series prediction. *Expert Systems with Applications*, 83:164–176, 2017. 27

[30] Senyao Wang and Jin Ma. A novel ensemble model for load forecasting: Integrating random forest, xgboost, and seasonal naive methods. In *2023 2nd Asian Conference on Frontiers of Power and Energy (ACFPE)*, pages 114–118. IEEE, 2023. 27, 29, 43, 44, 45

[31] Fan Zhang and Lauren J O'Donnell. Support vector regression. In *Machine learning*, pages 123–140. Elsevier, 2020. 28, 43

[32] Dimitrios D Thomakos and Konstantinos Nikolopoulos. Forecasting multivariate time series with the theta method. *Journal of Forecasting*, 34(3):220–229, 2015. 28, 42, 43, 44

[33] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 95–104, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210006. URL https://doi.org/10.1145/3209978.3210006. 31

[34] Indeed. Software Development Job Postings on Indeed in the United States [IHLIDXUSTPSOFTDEVE]. https://fred.stlouisfed.org/series/IHLIDXUSTPSOFTDEVE, 2025. Retrieved August 29, 2025. 31

[35] Nisarg Chodavadiya. Daily Gold Price (2015-2021) Time Series. https://www.kaggle.com/datasets/nisargchodavadiya/daily-gold-price-20152021-time-series, 2025. Accessed on August 29, 2025. 31

[36] Coinbase. Coinbase Litecoin [CBLTCUSD]. https://fred.stlouisfed.org/series/CBLTCUSD, 2025. Retrieved August 29, 2025. 31

[37] IgnacioQG. 2001-2022 Hourly Dataset of Pollution in Madrid. https://www.kaggle.com/datasets/ignacioqg/20012022-hourly-dataset-of-pollution-in-madrid, 2022. Accessed on August 29, 2025. 31

[38] DeltaTrup. LT 1-Minute Historical Stock Data (2003-2024). https://www.kaggle.com/datasets/deltatrup/lt-1-minute-historical-stock-data-2003-2024, may 2024. Accessed on August 29, 2025. 31

[39] GabrielSantello. Airline Baggage Complaints – Time Series Dataset. https://www.kaggle.com/datasets/gabrielsantello/airline-baggage-complaints-time-series-dataset, 2023. Accessed on August 29, 2025. 31

[40] U.S. Census Bureau. Manufacturers: Inventories to Sales Ratio [MNFCTRIRSA]. https://fred.stlouisfed.org/series/MNFCTRIRSA, 2025. Retrieved August 29, 2025. 31

[41] Bank for International Settlements. Real Residential Property Prices for Germany [QDER628BIS]. https://fred.stlouisfed.org/series/QDER628BIS, 2025. Retrieved August 29, 2025. 31

[42] Energy and Geoscience Institute at the University of Utah. Utah FORGE: Well 16A(78)-32 Drilling Data. Accessed via Data.gov, 2025. Accessed on August 29, 2025. 31

[43] Board of Governors of the Federal Reserve System (US). Federal Funds Effective Rate [FF]. https://fred.stlouisfed.org/series/FF, 2025. Retrieved August 29, 2025. 31

[44] U.S. Bureau of Economic Analysis. Personal Consumption Expenditures: Chain-type Price Index [DPCERG3A086NBEA]. https://fred.stlouisfed.org/series/DPCERG3A086NBEA, 2025. Retrieved August 29, 2025. 31

[45] SumanthVrao. Daily Climate Time Series Data. https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data, 2021. Accessed on August 29, 2025. 31

[46] BITS Pilani - Goa. SplitSmart: An Open Dataset for Enabling Research in Energy-Efficient Ductless-Split Air Conditioner, 2024. Accessed on August 29, 2025. 31

[47] City of New York. COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths, 2025. Accessed on August 29, 2025. Daily count of NYC residents who tested positive for SARS-CoV-2, hospitalized with COVID-19, and deaths among COVID-19 patients. 31

[48] U.S. Bureau of Labor Statistics. All Employees, Health Care [CES6562000101]. https://fred.stlouisfed.org/series/CES6562000101, 2025. Retrieved August 29, 2025. 31

[49] NoeyIsLearning. Soil and Environmental Monitoring. https://www.kaggle.com/datasets/noeyislearning/soil-and-environmental-monitoring, 2024. Accessed on August 29, 2025. 31

[50] Riccardo Taormina et al. The Battle of the Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks. *Journal of Water Resources Planning and Management*, 144(8):04018048, aug 2018. doi: 10.1061/(ASCE)WR.1943-5452.0000969. URL https://www.batadal.net/data.html. 31

[51] Jorge Bañuelos-Gimeno, Natalia Sobrino, and Rosa Arce-Ruiz. Initial Insights into Teleworking's Effect on Air Quality in Madrid City. *Environments*, 11(9):204, 2024. doi: 10.3390/environments11090204. URL https://www.mdpi.com/2076-3298/11/9/204. 31

[52] RaminHuseyn. Web Traffic Time Series Dataset. https://www.kaggle.com/datasets/raminhuseyn/web-traffic-time-series-dataset, 2024. Accessed on August 29, 2025. 31

[53] UCI Machine Learning Repository. Hungarian Chickenpox Cases. https://doi.org/10.24432/C5103B, 2021. 31

[54] Alistair Johnson et al. MIMIC-III Clinical Database Demo (version 1.4). https://doi.org/10.13026/C2HM2Q, 2019. RRID:SCR_007345. 31

[55] Andrea Martiniano and Ricardo Ferreira. Absenteeism at work. https://doi.org/10.24432/C5X882, 2012. 31

[56] Daqing Chen. Online Retail. https://doi.org/10.24432/C5BW33, 2015. 31

[57] Paulo Cortez and Aníbal Morais. Forest Fires. https://doi.org/10.24432/C5D88D, 2007. 31

[58] Adarsh Pal Singh and Sachin Chaudhari. Room Occupancy Estimation. https://doi.org/10.24432/C5P605, 2018. 31

[59] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 5 edition, 2015. 42

[60] Rob J. Hyndman, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer, 2008. 42

[61] Vassilis Assimakopoulos and Konstantinos Nikolopoulos. The theta model: A decomposition approach to forecasting. *International Journal of Forecasting*, 16(4):521–530, 2000. 42

[62] Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005. 42

[63] Valentin Flunkert, David Salinas, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks, 2017. 42

[64] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2020. 42

[65] Abhimanyu Das et al. Long-term forecasting with tide: Time-series dense encoder. In *Neural Information Processing Systems (NeurIPS)*, 2023. 42

[66] Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2019. 42

[67] Yifan Nie, Zhihan Huang, Li Wang, Yuheng Sun, Yating He, and Zhifeng Zhang. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023. 42

[68] Han Liu et al. itransformer: Inverted transformers are effective for time series forecasting, 2023. 42

[69] Kashif Rasul, V Ashkinazi, I Schuster, A Schneider, and A Mishkin. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Neural Information Processing Systems (NeurIPS)*, 2021. 42

[70] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Neural Information Processing Systems (NeurIPS)*, 2021. 42

[71] Kashif Rasul et al. Multivariate probabilistic time series forecasting via conditioned normalizing flows, 2020. 42

[72] Sangwoo Woo et al. Moirai: Foundation models for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024. 42

[73] Abhimanyu Das et al. Timesfm: Time series foundation models at scale, 2023. 42

[74] Kashif Rasul et al. Lag-llama: Towards foundation models for time series forecasting, 2023. 42

[75] Yan Liu et al. Timer: Efficient time-series foundation model, 2024. 42

[76] Tian Gao et al. Units: Universal time series foundation models, 2024. 42

[77] Vinay Ekambaram et al. Multi-level tiny time mixers for efficient time-series foundation models, 2024. 42

[78] Xing Chen et al. Visionts: Multimodal time-series foundation models, 2024. 42

[79] Spyros Makridakis, Evangelos Spiliotis, and Vassilis Assimakopoulos. The M5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting*, 38(4): 1346–1364, 2022. 42

[80] Haixu Zhang et al. Probts: Benchmarking probabilistic forecasting. In *Neural Information Processing Systems (NeurIPS)*, 2023. 42

[81] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1): 37–45, 2018. 42

[82] Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, et al. sktime: A unified interface for machine learning with time series. In *Proceedings of the 2nd Workshop on Systems for ML (NeurIPS)*, 2019. 42

[83] Alexander Alexandrov et al. Gluonts: Probabilistic time series models in python, 2020. 42

[84] Kashif Rasul. Pytorchts: A probabilistic deep learning library for time series, 2021. GitHub repository: https://github.com/zalandoresearch/pytorch-ts. 42

[85] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. 42

[86] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. 42

[87] Raj Kumar Tamatta. *Time series forecasting of hospital Inpatients and Day case waiting list using ARIMA, TBATS and Neural Network Models*. PhD thesis, Dublin, National College of Ireland, 2018. 42

[88] Caitlin Haskins. 5 Financial Forecasting Methods to Help Your Business. https://online.hbs.edu/blog/post/financial-forecasting-methods, apr 2021. Accessed on August 14, 2025. 43

A Additional Mathematical Background

A.1 Mathematical notation

We adopt the following calligraphic conventions to insist on the nature of the mathematical object at hand: We use calligraphic uppercase letters to denote sets (e.g., $\mathcal{X}$), bold uppercase letters to denote matrices (e.g., $\boldsymbol{X}$), bold lowercase letters to denote vectors (e.g., $\boldsymbol{p}$), lowercase letters to denote scalar quantities (e.g., x), and uppercase letters to denote random variables (e.g., X). We denote the ith row vector of a matrix (e.g., $\boldsymbol{X}$) by the corresponding bold lowercase letter with subscript i (e.g., $\boldsymbol{x}_i$). Similarly, we denote the jth entry of a vector (e.g., $\boldsymbol{p}$ or $\boldsymbol{x}_i$) by the corresponding lowercase letter with subscript j (e.g., p_j or x_{ij}). We denote functions by a letter determined by the value of the function, e.g., f if the mapping is scalar valued, $\boldsymbol{f}$ if the mapping is vector valued, and $\mathcal{F}$ if the mapping is set valued.

We denote the set $\{1, \ldots, n\}$ by $[n]$, the set $\{n, n+1, \ldots, m\}$ by $[n : m]$, the set of natural numbers by $\mathbb{N}$, and the set of real numbers by $\mathbb{R}$. We denote the positive and strictly positive elements of a set using a $+$ or $++$ subscript, respectively, e.g., $\mathbb{R}_+$ and $\mathbb{R}_{++}$. For any $n \in \mathbb{N}$, we denote the n-dimensional vector of zeros and ones by $\mathbf{0}_n$ and $\mathbf{1}_n$, respectively.

A.2 Mathematical Definitions

We let $\Delta_n = \{\boldsymbol{x} \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_i = 1\}$ denote the unit simplex in $\mathbb{R}^n$, and $\Delta(A)$ denote the set of all probability measures over a given set A. We also define the support of a probability density function $f \in \Delta(\mathcal{X})$ as $\operatorname{supp}(f) \doteq \{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) > 0\}$. Finally, we denote the orthogonal projection operator onto a set C by Π_C, i.e., $\Pi_C(\boldsymbol{x}) \doteq \arg\min_{\boldsymbol{y} \in C} \|\boldsymbol{x} - \boldsymbol{y}\|^2$.

A.3 Evaluation Metrics

An evaluation metric $\ell : \mathcal{Y}^h \times \mathcal{Y}^h \to \mathbb{R}_+$ is a positive-, scalar-valued function s.t. for any forecast $\widehat{\boldsymbol{Y}} \in \mathcal{Y}^h$ and realized future target values $\boldsymbol{Y}^* \in \mathcal{Y}^h$, $\ell(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}^*) \geq 0$ denotes the distance between the forecast and the realized values. We consider the following evaluation metrics at present. The *mean absolute error (MAE)* is defined as $\ell^{\mathrm{MAE}}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}^*) \doteq \frac{1}{mh} \sum_{i \in [m]} \sum_{t=1}^{h} |\widehat{y}_{it} - y^*_{it}|$. The *mean squared error (MSE)* is defined as $\ell^{\mathrm{MSE}}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}^*) \doteq \frac{1}{mh} \sum_{i \in [m]} \sum_{t=1}^{h} (\widehat{y}_{it} - y^*_{it})^2$. The *mean absolute scale error (MASE)* is defined as $\ell^{\mathrm{MASE}}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}^*) \doteq \frac{1}{mh} \sum_{i \in [m]} \sum_{t=1}^{h} \frac{|\widehat{y}_{it} - y^*_{it}|}{\frac{1}{h-1} \sum_{t=1}^{l-1} |y_{it+1} - y_{it}|}$.[5] The *mean absolute percentage error (MAPE)* is defined as $\ell^{\mathrm{MAPE}}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}^*) \doteq \frac{100}{mh} \sum_{i \in [m]} \sum_{t=1}^{h} \frac{|\widehat{y}_{it} - y^*_{it}|}{|y^*_{it}|}$.

[5] We note MAE is scale-dependent but less sensitive to outliers, MSE disproportionately penalizes large forecast errors and is therefore more outlier-sensitive, while MASE normalizes errors w.r.t. the forecasts of naive forecast method (i.e., setting the next time-step's forecast to be the current time-step realized value), making it scale-free and comparable across datasets or domains.

B Result Aggregation Procedure

After evaluating multiple forecasting models across a diverse set of benchmark tasks, we require aggregation methods to summarize and compare model performance at the aggregate level. This section describes two complementary aggregation procedures: *average win rate* and *skill score*.

B.1 Problem Setup

Let p denote the number of models under evaluation and q denote the number of benchmark tasks. For each model $i \in [p]$ and each benchmark task $j \in [q]$, we compute an error metric $\ell^{i,j}$ (e.g., MAE, RMSE, MASE, CRPS). The error values are organized into a matrix $E \in \mathbb{R}^{p \times q}$, where $E[i, j] = \ell^{i,j}$ represents the error of model i on task j.

In practice, some models may not produce valid results on certain tasks (e.g., due to computational failures or data incompatibilities), resulting in missing values. Our aggregation procedures handle these missing values gracefully by excluding unavailable comparisons.

B.2 Average Win Rate

The *average win rate* W_i for model i quantifies the probability that model i achieves lower error than another randomly chosen model $i' \neq i$ on a randomly chosen benchmark task. This metric provides a pairwise comparison perspective that is robust to the absolute scale of errors across different tasks.

Formally, for model i, the average win rate is computed as:

$$W_i = \frac{1}{|\mathcal{C}_i|} \sum_{j \in [q]} \sum_{\substack{i' \in [p] \\ i' \neq i}} w_{i,i',j}, \tag{1}$$

where $|\mathcal{C}_i|$ is the total number of valid comparisons involving model i, and the win indicator $w_{i,i',j}$ is defined as:

$$w_{i,i',j} = \begin{cases} 1 & \text{if } \ell^{i,j} < \ell^{i',j} \text{ and both values are valid,} \\ 0.5 & \text{if } \ell^{i,j} = \ell^{i',j} \text{ and both values are valid,} \\ 0 & \text{if } \ell^{i,j} > \ell^{i',j} \text{ and both values are valid,} \\ 0 & \text{if either value is missing.} \end{cases} \tag{2}$$

The normalization factor $|\mathcal{C}_i|$ accounts for the actual number of valid comparisons:

$$|\mathcal{C}_i| = \sum_{j \in [q]} \sum_{\substack{i' \in [p] \\ i' \neq i}} \mathbb{1}\{\ell^{i,j} \text{ and } \ell^{i',j} \text{ are both valid}\}, \tag{3}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

B.3 Skill Score

The *skill score* S_i for model i quantifies how much the model reduces forecasting error compared to a fixed baseline model β. Unlike win rate, which compares models in a pairwise manner, skill score provides an absolute measure of improvement relative to a reference model (typically a simple baseline such as seasonal naive forecasting).

For model i relative to baseline β, the skill score is computed as:

$$S_i = 1 - \left(\prod_{j \in \mathcal{R}_i} \text{clip}\left(\frac{\ell^{i,j}}{\ell^{\beta,j}}; \ell, u \right) \right)^{1/|\mathcal{R}_i|}, \tag{4}$$

where $\mathcal{R}_i = \{j \in [q] : \ell^{i,j} \text{ and } \ell^{\beta,j} \text{ are both valid}\}$ is the set of tasks where both model i and baseline β have valid results, and $\text{clip}(x; \ell, u) = \max(\ell, \min(x, u))$ clips the relative error ratio to the interval $[\ell, u]$ with $\ell = 10^{-2}$ and $u = 100$.

The clipping operation prevents extreme relative errors (e.g., division by near-zero baseline errors) from dominating the geometric mean. When $\ell^{\beta,j} = 0$, we handle this edge case as follows:

$$\frac{\ell^{i,j}}{\ell^{\beta,j}} = \begin{cases} 1 & \text{if } \ell^{i,j} = 0, \\ u & \text{if } \ell^{i,j} > 0. \end{cases} \tag{5}$$

The skill score interpretation is straightforward:

- $S_i > 0$: Model i performs better than the baseline (lower relative error).
- $S_i = 0$: Model i performs equivalently to the baseline.
- $S_i < 0$: Model i performs worse than the baseline.

B.4 Geometric Mean Rationale

The skill score uses a geometric mean (via the product raised to the reciprocal power) rather than an arithmetic mean for aggregating relative errors across tasks. This choice has several advantages:

- *Scale invariance*: The geometric mean is invariant to multiplicative scaling, ensuring that tasks with different error magnitudes contribute proportionally rather than being dominated by high-error tasks.
- *Symmetry*: The geometric mean treats improvements and degradations symmetrically (e.g., a $2\times$ improvement and a $2\times$ degradation cancel out in the geometric mean).
- *Robustness*: The geometric mean is less sensitive to outliers than the arithmetic mean, which is important when aggregating across diverse benchmark tasks.

B.5 Implementation Details

Both aggregation procedures are implemented in `https://github.com/Smlcrm/TempusBench`, which handles missing values gracefully by excluding unavailable comparisons from the computation. The aggregators accept a pivot table (DataFrame) where rows represent models $i \in [p]$, columns represent benchmark tasks $j \in [q]$, and values represent error metrics $\ell^{i,j}$. Missing values are automatically detected and excluded from the aggregation, ensuring that models are only compared on tasks where both models have valid results.

The implementation provides two aggregator classes: `WinRate` and `SkillScore`, both inheriting from `BaseAggregator`. Each aggregator can be instantiated with a pivot table and, in the case of `SkillScore`, a baseline model β (default: $\beta =$ seasonal_naive).

C Additional results.

C.1 Win Rate Results

Win rates are computed for all evaluated models across different metrics. Higher win rates indicate models that consistently outperform competitors.

C.1.1 Point Forecast Metrics

Point forecast metrics evaluate the accuracy of single-value predictions.

Mean Absolute Percentage Error (MAPE) Table 4a shows the average win rate for models evaluated on the MAPE metric.

(a) Average Win Rate for MAPE Metric. See Appendix B for additional details on computation.

Model Name	Average Win Rate
Lafn	0.7931
Timesfm	0.6730
Croston Classic	0.6164
Seasonal Naive	0.5849
Toto	0.5789
Varmax	0.5714
Arima	0.5346
Moment	0.5220
Lagllama	0.5031
Lstm	0.4969
Moirai	0.4966
Svr	0.4874
Tabpfn	0.4828
Random Forest	0.4748
Tiny Time Mixer	0.4151
Chronos	0.4025
Exponential Smoothing	0.3333
Prophet	0.3333
Theta	0.3300
Moirai Moe	0.0000

(b) Average Win Rate for MAE Metric

Model Name	Average Win Rate
Timesfm	0.9057
Toto	0.7368
Tiny Time Mixer	0.6604
Lafn	0.6414
Croston Classic	0.6164
Svr	0.5818
Tabpfn	0.5448
Lstm	0.5409
Moirai	0.5379
Chronos	0.5346
Random Forest	0.5000
Arima	0.4717
Seasonal Naive	0.5220
Lagllama	0.2704
Varmax	0.2571
Prophet	0.3836
Moment	0.3019
Theta	0.2600
Exponential Smoothing	0.2956
Moirai Moe	0.0000

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) Tables 4b and 5a show win rates for MAE and RMSE metrics respectively.

Mean Absolute Scaled Error (MASE) Table 5b shows win rates for the MASE metric. Note that MASE and MAE have identical win rates in this benchmark, indicating similar relative performance rankings.

C.1.2 Probabilistic Forecast Metrics

Probabilistic forecast metrics evaluate the quality of prediction intervals and distributions.

Continuous Ranked Probability Score (CRPS) Table 6a shows win rates for the CRPS metric, which evaluates probabilistic forecasts.

Weighted Interval Score and Quantile Score Tables 6b and 7a show win rates for weighted interval score and quantile score metrics.

(a) Average Win Rate for RMSE Metric

Model Name	Average Win Rate
Timesfm	0.8742
Tiny Time Mixer	0.6730
Toto	0.6316
Croston Classic	0.6855
Random Forest	0.5943
Arima	0.6038
Svr	0.5755
Prophet	0.5535
Lafn	0.5517
Lstm	0.5472
Tabpfn	0.4966
Chronos	0.4403
Varmax	0.4143
Moirai	0.4345
Seasonal Naive	0.3711
Exponential Smoothing	0.3648
Moment	0.3585
Lagllama	0.1950
Theta	0.1200
Moirai Moe	0.0000

(b) Average Win Rate for MASE Metric

Model Name	Average Win Rate
Timesfm	0.9057
Toto	0.7368
Tiny Time Mixer	0.6604
Lafn	0.6414
Croston Classic	0.6164
Svr	0.5818
Tabpfn	0.5448
Lstm	0.5409
Moirai	0.5379
Chronos	0.5346
Random Forest	0.5000
Arima	0.4717
Seasonal Naive	0.5220
Lagllama	0.2704
Varmax	0.2571
Prophet	0.3836
Moment	0.3019
Theta	0.2600
Exponential Smoothing	0.2956
Moirai Moe	0.0000

(a) Average Win Rate for CRPS Metric

Model Name	Average Win Rate
Toto	1.0000
Moirai	0.7857
Lafn	0.5714
Chronos	0.4000
Lagllama	0.2667
Moirai Moe	0.0000

(b) Average Win Rate for Weighted Interval Score

Model Name	Average Win Rate
Toto	1.0000
Moirai	0.8571
Lafn	0.4643
Chronos	0.5000
Lagllama	0.2000
Moirai Moe	0.0000

(a) Average Win Rate for Quantile Score

Model Name	Average Win Rate
Toto	1.0000
Moirai	0.8571
Lafn	0.5000
Chronos	0.4000
Lagllama	0.2667
Moirai Moe	0.0000

C.2 Skill Score Results

Skill scores compare model performance to a baseline model (Seasonal Naive). Positive skill scores indicate better performance than the baseline, while negative scores indicate worse performance.

Mean Absolute Percentage Error (MAPE) Skill Scores Table 8 shows skill scores for the MAPE metric relative to the Seasonal Naive baseline.

Table 8: Skill Score for MAPE Metric (Baseline: Seasonal Naive)

Model Name	Skill Score
Varmax	0.3264
Timesfm	0.2237
Croston Classic	0.1145
Seasonal Naive	0.0000
Prophet	-0.0240
Tabpfn	-0.1035
Arima	-0.0895
Lafn	-0.0933
Chronos	-0.1272
Tiny Time Mixer	-0.2292
Exponential Smoothing	-0.3066
Random Forest	-0.3410
Toto	-0.3500
Moment	-0.3883
Svr	-0.7162
Lstm	-0.7212
Lagllama	-0.9898
Theta	-2.4162
Moirai Moe	-1.6942
Moirai	-1.7595

Mean Absolute Error (MAE) Skill Scores Table 9 shows skill scores for the MAE metric.

Table 9: Skill Score for MAE Metric (Baseline: Seasonal Naive)

Model Name	Skill Score
Timesfm	0.5442
Toto	0.5081
Chronos	0.2852
Tiny Time Mixer	0.2510
Prophet	0.1179
Croston Classic	0.1016
Arima	0.0153
Lafn	0.0092
Seasonal Naive	0.0000
Tabpfn	-0.0741
Varmax	-0.0765
Random Forest	-0.0866
Lstm	-0.0516
Exponential Smoothing	-0.1032
Svr	-0.1683
Moment	-0.3546
Moirai Moe	-0.3957
Moirai	-0.4945
Lagllama	-1.9980
Theta	-3.4353

Root Mean Squared Error (RMSE) Skill Scores Table 10 shows skill scores for the RMSE metric.

Table 10: Skill Score for RMSE Metric (Baseline: Seasonal Naive)

Model Name	Skill Score
Timesfm	0.5739
Toto	0.4378
Tiny Time Mixer	0.3033
Prophet	0.2328
Chronos	0.1801
Croston Classic	0.1177
Arima	0.0944
Lafn	0.0792
Random Forest	0.0019
Seasonal Naive	0.0000
Lstm	-0.0049
Exponential Smoothing	-0.0063
Svr	-0.0094
Tabpfn	-0.0202
Varmax	-0.0446
Moirai Moe	-0.1578
Moment	-0.2504
Moirai	-0.9759
Lagllama	-1.4620
Theta	-2.5424

Mean Absolute Scaled Error (MASE) Skill Scores Table 11 shows skill scores for the MASE metric. Note that MASE and MAE have identical skill scores in this benchmark.

Table 11: Skill Score for MASE Metric (Baseline: Seasonal Naive)

Model Name	Skill Score
Timesfm	0.5442
Toto	0.5081
Chronos	0.2852
Tiny Time Mixer	0.2510
Prophet	0.1179
Croston Classic	0.1016
Arima	0.0153
Lafn	0.0092
Seasonal Naive	0.0000
Tabpfn	-0.0741
Varmax	-0.0765
Random Forest	-0.0866
Lstm	-0.0516
Exponential Smoothing	-0.1032
Svr	-0.1683
Moment	-0.3546
Moirai Moe	-0.3957
Moirai	-0.4945
Lagllama	-1.9980
Theta	-3.4353

C.3 Key Findings

C.3.1 Top Performing Models

- **MAPE Metric**: Lafn achieves the highest win rate (0.7931), followed by Timesfm (0.6730) and Croston Classic (0.6164).
- **MAE/RMSE/MASE Metrics**: Timesfm consistently achieves the highest win rates across MAE (0.9057), RMSE (0.8742), and MASE (0.9057), with Toto and Tiny Time Mixer also performing strongly.
- **Probabilistic Metrics**: Toto achieves perfect win rates (1.0000) for both CRPS and Quantile Score, while also achieving perfect win rate for Weighted Interval Score. Moirai shows strong performance on probabilistic metrics (0.7857 for CRPS, 0.8571 for WIS and Quantile Score).

C.3.2 Skill Score Insights

- **Positive Skill Scores**:
  - **MAPE**: Varmax (0.3264), Timesfm (0.2237), and Croston Classic (0.1145) show positive skill scores.
  - **MAE/RMSE/MASE**: Timesfm, Toto, Tiny Time Mixer, and Chronos consistently show positive skill scores across these metrics, indicating they outperform the Seasonal Naive baseline.
- **Negative Skill Scores**: Some models show negative skill scores, particularly Lagllama, Theta, Moirai, and Moirai Moe, which perform worse than the baseline across most metrics.
- **Baseline Performance**: The Seasonal Naive model serves as the baseline (skill score = 0.0) and provides competitive performance across many tasks. On MAPE, most models actually perform worse than the baseline, with only Varmax, Timesfm, and Croston Classic showing positive skill.

C.3.3 Model-Specific Observations

- **Toto**: Exceptional performance on probabilistic metrics with perfect win rates, while maintaining decent performance on point forecast metrics.
- **Moirai**: Strong probabilistic forecasting capabilities but weaker performance on point forecast metrics (MAPE win rate: 0.4966).
- **Moirai Moe**: Consistently shows zero win rate across all metrics, indicating it does not outperform other models in any evaluated scenario.
- **Lafn**: Best performer on MAPE metric but shows variable performance across other metrics.

D Forecasters

Table 12: Summary of forecasters included in TempusBench.

Category	Included Models	Core Characteristics
Foundation Models	*Moirai, Moirai-MoE, TimesFM, TimesFM-2.0, Chronos, Lag-Llama, Toto, MOMENT, TTM, TabPFN-TS*	**Paradigm:** Universal, zero-shot/few-shot forecasting. A single large model is pre-trained on massive, diverse datasets and generalizes to new tasks without retraining. **Architecture:** Primarily based on Transformers or other deep learning structures like MLP-Mixers. They process raw time series via patching or novel tokenization schemes. **I/O:** Often produce probabilistic forecasts and can natively handle univariate, multivariate, and covariate data.
Classic Machine Learning	*LSTM, Random Forest, XGBoost, SVR*	**Paradigm:** Supervised learning models trained per-dataset. They excel at capturing complex, non-linear relationships but require specific training for each task. **Architecture:** Diverse, including Recurrent Neural Networks (for sequence memory), Tree Ensembles (for interaction effects), and Kernel Methods. **I/O:** Typically require explicit feature engineering (e.g., lags, calendar variables) to create a tabular format. Most often produce point forecasts.
Statistical & Decomposable	*ARIMA, Holt-Winters, Prophet, Theta Method, Croston's Method, Seasonal Naive*	**Paradigm:** Assume the time series is generated by an underlying statistical process or can be decomposed into simpler, interpretable components like trend and seasonality. **Architecture:** An explicit mathematical formula is fitted directly to an individual time series. **I/O:** Highly interpretable point forecasts. Often specialized for particular data patterns (e.g., intermittency with *Croston's*).

In this section, we summarize the forecasting models which have been included in TempusBench. We summarize all models in Table 12, and provide and comparison of TSFMs, machine learning forecasting models, and statistical forecasting models in Table 13.

Table 13: Comparative Overview of Forecasting Models

Feature	Group 1: Foundation Models	Group 2: Machine Learning Models	Group 3: Classical & Statistical Models
Models Included	• *Moirai / Moirai-MoE* • *TimesFM / TimesFM-2.0* • *Chronos* • *Lag-Llama* • *Toto* • *MOMENT* • *TabPFN-TS* • *Tiny Time Mixers (TTM)*	• *LSTM* • *Random Forest* • *XGBoost* • *SVR*	• *ARIMA* • *Holt-Winters* • *Prophet* • *Theta Method* • *Croston's Method* • *Seasonal Naive*
Core Paradigm	Large, pre-trained models designed for universal, zero-shot forecasting. They learn general time-series patterns from massive, diverse datasets.	Models trained for a specific forecasting task, often relying on feature engineering. They leverage distinct architectures (e.g., recurrence, ensembles) rather than massive pre-training.	Model-based approaches assuming an underlying stochastic process or decomposable structure. Parameters are estimated directly from the target time series.
Architecture	Primarily Transformer-based (Encoder, Decoder, or both). Innovations include MoE layers, residual forecasting, and specialized attention mechanisms.	Diverse architectures: MLP-Mixer (*TTM*), RNN (*LSTM*), Tabular-Transformer (*TabPFN-TS*), Tree Ensembles (*RF, XGBoost*), and Kernel-based (*SVR*).	Mathematical formulations: State-space models (*ARIMA, Holt-Winters*), decomposable additive models (*Prophet, Theta*), and simple heuristics (*Croston's, S. Naive*).
Input Handling	Process raw time series, typically via patching (*Moirai, TimesFM*), lag-based tokenization (*Lag-Llama*), or value quantization (*Chronos*). Can natively handle uni/multivariate series.	Generally require explicit feature engineering (e.g., lags, calendar variables) to create a tabular dataset (*RF, XGBoost, SVR*). *LSTM* and *TTM* process raw sequences.	Operate directly on the univariate time series. May require stationarity (*ARIMA*) or be specialized for patterns like seasonality (*Holt-Winters*) or intermittency (*Croston's*).
Output Type	Mostly probabilistic, predicting the parameters of a flexible distribution. Point forecasts are derived from the distribution (e.g., median).	Primarily point forecasts. *TabPFN-TS* is a notable exception, providing a probabilistic output by approximating the posterior.	Primarily point forecasts. *Prophet* is an exception, generating uncertainty intervals. Probabilistic versions exist but are not standard.
Key Trait	Powerful zero-shot/few-shot performance. High model capacity and ability to generalize across domains without dataset-specific training.	Model-specific strengths: computational efficiency (*TTM*), modeling long-term dependencies (*LSTM*), and capturing complex non-linear interactions (*RF, XGBoost*).	High interpretability and strong statistical foundations. Often specialized and highly efficient for specific data patterns (e.g., trend, seasonality, intermittency).

D.0.1 Moirai

Moirai is a universal time series forecasting model developed by Salesforce AI Research, built upon a masked encoder-only Transformer architecture. It is designed as a single, large pre-trained model capable of handling diverse forecasting tasks without dataset-specific retraining. The model is pre-trained on LOTSA, a large-scale archive of over 27 billion observations, enabling it to perform powerful zero-shot forecasting. [18]

- **Input:** Accepts univariate or multivariate time series with an arbitrary number of variates and covariates.
- **Output:** Produces a probabilistic forecast by predicting the parameters of a flexible mixture distribution (composed of Student's t, Negative Binomial, Log-Normal, and low-variance Normal distributions).
- **Architecture:** Employs a masked encoder-only Transformer. Its key innovations include:
  - **Multi Patch Size Projection:** Uses different patch sizes to effectively process time series of varying frequencies.
  - **Any-variate Attention:** Flattens multivariate series into a single sequence and uses binary attention biases to manage an arbitrary number of variates while maintaining permutation equivariance.
- **Forecasting Type:** A universal, zero-shot, probabilistic forecaster. It can generate point forecasts by taking the median of the predicted distribution.

D.0.2 Moirai-MoE

Moirai-MoE is an advanced version of the Moirai foundation model that integrates a Sparse Mixture of Experts (MoE) architecture. Instead of relying on heuristic-based, frequency-specific projection layers, Moirai-MoE delegates the task of modeling diverse time series patterns to specialized "expert" networks within its Transformer layers. This allows for automatic, token-level specialization in a data-driven manner, leading to improved accuracy and greater efficiency in terms of activated parameters. [19]

- **Input:** Accepts univariate or multivariate time series with an arbitrary number of variates and covariates.
- **Output:** Produces a probabilistic forecast by predicting the parameters of a flexible mixture distribution for the next token in an autoregressive manner.
- **Architecture:** Employs a decoder-only Transformer that replaces the standard Feed-Forward Network (FFN) layers with MoE layers. Key architectural changes from the original Moirai include:
  - **Mixture of Experts (MoE):** A gating function routes each time series token to a small subset of specialized expert networks, allowing the model to handle diverse patterns at a granular level.
  - **Single Projection Layer:** It uses a single input/output projection layer for all time series, removing the dependency on frequency-based heuristics.
- **Forecasting Type:** A universal, zero-shot, probabilistic forecaster that is more accurate and efficient (in terms of activated parameters) than the original Moirai model. It can generate point forecasts by taking the median of the predicted distribution.

D.0.3 TimesFM

TimesFM is a time-series foundation model developed by Google Research, designed for zero-shot forecasting. It is based on a decoder-only Transformer architecture and is pretrained on a very large corpus of time series data, combining both real-world and synthetic sources. The model's key objective is to provide accurate out-of-the-box point forecasts on unseen datasets without requiring any dataset-specific training. [20]

- **Input:** Accepts a univariate time series context window.
- **Output:** Produces a point forecast for a given prediction horizon.

- **Architecture:** Employs a decoder-only Transformer architecture that processes the time series in patches. Key architectural features include:
  - **Decoder-Only Transformer:** Utilizes a standard decoder-style attention mechanism to autoregressively predict future values patch by patch.
  - **Input Patching:** The input time series is segmented into non-overlapping patches, which are then embedded using a residual block of MLPs before being fed to the Transformer.
- **Forecasting Type:** A universal, zero-shot, point forecaster designed primarily for long-horizon forecasting tasks.

D.0.4 TimesFM-2.0

TimesFM-2.0 is an improved version of the original foundation model from Google Research. While retaining the same decoder-only Transformer architecture, its key innovation lies in forecasting the residual component of a time series after performing a seasonal-trend decomposition. This approach makes the model significantly more accurate, particularly for time series that exhibit clear trends. [20]

- **Input:** Accepts a univariate time series context window.
- **Output:** Produces a point forecast for a given prediction horizon.
- **Architecture:** Based on the original decoder-only Transformer with input patching. The primary architectural update is its **residual forecasting** methodology:
  - **Seasonal-Trend Decomposition:** The model first decomposes the input series to separate its trend and seasonal components.
  - **Residual Forecasting:** The core Transformer then forecasts the residual (the signal remaining after decomposition). This forecast is added back to the projected trend to produce the final prediction.
- **Forecasting Type:** A universal, zero-shot, point forecaster with enhanced performance on trended time series compared to its predecessor.

D.0.5 Chronos

Chronos is a family of pretrained time series models developed by Amazon Science that frames forecasting as a language modeling task. The core idea is to "tokenize" time series values by scaling and quantizing them into a fixed vocabulary. By doing so, standard Transformer-based language model architectures can be trained on sequences of these tokens using a cross-entropy loss, effectively learning the "language" of time series. [21]

- **Input:** Accepts a univariate time series context window.
- **Output:** Produces a probabilistic forecast by generating multiple sample future trajectories. A point forecast can be derived from the median of these samples.
- **Architecture:** Based on standard language model architectures (specifically the T5 encoder-decoder family). Its defining characteristic is its unique data preprocessing pipeline:
  - **Tokenization via Quantization:** The model first applies mean scaling to the input time series. It then quantizes these scaled values into a finite set of discrete tokens, converting the continuous series into a sequence of categorical variables.
  - **Language Model Training:** The model is trained to predict the next token in a sequence using a standard cross-entropy loss, analogous to how a language model predicts the next word.
- **Forecasting Type:** A universal, zero-shot, probabilistic forecaster.

D.0.6 TabPFN

TabFPN is a forecasting framework that adapts feature pyramid networks (FPN), originally developed for computer vision tasks, to tabular time-series data. The approach builds hierarchical feature representations across multiple temporal resolutions, enabling the model to capture both short- and long-range dependencies. Unlike traditional time-series architectures, TabFPN treats forecasting as a

structured feature-learning problem on tabularized sequences, combining multiscale decomposition with probabilistic prediction.

- **Input:** A univariate or multivariate time series, converted into tabular form with hierarchical features at multiple temporal resolutions.
- **Output:** Produces probabilistic forecasts by estimating distributions over future values at each horizon; point forecasts can be obtained from the distribution mean or median.
- **Architecture:**
  - *Feature Pyramids:* The series is decomposed into multiple temporal scales (e.g., short-term, medium-term, seasonal) using windowed transformations. Each scale yields a feature representation.
  - *FPN Backbone:* These features are passed into a feature pyramid network adapted for tabular regression, allowing cross-scale information flow and refinement.
  - *Prediction Head:* Aggregates multiscale features to generate forecasts, with uncertainty quantification via distributional outputs.
- **Forecasting Type:** A universal, zero-shot, probabilistic forecaster with explicit multiscale feature integration.

D.0.7 TabPFN-TS

TabPFN-TS is a novel approach that adapts TabPFN-v2, a general-purpose tabular foundation model, for time series forecasting. The core methodology involves recasting the forecasting problem as a tabular regression task. This is achieved through lightweight feature engineering on the time index, without relying on lagged values. Notably, the underlying TabPFN-v2 model was pretrained exclusively on synthetic tabular data and has not seen any time series data. [22]

- **Input:** A univariate time series, which is converted into a feature matrix based on timestamps.
- **Output:** Produces a probabilistic forecast by approximating the posterior predictive distribution for each future time step. Point forecasts can be derived from the mean or median of this distribution.
- **Architecture:** It does not use a time-series-specific architecture. Instead, it relies on:
  - **Feature Engineering:** The time series is transformed into a tabular dataset by creating features from timestamps. These include standard calendar features (e.g., hour of day, day of week), automatically detected seasonal features via a Fourier transform, and a simple running index.
  - **TabPFN-v2 Model:** The generated tabular data is fed into the pretrained TabPFN-v2 model, which performs the regression task to predict future values.
- **Forecasting Type:** A universal, zero-shot, probabilistic forecaster.

D.0.8 Tiny Time Mixers (TTM)

Tiny Time Mixers (TTM) is a family of lightweight pre-trained models from IBM Research, based on the efficient TSMixer architecture. In contrast to large, LLM-based approaches, TTMs are designed to be extremely small (<1M parameters) and fast, while still providing strong zero-shot and few-shot forecasting performance. The models are pre-trained exclusively on a large corpus of public time series datasets, making them a highly efficient alternative for universal forecasting. [23]

- **Input:** Accepts univariate or multivariate time series, with optional support for exogenous variables during the fine-tuning stage.
- **Output:** Produces a point forecast for a given prediction horizon.
- **Architecture:** Based on the MLP-Mixer architecture. The model is pre-trained in a channel-independent manner and uses a multi-level structure to handle diverse data and tasks.
  - **TSMixer Backbone:** The core of the model uses simple MLP blocks for temporal and feature mixing, avoiding the computational overhead of Transformer-based attention.
  - **Multi-Resolution Pre-training:** Employs several novel techniques to handle heterogeneous datasets, including **adaptive patching** (using different patch configurations at different layers) and **data augmentation via downsampling**.

  - **Multi-level Modeling:** Uses a frozen pre-trained backbone and a smaller, fine-tunable decoder, which can incorporate channel-mixing and an **exogenous mixer** to fuse external signals for target-specific tasks.
- **Forecasting Type:** A universal, zero-shot/few-shot, point forecaster, notable for its small size and computational efficiency.

D.0.9 Lag-Llama

Lag-Llama is a foundation model for univariate probabilistic time series forecasting. It is built upon a decoder-only Transformer architecture, similar to LLaMA, and is pretrained on a large, diverse corpus of open-source time series data. The model's key innovation is its tokenization strategy, which uses lagged values of the time series as input features, allowing it to generalize across different frequencies and domains. [24]

- **Input:** Accepts a univariate time series context window.
- **Output:** Produces a probabilistic forecast by outputting the parameters of a Student's t-distribution for the next time step. Future trajectories are generated autoregressively.
- **Architecture:** Based on a decoder-only Transformer (LLaMA). Its defining characteristic is its input representation:
  - **Tokenization via Lag Features:** Instead of patching, the input token for each time step is a vector composed of lagged values from the time series history (e.g., values from 1, 7, and 14 days prior). This is augmented with standard date-time features.
  - **Value Scaling:** Applies robust scaling (using median and IQR) to normalize the input values and includes the scaling parameters as additional features.
- **Forecasting Type:** A universal, zero-shot/few-shot, probabilistic forecaster.

D.0.10 Toto

Toto (Time Series Optimized Transformer for Observability) is a foundation model from Datadog, specifically designed for multivariate time series forecasting with a focus on observability metrics. It is built on a decoder-only Transformer architecture and incorporates several novel components to handle the unique challenges of observability data, such as high non-stationarity and heavy-tailed distributions. The model is pretrained on a large and diverse corpus that includes real-world observability data, public datasets, and synthetic data. [25]

- **Input:** Accepts multivariate time series.
- **Output:** Produces a probabilistic forecast by predicting the parameters of a Student-T mixture model.
- **Architecture:** A decoder-only Transformer with several key innovations tailored for observability data:
  - **Patch-based Causal Normalization:** A novel per-patch scaling method that computes normalization statistics from current and past data to handle highly nonstationary series.
  - **Proportional Factorized Attention:** An efficient attention mechanism that uses a mix of time-wise and variate-wise attention blocks to judiciously model interactions in high-dimensional multivariate data.
  - **Student-T Mixture Model Head:** An output layer that models the predictive distribution as a mixture of Student-T distributions to better capture the complex, heavy-tailed nature of observability metrics.
  - **Composite Robust Loss:** A hybrid loss function combining negative log-likelihood with a robust point-wise loss to stabilize training in the presence of outliers.
- **Forecasting Type:** A universal, zero-shot, probabilistic forecaster for multivariate time series.

D.0.11 MOMENT

MOMENT (Multi-task, Open-source, Foundation Model for Time-series) is a family of open-source foundation models from Carnegie Mellon University designed for general-purpose time series analysis.

The models are built on a Transformer encoder architecture and are pretrained on a large, diverse collection of public time series called the "Time Series Pile." A key characteristic of MOMENT is its versatility; it is designed to serve as a building block for a wide range of downstream tasks, including forecasting, classification, anomaly detection, and imputation, often with minimal task-specific fine-tuning. [26]

- **Input:** Accepts a univariate time series of a fixed length. Multivariate time series are handled by treating each channel independently.
- **Output:** Produces a reconstructed version of the input time series. This output can be adapted for various downstream tasks, such as generating forecasts by masking future values or extracting embeddings for classification.
- **Architecture:** A standard Transformer encoder that processes time series data in patches.
  - **Masked Pre-training:** The model is pretrained using a masked time series prediction task. It learns to reconstruct randomly masked patches of the input time series, enabling it to learn robust representations.
  - **Patching:** The input time series is segmented into non-overlapping patches, which are then linearly projected into embeddings for the Transformer.
  - **Lightweight Prediction Head:** A simple linear layer is used to reconstruct the time series from the Transformer's output embeddings. This head can be easily replaced or adapted for different downstream tasks.
- **Forecasting Type:** A universal foundation model for general time series analysis. It can be used for zero-shot or few-shot forecasting (point-based), classification, anomaly detection, and imputation.

D.0.12 ARIMA

The Autoregressive Integrated Moving Average (ARIMA) model is a class of statistical models for analyzing and forecasting time series data. It is a generalization of the simpler Autoregressive Moving Average (ARMA) model that can be applied to non-stationary time series. The model's name reflects its three core components: Autoregression (AR), Integrated (I), and Moving Average (MA). These components capture the key temporal structures within the data, such as dependencies on past observations and past forecast errors. [27]

- **Input:** A univariate time series.
- **Output:** A point forecast for future time steps. While classical ARIMA produces point forecasts, probabilistic forecasts can be generated by assuming a distribution for the error term.
- **Mathematical Formulation:** An ARIMA(p, d, q) model is defined by three parameters: the order of the autoregressive component (p), the degree of differencing (d), and the order of the moving average component (q). The model assumes that the differenced time series, $\widetilde{\boldsymbol{y}}_t = (1-B)^d \boldsymbol{y}_t$, is stationary, where B is the backshift operator. The formulation for the stationary series $\widetilde{\boldsymbol{y}}_t$ is:

$$\widetilde{\boldsymbol{y}}_t = c + \sum_{i=1}^{p} \phi_i \widetilde{\boldsymbol{y}}_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{6}$$

  where:
  - p is the autoregressive order, representing the number of lagged observations included in the model.
  - d is the degree of differencing, representing the number of times the raw observations are differenced to achieve stationarity.
  - q is the moving average order, representing the size of the moving average window applied to past forecast errors.
  - $\boldsymbol{\phi}$ is the vector of autoregressive coefficients.
  - $\boldsymbol{\theta}$ is the vector of moving average coefficients.
  - c is a constant term.
  - ϵ_t is the white noise error term at time t, typically assumed to be drawn from a Gaussian distribution with zero mean.
- **Forecasting Type:** A statistical model that provides point forecasts. It is often used as a baseline in forecasting tasks. Seasonal variations can be included by using a Seasonal ARIMA (SARIMA) model.

D.0.13 Croston's Method

Croston's method is a forecasting technique specifically designed for intermittent demand time series, which are characterized by sporadic, non-zero values interspersed with periods of zero demand. The method decomposes the original time series into two separate components: the magnitude of the non-zero demand and the time interval between consecutive demands. By forecasting these two components separately using Simple Exponential Smoothing and then combining them, the model provides a more accurate estimate of the mean demand per period compared to standard smoothing methods, which can be biased when applied to intermittent data. [28]

- **Input:** A univariate time series with intermittent demand.
- **Output:** A point forecast for the average demand per period.
- **Mathematical Formulation:** The method maintains and updates two estimates: one for the non-zero demand size ($\widehat{\boldsymbol{z}}$) and one for the interval between demands ($\widehat{\boldsymbol{p}}$). Let $\boldsymbol{y}_t$ be the demand at time t, and let q be the time elapsed since the last demand. The updates occur only when a non-zero demand is observed ($\boldsymbol{y}_t > 0$):

$$\widehat{\boldsymbol{z}}_t = \widehat{\boldsymbol{z}}_{t-1} + \alpha(\boldsymbol{y}_t - \widehat{\boldsymbol{z}}_{t-1}) \tag{7}$$

$$\widehat{\boldsymbol{p}}_t = \widehat{\boldsymbol{p}}_{t-1} + \alpha(q - \widehat{\boldsymbol{p}}_{t-1}) \tag{8}$$

  If demand at time t is zero, the estimates are not updated ($\widehat{\boldsymbol{z}}_t = \widehat{\boldsymbol{z}}_{t-1}$, $\widehat{\boldsymbol{p}}_t = \widehat{\boldsymbol{p}}_{t-1}$) and the interval counter q is incremented. After a demand occurs, q is reset to 1. The final forecast for the mean demand per period, $\widehat{\boldsymbol{y}}_t$, is the ratio of the two smoothed components:

$$\widehat{\boldsymbol{y}}_t = \frac{\widehat{\boldsymbol{z}}_t}{\widehat{\boldsymbol{p}}_t} \tag{9}$$

  where α is the smoothing parameter.
- **Forecasting Type:** A statistical model for point forecasting, specialized for intermittent or "lumpy" demand patterns.

D.0.14 Holt-Winters Exponential Smoothing

Holt-Winters is an extension of exponential smoothing that explicitly models trend and seasonality. It is a widely used statistical method for forecasting time series data that exhibit these components. The method operates by applying exponential smoothing to three components: the level, the trend, and the seasonality. There are two primary variations of the model, additive and multiplicative, which differ in how they incorporate the seasonal component. [5]

- **Input:** A univariate time series with trend and seasonality.
- **Output:** A point forecast for future time steps.
- **Mathematical Formulation:** The model provides separate updating equations for the level ($\widehat{l}_t$), trend ($\widehat{b}_t$), and seasonal ($\widehat{s}_t$) components, using smoothing parameters α, β, and γ, respectively. Let L be the length of the seasonal period.

  Additive Method: Used when the seasonal variation is roughly constant throughout the series.

$$\text{Level:} \quad \widehat{l}_t = \alpha(\boldsymbol{y}_t - \widehat{s}_{t-L}) + (1-\alpha)(\widehat{l}_{t-1} + \widehat{b}_{t-1}) \tag{10}$$

$$\text{Trend:} \quad \widehat{b}_t = \beta(\widehat{l}_t - \widehat{l}_{t-1}) + (1-\beta)\widehat{b}_{t-1} \tag{11}$$

$$\text{Seasonality:} \quad \widehat{s}_t = \gamma(\boldsymbol{y}_t - \widehat{l}_t) + (1-\gamma)\widehat{s}_{t-L} \tag{12}$$

  The forecast for h steps ahead is given by:

$$\widehat{\boldsymbol{y}}_{t+h|t} = \widehat{l}_t + h\widehat{b}_t + \widehat{s}_{t-L+h_L^+} \quad \text{where } h_L^+ = \lfloor (h-1) \pmod L \rfloor + 1 \tag{13}$$

  Multiplicative Method: Used when the seasonal variation changes in proportion to the level of the series.

$$\text{Level:} \quad \widehat{l}_t = \alpha\left(\frac{\boldsymbol{y}_t}{\widehat{s}_{t-L}}\right) + (1-\alpha)(\widehat{l}_{t-1} + \widehat{b}_{t-1}) \tag{14}$$

$$\text{Trend:} \quad \widehat{b}_t = \beta(\widehat{l}_t - \widehat{l}_{t-1}) + (1-\beta)\widehat{b}_{t-1} \tag{15}$$

$$\text{Seasonality:} \quad \widehat{s}_t = \gamma\left(\frac{\boldsymbol{y}_t}{\widehat{l}_t}\right) + (1-\gamma)\widehat{s}_{t-L} \tag{16}$$

The forecast for h steps ahead is given by:

$$\widehat{\boldsymbol{y}}_{t+h|t} = (\widehat{l}_t + h\widehat{b}_t)\widehat{s}_{t-L+h_L^+} \quad \text{where } h_L^+ = \lfloor (h-1) \pmod{L} \rfloor + 1 \tag{17}$$

- **Forecasting Type:** A statistical model for point forecasting that can handle various combinations of trend and seasonality.

D.0.15 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) architecture specifically designed to address the vanishing gradient problem, allowing it to learn and remember long-term dependencies in sequential data. Unlike traditional neural networks, LSTMs have internal mechanisms called "gates" that regulate the flow of information. These gates enable the network to selectively remember or forget information over long periods, making it particularly well-suited for time series forecasting. [27]

- **Input:** A sequence of historical time series observations.
- **Output:** A point forecast for one or more future time steps.
- **Mathematical Formulation:** The core of an LSTM unit is its cell state, $\widehat{\boldsymbol{c}}_t$, which acts as a memory. The flow of information into and out of the cell is controlled by three gates: the forget gate ($\boldsymbol{f}_t$), the input gate ($\boldsymbol{i}_t$), and the output gate ($\boldsymbol{o}_t$). At each time step t, these gates update the cell state and produce a hidden state, $\widehat{\boldsymbol{h}}_t$.

$$\text{Forget Gate:} \quad \boldsymbol{f}_t = \boldsymbol{\sigma}(\boldsymbol{w}_f \cdot [\widehat{\boldsymbol{h}}_{t-1}, \boldsymbol{y}_t] + b_f) \tag{18}$$

$$\text{Input Gate:} \quad \boldsymbol{i}_t = \boldsymbol{\sigma}(\boldsymbol{w}_i \cdot [\widehat{\boldsymbol{h}}_{t-1}, \boldsymbol{y}_t] + b_i) \tag{19}$$

$$\text{Candidate State:} \quad \widetilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{w}_c \cdot [\widehat{\boldsymbol{h}}_{t-1}, \boldsymbol{y}_t] + b_c) \tag{20}$$

$$\text{Cell State Update:} \quad \widehat{\boldsymbol{c}}_t = \boldsymbol{f}_t \odot \widehat{\boldsymbol{c}}_{t-1} + \boldsymbol{i}_t \odot \widetilde{\boldsymbol{c}}_t \tag{21}$$

$$\text{Output Gate:} \quad \boldsymbol{o}_t = \boldsymbol{\sigma}(\boldsymbol{w}_o \cdot [\widehat{\boldsymbol{h}}_{t-1}, \boldsymbol{y}_t] + b_o) \tag{22}$$

$$\text{Hidden State Update:} \quad \widehat{\boldsymbol{h}}_t = \boldsymbol{o}_t \odot \tanh(\widehat{\boldsymbol{c}}_t) \tag{23}$$

  where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weight matrices and bias vectors for each gate, $\boldsymbol{\sigma}$ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The final prediction is typically generated by passing the hidden state $\widehat{\boldsymbol{h}}_t$ through a dense output layer.
- **Forecasting Type:** A neural network model for point forecasting that can capture complex non-linear patterns in time series data.

D.0.16 Prophet

Prophet is a forecasting procedure developed by Meta, based on a decomposable time series model. It is designed to be robust to missing data and shifts in the trend, and it typically handles holidays and seasonal effects well. The model fits an additive model with components for trend, seasonality, and holidays. [8]

- **Input:** A univariate time series with timestamps.
- **Output:** A point forecast, along with uncertainty intervals.
- **Mathematical Formulation:** The Prophet model is specified as a sum of three components:

$$\boldsymbol{y}_t = g(t) + s(t) + h(t) + \epsilon_t \tag{24}$$

  where:
  - $g(t)$ is the trend component, which is modeled as either a piecewise linear or logistic growth function. This allows the model to capture non-periodic changes in the time series.
  - $s(t)$ is the seasonality component, which models periodic changes (e.g., yearly, weekly, daily). It is approximated by a Fourier series:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right) \tag{25}$$

    where P is the period of the seasonality (e.g., 365.25 for yearly).

- $h(t)$ is the holiday component, which represents the effects of holidays and special events. It is modeled as a sum of indicator functions for each holiday.
- ϵ_t is the error term, assumed to be normally distributed white noise.

- **Forecasting Type:** A decomposable statistical model for point and probabilistic forecasting, particularly effective for business time series with strong seasonal patterns and holiday effects.

D.0.17 Random Forest

Random Forest is an ensemble machine learning model that operates by constructing a multitude of decision trees at training time. For time series forecasting, it is applied as a regression model to a featurized dataset. By fitting numerous trees on various sub-samples of the data and employing randomness in feature selection, it improves predictive accuracy and controls over-fitting. The final prediction is an average of the outputs from all individual trees, making the model robust and capable of capturing complex, non-linear relationships. [29]

- **Input:** A feature matrix $\boldsymbol{X}$ where rows are observations and columns are engineered features (e.g., lags, calendar variables), and a corresponding target vector $\boldsymbol{y}$.
- **Output:** A point forecast for each input feature vector.
- **Architecture and Formulation:** A Random Forest is an ensemble of B decision trees. Its predictive power comes from two sources of randomness introduced during training:
  - **Bagging (Bootstrap Aggregating):** Each individual tree, f_b, is trained on a bootstrap sample (a random sample drawn with replacement) from the original training dataset.
  - **Feature Randomness:** When splitting a node in a tree, the algorithm considers only a random subset of the total features, which decorrelates the trees in the forest.

  For a new input feature vector $\boldsymbol{x}$, the forecast is the average of the predictions from all B trees in the ensemble:

$$\widehat{\boldsymbol{y}}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} f_b(\boldsymbol{x}) \tag{26}$$

- **Forecasting Type:** An ensemble machine learning model for point forecasting. It is non-parametric and highly effective at modeling non-linear relationships between features and the target variable.

D.0.18 Seasonal Naive

The Seasonal Naive model is a simple yet effective baseline method for forecasting time series with a strong seasonal component. Its core principle is that the forecast for a future period is equal to the last observed value from the same season. For example, the forecast for this Monday would be the value from last Monday. Despite its simplicity, it serves as a crucial benchmark for more complex models. [30]

- **Input:** A univariate time series with a known seasonal period.
- **Output:** A point forecast for future time steps.
- **Mathematical Formulation:** The forecast for h steps ahead from time t, denoted $\widehat{\boldsymbol{y}}_{t+h|t}$, is the last observed value from the corresponding season. Let L be the seasonal period (e.g., $L = 7$ for daily data with weekly seasonality). The forecast is given by:

$$\widehat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{y}_{t+h-L\cdot k} \tag{27}$$

  where $k = \lceil h/L \rceil$ is an integer that ensures the lagged time index refers to the most recent observation from the target season. For a one-season-ahead forecast ($h = L$), this simplifies to $\widehat{\boldsymbol{y}}_{t+L|t} = \boldsymbol{y}_t$.
- **Forecasting Type:** A simple statistical baseline for seasonal point forecasting.

D.0.19 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a supervised learning algorithm that extends the principles of Support Vector Machines (SVMs) to regression problems. Instead of finding a hyperplane that separates classes, SVR aims to find a function that deviates from the target values by a value no greater than a specified margin, ϵ, for as many of the training points as possible. It is particularly effective in high-dimensional spaces and is robust to some outliers due to its use of an ϵ-insensitive loss function, which ignores errors within this margin. [31]

- **Input:** A feature matrix $\boldsymbol{X}$ and a corresponding target vector $\boldsymbol{y}$.
- **Output:** A point forecast for each input feature vector.
- **Mathematical Formulation:** The goal of SVR is to find a function $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ that is as "flat" as possible. This is achieved by minimizing the norm of the weight vector, $||\boldsymbol{w}||^2$. The optimization problem is formulated to tolerate errors up to a margin ϵ while penalizing points that fall outside this margin using slack variables ξ_i and ξ_i^*. The primal optimization problem is:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{28}$$

  subject to the constraints:

$$\boldsymbol{y}_i - (\boldsymbol{w}^T\boldsymbol{x}_i + b) \leq \epsilon + \xi_i \tag{29}$$

$$(\boldsymbol{w}^T\boldsymbol{x}_i + b) - \boldsymbol{y}_i \leq \epsilon + \xi_i^* \tag{30}$$

$$\xi_i, \xi_i^* \geq 0 \tag{31}$$

  where C is a regularization parameter that controls the trade-off between the flatness of the model and the amount up to which deviations larger than ϵ are tolerated. Non-linear relationships are handled by mapping the data to a higher-dimensional space using a kernel function.
- **Forecasting Type:** A machine learning model for point forecasting that is robust to some outliers and effective in high-dimensional feature spaces.

D.0.20 Theta Method

The Theta method is a statistical forecasting technique based on the concept of decomposition. It models a time series by breaking it down into two components, or "theta lines." The first line represents the long-term trend of the data, while the second line is constructed to capture the short-term dynamics by modifying the curvature of the original series. These two lines are forecasted independently and then combined to produce the final forecast. The standard Theta model has been shown to be equivalent to Simple Exponential Smoothing with a drift term. [32]

- **Input:** A univariate time series.
- **Output:** A point forecast for future time steps.
- **Mathematical Formulation:** The method decomposes the original time series, $\boldsymbol{y}_t$, into two theta lines.
  - **Line 1 (Trend Component):** This line is the simple linear trend fitted to the data, which is found by ordinary least squares regression:

$$\widetilde{\boldsymbol{y}}_t^{(1)} = \widehat{a} + \widehat{b}t \tag{32}$$

    This line is extrapolated linearly to produce its forecast.
  - **Line 2 (Short-term Component):** This line is constructed by modifying the original series with a coefficient θ. A common and effective choice is $\theta = 2$, which doubles the local curvatures of the series. This modified series, $\widetilde{\boldsymbol{y}}_t^{(2)}$, is then forecasted using Simple Exponential Smoothing (SES).

  The final forecast, $\widehat{\boldsymbol{y}}_{t+h}$, is a simple average of the forecasts from the two lines:

$$\widehat{\boldsymbol{y}}_{t+h} = \frac{1}{2}\left(\widehat{\boldsymbol{y}}_{t+h}^{(1)} + \widehat{\boldsymbol{y}}_{t+h}^{(2)}\right) \tag{33}$$

- **Forecasting Type:** A statistical decomposition model for point forecasting, often used as a strong baseline for its simplicity and performance.

D.0.21 XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful and efficient implementation of the gradient boosting framework. It is an ensemble model that builds decision trees sequentially, where each new tree is trained to correct the errors made by the previous ones. For time series forecasting, XGBoost is used as a regression model on a featurized dataset, making it highly effective at capturing complex, non-linear relationships between the engineered features (e.g., lags, calendar variables) and the target. [30]

- **Input:** A feature matrix $\boldsymbol{X}$ and a corresponding target vector $\boldsymbol{y}$.
- **Output:** A point forecast for each input feature vector.
- **Architecture and Formulation:** XGBoost builds an additive model where the final prediction is the sum of the predictions from K decision trees:

$$\widehat{\boldsymbol{y}}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i) \tag{34}$$

  The trees are added one at a time in a greedy fashion. The k-th tree, f_k, is chosen to minimize a regularized objective function:

$$\mathcal{L}^{(k)} = \sum_{i=1}^{n} l(\boldsymbol{y}_i, \widehat{\boldsymbol{y}}_i^{(k-1)} + f_k(\boldsymbol{x}_i)) + \Omega(f_k) \tag{35}$$

  where l is a differentiable loss function, $\widehat{\boldsymbol{y}}_i^{(k-1)}$ is the prediction from the first $k-1$ trees, and Ω is a regularization term that penalizes the complexity of the tree:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{36}$$

  Here, T is the number of leaves in the tree, $\boldsymbol{w}$ is the vector of scores on the leaves, and γ and λ are regularization parameters.
- **Forecasting Type:** An ensemble machine learning model for point forecasting, known for its high performance, speed, and regularization capabilities.

D.1 Benchmark Tasks Included in TempusBench

In this section, we describe the datasets that have been used for each benchmark task. We summarize the dataset used for each benchmark task in Table 2.

D.2 Synthetic Data: Cyclic Seasonality with Additive Trends

D.2.1 Description

This category of synthetic data models a time series that exhibits both a complex seasonal pattern and a persistent, long-term trend. The data is generated using two related methods. Both methods start with a foundational signal that combines multi-frequency sinusoids with a linear trend. The second, more complex method builds upon this foundation by introducing an additional, randomized sinusoidal component to the signal.

In both cases, non-negative noise from an exponential distribution is added to the deterministic signal. These datasets are ideal for testing a model's ability to identify and separate periodicities from an underlying linear trend, with the second method providing a more complex seasonal structure.

D.2.2 Mathematical Formulation

The generation process for both methods is based on a primary signal, $y_{\text{base}}(t)$, which includes seasonal, trend, and offset components:

$$y_{\text{base}}(t) = \underbrace{2\sin(t) + 2\cos\left(\frac{t}{2}\right)}_{\text{Seasonality}} + \underbrace{\frac{1}{4}t}_{\text{Trend}} + \underbrace{4}_{\text{Offset}} \tag{37}$$

Method 1: Fixed Additive Trend In the first method, the true signal, $y_1(t)$, is simply the base function. The final observed value, Y_t, is this signal plus an additive noise term, ϵ_t.

$$Y_t = y_1(t) + \epsilon_t = y_{\text{base}}(t) + \epsilon_t \tag{38}$$

Method 2: Randomized Additive Trend The second method introduces additional complexity. For each generated time series, a random frequency parameter, α, is sampled once from a continuous uniform distribution:

$$\alpha \sim U(a, b) \tag{39}$$

In the provided code, this range is fixed from $a = 0$ to $b = 5$. This parameter is used to create an additional sinusoidal component that is added to the base signal. The true signal, $y_2(t)$, is therefore:

$$y_2(t) = y_{\text{base}}(t) + \sin(\alpha t) \tag{40}$$

The final observed value, Y_t, is this enhanced signal plus the noise term:

$$Y_t = y_2(t) + \epsilon_t \tag{41}$$

Noise Model For both methods, the noise term, ϵ_t, is drawn from an exponential distribution with a scale parameter β:

$$\epsilon_t \sim \text{Exponential}(\beta) \tag{42}$$

D.2.3 Adjustable Parameters

The data generation process is controlled by the following parameters.

- **Number of Points (`num_points`, N):** This integer parameter sets the total number of data points, defining the length of the time series.
- **Start Time (`start_time`, t_0):** This parameter defines the initial time value for the series.
- **Noise Scale (`noise_std`, β):** This parameter represents the scale (and mean) of the exponential noise distribution. A larger value for β increases the average magnitude of the positive noise added to the base signal.

Table 14: Summary of datasets used for benchmark tasks.

Benchmark	Task	l	h	n	m
Trend					
Multivariate (Non-stationary)	Electricity Consumption [33]	512	64	1741	44
Univariate (Non-stationary)	Software Development Job Postings [34]	512	64	1827	1
Decomposition					
Univariate (Additive)	Synthetically Generated Additive (Appendix D.2)	1024	64	3000	1
Univariate (Multiplicative)	Synthetically Generated Multiplicative (Appendix D.3)	1024	64	3000	1
Frequency					
Multivariate (Days)	Gold Price in India [35]	1024	64	4024	5
Univariate (Days)	Coinbase Litecoin [36]	512	64	1827	1
Multivariate (Hours)	Madrid Transport Pollution [37]	2048	64	181753	14
Multivariate (Minutes)	Historical Stock Data (2003-2024) [38]	2048	64	122110	6
Multivariate (Minutes)	Historical Stock Data (2003-2024, Longest) [38]	2048	64	122110	6
Multivariate (Months)	Airlines Baggage Complains [39]	32	8	84	4
Univariate (Months)	Inventories to Sales Ratio [40]	64	64	402	1
Univariate (Quarters)	German House Prices [41]	32	32	221	1
Multivariate (Seconds)	Utah Drilling [42]	2048	64	9661	35
Univariate (Weeks)	Federal Funds Effective Rate [43]	1024	64	3713	1
Univariate (Years)	Personal Consumption Expenditures [44]	32	8	96	1
Seasonality					
Multivariate (Periodic)	Madrid Transport (Cyclical) [37]	2048	64	181753	14
Univariate (Periodic)	Synthetic Cyclic	1024	64	3000	1
Univariate (Quasiperiodic)	Synthetic Non-stationary	1024	64	3000	1
Domain					
Univariate (Climate)	Delhi Climate [45]	512	64	1462	1
Multivariate (Economics/Finance)	Gold Price in India [35]	1024	64	4024	5
Multivariate (Economics/Finance)	Gold Price in India (Real) [35]	1024	64	4024	5
Univariate (Economics/Finance)	Coinbase Litecoin [36]	512	64	1827	1
Multivariate (Energy)	Room SplitSmart [46]	2048	64	10603	2
Univariate (Energy)	Room SplitSmart [46]	128	64	561	1
Multivariate (Healthcare)	NYC Covid Cases [47]	512	64	2005	54
Univariate (Healthcare)	Employees Health Care [48]	64	64	427	1
Univariate (Manufacturing)	Inventories to Sales Ratio [40]	64	64	402	1
Multivariate (Nature)	Soil Monitoring [49]	1024	64	4323	127
Multivariate (Nature)	Soil Monitoring (500) [49]	1024	64	4323	127
Univariate (Nature)	Soil Monitoring [49]	128	64	679	1
Multivariate (Sales)	Airlines Baggage Complains [39]	32	8	84	4
Univariate (Sales)	German House Prices [41]	32	32	221	1
Multivariate (Software)	Cyber Attacks on Water Networks [50]	512	64	1741	44
Univariate (Software)	Software Development Job Postings [34]	512	64	1827	1
Multivariate (Transport)	Airlines Baggage Complains (100) [39]	32	8	84	4
Multivariate (Transport)	Madrid BEN pollution [51]	2048	64	181753	14
Multivariate (Transport)	Madrid BEN pollution (Noisy) [37]	2048	64	181753	14
Univariate (Transport)	Madrid BEN pollution [51]	2048	64	172622	1
Univariate (Web)	Web Traffic [52]	1024	64	2793	1
Data sparsity					
Multivariate (Dense)	Gold Price in India [35]	1024	64	4024	5
Univariate (Dense)	Chicken Pox [53]	128	64	522	1
Univariate (Sparse)	Patient Chart [54]	2048	64	8093	1
Value type					
Univariate (Binary)	Absenteeism at Work [55]	128	64	740	1
Univariate (Categorical)	Online Retail [56]	2048	64	541909	1
Multivariate (Continuous)	Gold Price in India [35]	1024	64	4024	5
Univariate (Continuous)	Forest Fires [57]	128	64	517	1
Multivariate (Count)	Madrid BEN pollution [51]	2048	64	181753	14
Univariate (Count)	Occupancy [58]	2048	64	10129	1

- **Random Frequency (`alpha`, α):** (Method 2 only) This parameter is not set by the user but is sampled internally from a uniform distribution $U(0, 5)$ for each generated series. It introduces variability in the seasonal component across different datasets created by the second method.

D.3 Synthetic Data: Cyclic Seasonality with Multiplicative and Additive Trends

D.3.1 Description

This category of synthetic data models a time series characterized by a complex interaction of seasonal components and trends. A key feature is a multiplicative trend, where the amplitude of one of the seasonal components grows exponentially over time. This is combined with another stable seasonal component and a linear additive trend.

The data is generated using two related methods. The first method uses a fixed, deterministic signal. The second method introduces additional complexity by adding another sinusoidal component with a randomized frequency to the base signal. In both cases, non-negative noise from an exponential distribution is added. These datasets are particularly useful for testing a model's ability to handle heteroscedasticity, where the variance of the series changes over time, in the presence of other seasonalities and trends.

D.3.2 Mathematical Formulation

Both methods are built upon a primary signal, $y_{\text{base}}(t)$, which is a composite of several functions:

$$y_{\text{base}}(t) = \underbrace{e^{t/100}\sin(t)}_{\text{Multiplicative Seasonality}} + \underbrace{3\cos\left(\frac{t}{2}\right)}_{\text{Additive Seasonality}} + \underbrace{\frac{1}{2}t}_{\text{Linear Trend}} \tag{43}$$

Method 1: Fixed Multiplicative Trend In the first method, the true signal, $y_1(t)$, is simply the base function. The final observed value, Y_t, is this signal plus an additive noise term, ϵ_t.

$$Y_t = y_1(t) + \epsilon_t = y_{\text{base}}(t) + \epsilon_t \tag{44}$$

Method 2: Randomized Additive Component The second method adds another layer of seasonality. For each generated time series, a random frequency parameter, α, is sampled once from a continuous uniform distribution:

$$\alpha \sim U(a, b) \tag{45}$$

In the provided code, this range is fixed from $a = 5$ to $b = 10$. The true signal, $y_2(t)$, is the base signal plus this new randomized sinusoidal component:

$$y_2(t) = y_{\text{base}}(t) + \sin(\alpha t) \tag{46}$$

The final observed value, Y_t, is this enhanced signal plus the noise term:

$$Y_t = y_2(t) + \epsilon_t \tag{47}$$

Noise Model For both methods, the noise term, ϵ_t, is drawn from an exponential distribution with a scale parameter β:

$$\epsilon_t \sim \text{Exponential}(\beta) \tag{48}$$

D.3.3 Adjustable Parameters

The data generation process is controlled by the following parameters.

- **Number of Points (`num_points`, N):** This integer parameter sets the total number of data points, defining the length of the time series.
- **Start Time (`start_time`, t_0):** This parameter defines the initial time value for the series.

- **Noise Scale (`noise_std`, β):** This parameter represents the scale (and mean) of the exponential noise distribution. A larger value for β increases the average magnitude of the positive noise added to the base signal.
- **Random Frequency (`alpha`, α):** (Method 2 only) This parameter is not set by the user but is sampled internally from a uniform distribution $U(5, 10)$ for each generated series. It introduces variability in the seasonal component across different datasets created by the second method.

E Benchmark Evaluations

Our benchmark evaluation across multiple deterministic and stochastic forecasting metrics reveals several key insights regarding model performance and efficiency. Notably, LAFN, with only 0.4M (400K) parameters, demonstrates remarkable performance efficiency compared to significantly larger foundation models such as TimesFM (200M parameters) [20], Chronos (20M parameters) [21], and Moirai (approximately 91M for the base variant) [18].

On deterministic metrics, LAFN achieves the best performance on several tasks: it attains the lowest MAE on the Baggage dataset (0.358) and SplitSmart Energy task (0.645) (see Table 15), the lowest RMSE on Baggage (0.483) (Table 18), and the lowest MASE on both Baggage (0.689) and SplitSmart (1.524) (Table 17). This performance persists despite LAFN's compact architecture, achieving competitive or superior results compared to models with up to $2000\times$ more parameters. For instance, while LAFN achieves the best MAE on Baggage, TimesFM (200M parameters) achieves 0.526 on the same task, and Chronos (depending on variant) achieves 0.841. Similarly, on SplitSmart, LAFN's MASE of 1.524 outperforms TimesFM (1.617), demonstrating that parameter efficiency does not necessarily compromise forecasting accuracy.

On stochastic metrics, LAFN also shows strong performance, achieving the best CRPS scores on Madrid Count and Madrid Hours datasets (both 0.798) (Table 21), despite competing against large-scale foundation models. The quantile score (Table 19) and weighted interval score (Table 20) results further validate LAFN's capability to provide well-calibrated probabilistic forecasts with minimal model complexity.

These findings align with recent research on efficient forecasting architectures [4] and suggest that architectural design and training methodology are as important as model scale for time series forecasting [18, 21]. LAFN's success highlights the potential for lightweight yet effective forecasting models suitable for resource-constrained environments, while maintaining competitive performance against much larger foundation models [4].

Table 15: MAE Results

Model	Baggage	GoldIndia	IndiaGold	LtStock	Madrid Count	Madrid Hours	Madrid Trans.	Soil	SplitSmart	Utah Mfg.
prophet	1.3503	0.1703	0.1703	**1.129e-5**	0.8257	0.8257	0.8257	1.7661	0.7788	0.05313
timesfm	0.5265	0.08253	0.08253	1.631e-4	0.75	0.75	0.75	**0.3798**	0.6845	0.00557
chronos	0.8411	0.09603	0.0961	0.0034	1.0979	1.1446	1.1189	0.5629	0.6915	**0.00555**
moirai	0.6021	0.1127	0.1132	0.00418	4074.9799	—	12.2493	0.4652	0.6824	0.0103
exponential_smoothing	1.56	0.1322	0.1322	0.03214	1.1162	1.1162	1.1162	1.1566	0.7413	0.00588
arima	1.451	0.08869	0.08869	0.0377	0.9674	0.9674	0.9674	1.1532	0.7394	0.00594
croston_classic	1.3627	0.09168	0.09168	0.04007	**0.6607**	**0.6607**	**0.6607**	1.3323	0.7505	0.00595
seasonal_naive	1.2866	0.1089	0.1089	0.04068	0.8156	0.8156	0.8156	1.486	0.7609	0.00604
tiny_time_mixer	0.5614	0.1095	0.1095	0.00738	0.8694	0.8694	0.8694	0.4273	0.6896	0.01328
lstm	0.7253	0.1229	0.08364	0.09176	0.9037	0.9174	0.8942	0.6005	0.6949	0.01797
random_forest	0.7274	0.09343	0.09343	0.08293	0.9408	0.9408	0.9408	0.6318	0.7001	0.0272
moment	1.101	0.1355	0.1341	0.115	0.9741	0.965	0.9727	1.0086	0.712	0.03167
LAFN	**0.3584**	0.1259	0.1259	—	0.8623	0.8623	0.8623	0.5147	**0.645**	0.04306
tabpfn	0.523	0.08237	0.08237	—	1.2051	1.2051	1.2051	0.5751	0.6811	0.04421
svr	0.7274	**0.0807**	**0.0807**	0.1292	0.925	0.925	0.925	0.566	0.6847	0.05803
varmax	1.4483	0.1188	0.1188	—	—	—	—	—	0.763	—
lagllama	0.975	1.0764	0.9712	1.0818	1.0346	0.7924	1.0575	0.4772	0.8107	0.3648
toto	0.6329	—	—	—	—	—	—	—	—	—
theta	1.0456	0.7315	0.7315	3.6515	—	—	—	4.033	0.6483	—
moirai_moe	1.7957	—	—	—	—	—	—	—	—	—

Table 16: MAPE Results

Model	Baggage	GoldIndia	IndiaGold	LtStock	Madrid Count	Madrid Hours	Madrid Trans.	Soil	SplitSmart	Utah Mfg.
prophet	326.9251	180.2845	180.2845	**5.914e-4**	759.798	759.798	759.798	1977.3261	262.7489	12.6987
timesfm	**141.1057**	98.0174	98.0174	0.03857	711.3051	711.3051	711.3051	220.7421	113.7671	6.929e+8
chronos	312.9438	125.4187	127.6942	0.2912	839.1958	895.2794	828.343	269.6401	123.8572	3.516e+9
moirai	623.2775	81.6825	82.0073	0.5541	1.116e+7	—	1.656e+5	229.2512	109.6988	2.574e+8
tiny_time_mixer	223.935	161.4929	161.4929	1.0542	841.4785	841.4785	841.4785	**162.0339**	121.426	1.050e+9
croston_classic	377.9797	95.3737	95.3737	5.8565	**353.8011**	**353.8011**	**353.8011**	1396.8504	221.4088	**1.5506**
seasonal_naive	234.6152	113.4186	113.4186	5.9475	526.714	526.714	526.714	1676.5477	208.3699	1.5745
exponential_smoothing	517.0076	197.8321	197.8321	4.7063	740.4425	740.4425	740.4425	1430.5323	211.1025	1.792
arima	444.6815	91.3389	91.3389	5.5797	728.6722	728.6722	728.6722	1125.1788	209.3847	1.8021
random_forest	624.1922	92.6297	92.6297	12.185	732.1473	732.1473	732.1473	651.4781	147.4979	11.0236
tabpfn	480.2252	45.1138	45.1138	—	939.0734	939.0734	939.0734	547.5877	113.5734	11.693
lstm	577.3367	87.638	115.776	13.5608	701.1755	667.6615	731.9914	582.6126	131.8695	5.806e+7
moment	216.8776	**32.0207**	**25.6073**	16.8106	807.9779	796.8301	815.8814	889.9217	174.3516	4.780e+4
svr	624.1922	89.9103	89.9103	28.163	732.3681	732.3681	732.3681	397.0512	114.873	125.579
varmax	475.8379	33.7242	33.7242	—	—	—	—	—	239.2184	—
LAFN	205.5096	41.5552	41.5552	—	699.4932	699.4932	699.4932	317.9451	89.0227	1.587e+5
lagllama	262.3012	209.8353	145.1316	74.9169	515.0667	386.8356	627.3051	730.9717	163.071	159.1523
theta	218.2205	381.5479	381.5479	1054.1567	—	—	—	6122.3174	**86.1676**	—
toto	316.7353	—	—	—	—	—	—	—	—	—
moirai_moe	632.1069	—	—	—	—	—	—	—	—	—

Table 17: MASE Results

Model	Baggage	GoldIndia	IndiaGold	LtStock	Madrid Count	Madrid Hours	Madrid Trans.	Soil	SplitSmart	Utah Mfg.
prophet	2.5965	4.3376	4.3376	**0.02839**	2.7294	2.7294	2.7294	13.2827	1.8398	899.244
timesfm	1.0124	2.1017	2.1017	0.41	2.479	2.479	2.479	**2.8563**	1.6168	94.2659
LAFN	**0.6892**	3.2052	3.2052	—	2.8502	2.8502	2.8502	3.8706	**1.5237**	728.785
tabpfn	1.0057	2.0977	2.0977	—	3.9834	3.9834	3.9834	4.3253	1.609	748.3392
tiny_time_mixer	1.0795	2.7885	2.7885	18.5563	2.8738	2.8738	2.8738	3.2137	1.6291	224.8232
moirai	1.1578	2.8695	2.8826	10.509	1.347e+4	—	40.4898	3.499	1.6119	174.2673
toto	1.2169	—	—	—	—	—	—	—	—	—
lstm	1.3948	3.1305	2.13	230.634	2.987	3.0325	2.9556	4.5162	1.6415	304.1776
random_forest	1.3986	2.3795	2.3795	208.4371	3.1097	3.1097	3.1097	4.7518	1.6538	460.3945
svr	1.3986	**2.0552**	**2.0552**	324.7135	3.0577	3.0577	3.0577	4.2569	1.6174	982.1705
theta	2.0105	18.6298	18.6298	9177.467	—	—	—	30.3313	1.5314	—
chronos	1.6172	2.4457	2.4475	8.5428	3.6289	3.7835	3.6986	4.2335	1.6335	**93.8688**
moment	2.1171	3.4509	3.4155	288.9883	3.2198	3.1898	3.2153	7.5856	1.6818	535.9909
arima	2.79	2.2586	2.2586	94.7489	3.1976	3.1976	3.1976	8.6726	1.7467	100.5149
exponential_smoothing	2.9998	3.367	3.367	80.7743	3.6897	3.6897	3.6897	8.6984	1.751	99.4697
croston_classic	2.6202	2.3348	2.3348	100.7	**2.1839**	**2.1839**	**2.1839**	10.0198	1.7727	100.7
seasonal_naive	2.4739	2.773	2.773	102.25	2.6959	2.6959	2.6959	11.1759	1.7973	102.25
varmax	2.785	3.0245	3.0245	—	—	—	—	—	1.8023	—
lagllama	1.8748	27.4121	24.7327	2718.8557	3.4198	2.6191	3.4956	3.5888	1.9151	6173.8657
moirai_moe	3.4529	—	—	—	—	—	—	—	—	—

Table 18: RMSE Results

Model	Baggage	GoldIndia	IndiaGold	LtStock	Madrid Count	Madrid Hours	Madrid Trans.	Soil	SplitSmart	Utah Mfg.
prophet	1.4893	0.2411	0.2411	**1.892e-5**	1.291	1.291	1.291	1.933	**2.3175**	0.1627
timesfm	0.669	0.1738	0.1738	1.940e-4	**1.1987**	**1.1987**	**1.1987**	**0.623**	2.3527	**0.00743**
chronos	0.967	0.1911	0.1914	0.00495	1.7918	1.8709	1.7704	1.0103	2.3504	0.05167
moirai	0.9712	0.2037	0.2055	0.00499	1.030e+5	—	154.0111	0.7218	2.3553	0.03645
tiny_time_mixer	0.6816	0.2051	0.2051	0.01059	1.309	1.309	1.309	0.6702	2.3501	0.02331
exponential_smoothing	1.6599	0.2596	0.2596	0.03796	1.5877	1.5877	1.5877	1.368	2.3273	0.01826
arima	1.5465	0.182	0.182	0.04549	1.3808	1.3808	1.3808	1.3067	2.3274	0.01846
croston_classic	1.4597	0.1829	0.1829	0.04758	1.2076	1.2076	1.2076	1.4728	2.3258	0.01894
seasonal_naive	1.6317	0.2179	0.2179	0.0483	1.5016	1.5016	1.5016	1.6287	2.3393	0.01923
lstm	0.8749	0.2134	0.1721	0.1079	1.4227	1.458	1.4236	0.8143	2.3458	0.04982
random_forest	0.8789	0.1831	0.1831	0.09686	1.3707	1.3707	1.3707	0.8463	2.3412	0.06252
svr	0.8789	**0.1507**	**0.1507**	0.131	1.4476	1.4476	1.4476	0.7478	2.3468	0.07317
LAFN	**0.4835**	0.2483	0.2483	—	1.3087	1.3087	1.3087	0.7026	2.3899	0.08154
moment	1.2252	0.2642	0.2622	0.1347	1.4038	1.4048	1.4095	1.1703	2.3353	0.09986
tabpfn	0.6452	0.1607	0.1607	—	1.7946	1.7946	1.7946	0.7316	2.3536	0.1387
varmax	1.5476	0.2451	0.2451	—	—	—	—	—	2.3212	—
lagllama	1.1522	1.1101	1.0398	1.2697	1.9641	1.4825	1.9135	0.6247	2.4138	0.5353
theta	1.1915	0.775	0.775	3.7085	—	—	—	4.4442	2.3882	—
toto	0.9174	—	—	—	—	—	—	—	—	—
moirai_moe	1.8892	—	—	—	—	—	—	—	—	—

Table 19: Quantile Score Results

Model	Baggage	GoldIndia	IndiaGold	LtStock	Madrid Count	Madrid Hours	Madrid Trans.	Soil	SplitSmart	Utah Mfg.
moirai	1.4656	0.3607	0.3692	**0.0116**	**2.2638**	—	**2.2609**	1.6067	**2.9823**	**0.00498**
chronos	3.1579	**0.3429**	**0.3513**	0.01619	3.4517	3.7626	3.5154	2.1077	3.3491	0.00762
LAFN	1.8731	0.4775	0.4775	—	2.8502	2.8502	2.8502	2.0161	3.1046	0.2372
lagllama	3.1239	3.1333	2.9102	3.5586	3.5871	**2.6475**	3.5033	**1.5105**	3.469	1.1164
toto	**1.3928**	—	—	—	—	—	—	—	—	—
moirai_moe	6.4939	—	—	—	—	—	—	—	—	—

Table 20: Weighted Interval Score Results

Model	Baggage	GoldIndia	IndiaGold	LtStock	Madrid Count	Madrid Hours	Madrid Trans.	Soil	SplitSmart	Utah Mfg.
moirai	6.4414	1.5728	1.6072	**0.05126**	**9.9283**	—	**9.8992**	6.8191	**11.5368**	**0.02172**
chronos	12.4337	**1.398**	**1.4334**	0.067	14.426	15.7645	14.5568	8.2637	12.2022	0.03079
LAFN	8.3478	2.1039	2.1039	—	12.4137	12.4137	12.4137	8.699	11.7468	1.0216
lagllama	13.528	13.7197	12.7306	15.7445	15.3557	**11.5445**	15.3237	**6.6222**	13.2707	4.8724
toto	**6.128**	—	—	—	—	—	—	—	—	—
moirai_moe	26.8475	—	—	—	—	—	—	—	—	—

Table 21: CRPS Results

Model	Baggage	GoldIndia	IndiaGold	LtStock	Madrid Count	Madrid Hours	Madrid Trans.	Soil	SplitSmart	Utah Mfg.
moirai	0.5905	**0.1316**	**0.1344**	**0.00349**	1.497	—	**0.6881**	0.8374	**2.3426**	**0.00188**
chronos	1.8193	0.1365	0.1395	0.00611	1.1359	1.312	1.2141	1.3624	2.477	0.00448
LAFN	1.3921	0.1651	0.1651	—	**0.7978**	**0.7978**	0.7978	0.775	2.3691	0.08721
lagllama	1.1077	1.0429	0.9192	1.858	1.1399	0.9721	1.4797	**0.6107**	2.5584	0.3673
toto	**0.489**	—	—	—	—	—	—	—	—	—
moirai_moe	4.1489	—	—	—	—	—	—	—	—	—

F Additional Related Works

Classical time-series forecasting began with statistical models that exploit stochastic structure and domain priors, including ARIMA and its Box–Jenkins methodology [59], exponential-smoothing state-space ETS [60], the Theta method [61], and multivariate VAR models [62]. Deep learning methods later advanced accuracy and scale by learning nonlinear temporal dependencies from large corpora: DeepAR [63], N-BEATS [64], DLinear [15], TiDE [65], TFT [66], PatchTST [67], and iTransformer [68]. Probabilistic forecasters further model predictive distributions, e.g., diffusion-based TimeGrad [69], score-based CSDI for imputation and forecasting [70], and conditional-flow GRU-NVP [71].

TSFMs. Inspired by NLP/vision pretraining, TSFMs train on heterogeneous corpora and evaluate in zero/few-shot settings across domains and horizons. Representative models include Moirai [72], Chronos [21], TimesFM [73], Lag-Llama [74], Timer [75], UniTS [76], TTM (Tiny Time Mixers) [77], Moment [26], and multimodal VisionTS [78]. Collectively, they demonstrate strong zero-shot point and probabilistic accuracy on diverse benchmarks while revealing open challenges at long horizons (error accumulation) and at very high frequencies.

Public datasets and repositories. Public corpora have underpinned progress from statistical to foundation-model eras. The M-competitions (M3 and M4) provided broad univariate benchmarks across domains and frequencies [6, 7], followed by the retail-demand M5 competition [79]. The Monash Time-Series Forecasting Archive curates a large, standardized repository spanning many domains and sampling granularities [13]. Large-scale pretraining/evaluation collections include LOTSA (released with Moirai) [72], the Chronos corpus with in-domain/zero-shot splits [21], and the diverse univariate corpus aggregated in Lag-Llama [74]. Task-focused collections such as the LTSF suite [15] (e.g., ETT datasets) and broader benchmarks like TFB [14] and ProbTS [80] assemble datasets emphasizing horizon length, covariates, and probabilistic outputs.

Evaluation frameworks and benchmarks. Tooling and standardized evaluation have evolved in parallel. Practitioner libraries such as Prophet [81] and sktime [82] offer classical and ML baselines with unified interfaces, while GluonTS [83] and PyTorchTS [84] provide probabilistic deep-learning pipelines. Benchmarking efforts including LTSF [15], BasicTS+ [**?**], TFB [14], and ProbTS [80] compare statistical, deep, and (in some cases) foundation models, but differ in task taxonomies, splits, and leakage controls. Standardized metrics such as MASE [85] and CRPS [86] enable cross-dataset aggregation of point and probabilistic performance, yet consistent pretraining/evaluation protocols and leakage-free large-scale corpora remain key needs for fair TSFM assessment.

The collective consequence of these issues is a research environment where it is difficult to distinguish genuine methodological advances from circumstantial performance on a narrow, and potentially contaminated, set of tasks. This is particularly damaging for the development of TSFMs. The significant computational and financial resources required to pre-train these models demand a rigorous, fair, and comprehensive evaluation framework to justify their development and guide future research [32]. The current state of affairs falls short of this standard. Indeed, studies have shown that existing TSFMs, often pre-trained on general-purpose academic datasets, can struggle to generalize to the unique and challenging characteristics of specialized domains like observability data [87].

The field has thus reached an inflection point. Progress is no longer primarily limited by our ability to design novel model architectures, but by our inability to reliably and fairly measure their performance. Recognizing this crisis, recent efforts have focused on creating the next generation of evaluation infrastructure. The development of large-scale, standardized benchmarks such as GIFT-Eval and the domain-specific Benchmark of Observability Metrics (BOOM) represent a direct and necessary response [27]. These initiatives introduce carefully curated and decontaminated pre-training and evaluation sets, standardized protocols, and data that reflects the complexity of real-world applications. They treat the benchmark not as a mere dataset, but as a carefully designed scientific instrument [8]. This establishes a clear and urgent research gap: the critical need for a new, large-scale, and meticulously curated public benchmark that can serve as a gold standard for evaluating the next generation of time-series models. Such a contribution is not merely a prerequisite for future research; it is a foundational scientific contribution in its own right, providing the essential infrastructure required to move the field from an era of fragmented claims to one of robust, reproducible, and generalizable progress [26].

Contemporary time-series data seldom conform to the idealized assumptions of stationarity and linearity that underpin classical models. Instead, real-world data streams are characterized by a confluence of complex, interacting properties that present formidable modeling challenges [30].

- **Non-Linearity:** Perhaps the most fundamental challenge is the prevalence of non-linear relationships. Economic systems, biological processes, and energy grids are governed by complex feedback loops and interactions that cannot be adequately captured by linear models [27]. Traditional methods like Autoregressive Integrated Moving Average (ARIMA) are, by their construction, limited in their ability to model such non-linear dynamics, which is a primary reason for their performance ceiling on complex, real-world problems [30].
- **Multi-Regime Behavior:** Many time series exhibit structural breaks or distinct operational regimes, where the underlying data-generating process changes over time [5]. Examples include the shift between bull and bear markets in financial data or the different performance characteristics of an industrial machine under varying loads and environmental conditions. A single, global model often fails to capture this complex inner structure, leading to significant predictive errors when the system transitions between regimes [88].
- **Intermittency:** As noted previously, intermittent demand patterns are characterized by a high proportion of zero-valued observations, with non-zero demands occurring sporadically. This dual source of randomness—in both the timing and the magnitude of events—violates the assumptions of continuity and regular sampling inherent in many classical smoothing and regression-based techniques [5].
- **Heightened Volatility and Novel Data Sources:** The modern data ecosystem is characterized by the emergence of new data sources that introduce unprecedented levels of volatility and complexity [5]. The integration of renewable energy sources into power grids is a prime example, creating load patterns with high-frequency noise and non-stationary behavior that challenge traditional forecasting approaches. A parallel development is the explosion of "observability data" generated by large-scale distributed software and cloud computing systems. This data, which includes metrics on CPU load, network latency, and application error rates, is often characterized by extreme non-stationarity, high dimensionality (thousands of correlated variables), heavy-tailed distributions, and sparsity, posing a unique and difficult set of modeling challenges [5].

F.0.1 An Arms Race of Methodological Innovation

The progression of forecasting methodologies can be understood as a direct response to this escalating data complexity. Each new paradigm has sought to overcome the limitations of its predecessors, leading to the current diverse and powerful toolkit available to researchers and practitioners [5].

- **The Classical Foundation:** The field was built upon a foundation of statistical methods developed primarily in the mid-20th century. Models such as ARIMA and its variants [8], Holt-Winters Exponential Smoothing [5], and the Theta method [32] became the workhorses of the discipline. These models excel at capturing and extrapolating clear patterns of trend and seasonality from univariate time series. Their enduring appeal lies in their statistical rigor, interpretability, and computational efficiency. However, their reliance on strong assumptions about the underlying data-generating process, particularly linearity and stationarity, fundamentally limits their applicability to the more complex data common today [8].
- **The Machine Learning Advance:** The rise of machine learning in the late 20th and early 21st centuries provided a new set of tools capable of addressing the challenge of non-linearity. Non-parametric models like Support Vector Regression (SVR) [31] offered a principled approach, grounded in statistical learning theory, to model non-linear relationships in high-dimensional spaces via the "kernel trick" [27]. Concurrently, ensemble methods, particularly those based on decision trees like Random Forest and Gradient Boosting (e.g., XGBoost), proved to be exceptionally powerful and robust [30]. By combining the predictions of many weak learners, these models can capture complex, non-linear interactions and have consistently demonstrated state-of-the-art performance in a wide array of forecasting competitions and applications.
- **The Deep Learning Revolution:** While machine learning ensembles excelled at capturing complex feature interactions, they were not explicitly designed to model the long-range

temporal dependencies inherent in sequential data. This limitation was addressed by the deep learning revolution. Recurrent Neural Networks (RNNs) , and more specifically architectures like Long Short-Term Memory (LSTM) networks, were developed with internal memory mechanisms (gates) that allow them to capture and retain information over long sequences [27]. Empirical studies have shown that on complex financial and economic data, LSTMs can significantly outperform classical models like ARIMA by better modeling non-linear temporal dynamics [5]. Following the success of LSTMs, Transformer-based architectures, with their self-attention mechanism, have emerged as the next frontier, offering a powerful alternative for capturing dependencies across time without the sequential processing limitations of RNNs [27].

This co-evolution of data challenges and modeling paradigms, summarized in Table 22, illustrates a clear trajectory towards models of increasing complexity and representational power.

Table 22: The Co-evolution of Time-Series Challenges and Modeling Paradigms [5] [32] [8] [27] [30].

Era	Primary Data Challenge(s)	Dominant Model Paradigm	Key Models	Inherent Limitations
Statistical	Trends, Seasonality, Stationarity	Time-Domain Statistical Models	ARIMA, Holt-Winters, Theta	Struggle with non-linearity and complex dependencies
Machine Learning	Non-Linearity, Complex Interactions	Non-parametric	Ensemble Models (SVR, Random Forest, XGBoost)	Limited handling of long-range temporal dependencies
Deep Learning	Long-Range Dependencies, Sequential Patterns	Recurrent, Attention-based Networks	LSTMs, Transformers	Data-hungry, computationally intensive, task-specific
Foundation Models	Heterogeneity, Scale, Task Generalization	Large Pre-trained Models	MOMENT, TOTO, Chronos	Reliance on massive, curated datasets; evaluation bottleneck

However, this progression is not a simple linear march where newer, more complex models invariably render older ones obsolete. Empirical evidence reveals a more nuanced reality, one that aligns with the well-known "No Free Lunch" theorem in machine learning. While deep learning models like LSTMs have been shown to decisively outperform ARIMA on certain complex datasets, recent large-scale studies have also found that in zero-shot or limited-supervision settings, simpler statistical methods often outperform sophisticated deep learning models [27]. Furthermore, in production environments like large-scale observability systems, classical models remain prevalent due to the operational infeasibility of training and maintaining millions of distinct, complex neural network models [5]. This apparent contradiction is not a flaw in the research, but rather a reflection of a fundamental truth: the performance of any given forecasting model is highly contingent on the specific characteristics of the data, the length of the forecast horizon, the availability of computational resources, and the degree of supervision. This recognition implies that the central problem in the field is not merely the invention of more powerful algorithms, but the development of a deeper, more systematic understanding of the complex performance landscape that governs the interaction between data characteristics and model architectures [27].

F.1 The New Frontier: Pre-trained Foundation Models for Time Series

In response to the challenges of data heterogeneity and the high cost of developing task-specific models, the field is currently undergoing another paradigm shift, mirroring recent transformations in natural language processing and computer vision: the move towards large, pre-trained Time-Series Foundation Models (TSFMs). This new frontier aims to leverage the power of large-scale, self-supervised learning to create general-purpose models that can be adapted to a wide range of downstream tasks with minimal fine-tuning [27].

The core premise of the foundation model paradigm is to pre-train a single, high-capacity model (typically a Transformer) on a massive and diverse corpus of unlabeled data. This process allows the model to learn a rich, generalizable representation of temporal patterns. Subsequently, this pre-trained model can serve as a powerful building block for various downstream applications, including long- and short-horizon forecasting, time-series classification, anomaly detection, and missing value imputation. Models such as MOMENT, Chronos, and TOTO are at the vanguard of this movement. They are designed to be effective "out-of-the-box," providing strong zero-shot or few-shot performance without the need for extensive task-specific training [26] [25]. This approach holds particular promise for domains like observability, where the sheer scale and diversity of time series—often numbering in the millions or billions—make the traditional approach of training one model per series operationally intractable [30].

The primary enabler of this paradigm, and simultaneously its greatest bottleneck, is the availability of data. The success of foundation models in other domains was built on the existence of vast, cohesive, and publicly accessible datasets like The Pile for text and ImageNet for vision [26]. The time-series domain, by contrast, has historically been characterized by a fragmented landscape of smaller, scattered, and task-specific public datasets [5]. This data scarcity has been a major impediment to large-scale pre-training. To overcome this, pioneering research efforts have begun the monumental task of data curation. The creators of MOMENT compiled *The Time Series Pile*, a large collection of public repositories, while the TOTO model was pre-trained on a corpus containing a mixture of public, synthetic, and large-scale proprietary observability data, resulting in a dataset 4 to 10 times larger than those used for other leading TSFMs [25].

This focus on data curation signals a significant maturation of the field. In earlier eras, the primary axis of innovation was model architecture—for example, the design of the gating mechanisms in an LSTM or a novel attention variant in a Transformer [27]. The advent of the TSFM paradigm, however, shifts the research bottleneck. While architectural innovation remains important, the most critical and scientifically challenging work is now increasingly centered on the curation of massive, diverse, and clean datasets, and on the development of robust frameworks for evaluating the models trained on them. The value proposition of a new TSFM is now as much about the data it was trained on and the benchmark it was tested against as it is about its internal architecture. This implies that the most impactful contributions in this new era may not be designing a marginally better model, but rather creating the foundational data and evaluation infrastructure that enables the entire field to advance [8].