TS-ARENA: A LIVE FORECAST PRE-REGISTRATION PLATFORM FOR LEAKAGE-FREE EVALUATION OF TIME SERIES FOUNDATION MODELS

Marcel Meyer, Sascha Kaltenpoth, Henrik Albers, Kevin Zalipski & Oliver Müller
Data Analytics Group, Paderborn University
33098 Paderborn, Germany
firstname.lastname@uni-paderborn.de

ABSTRACT

Time Series Foundation Models (TSFMs) are transforming forecasting, yet evaluating them on historical data is increasingly compromised by train-test sample overlaps and temporal overlaps between correlated training and test series. This paper presents TS-Arena, an infrastructure designed to benchmark TSFMs by transitioning from retrospective historical testing to an environment of continuous, prospective evaluation. Our core contribution is a strict *Forecast Pre-Registration Protocol* (FPRP): models must submit predictions before the ground-truth data physically exists, making test-set contamination impossible by design. The platform relies on a modular microservice architecture that ingests real-time data streams and orchestrates containerized model submissions under enforced registration windows. By combining pre-registration with continuous evaluation rounds, TS-Arena aims to prevent both direct and indirect information leakage while enabling fast, ongoing model comparison. First results from simulating TS-Arena over one year of energy time series (e.g. renewable energy generation, electricity prices, district heating cogeneration) demonstrate the viability and discriminative power of leakage-free live evaluation. The live forecasting benchmark is available at ts-arena.live.

Track: Research

1 INTRODUCTION

Time series forecasting is vital across domains such as finance (Zhang et al., 2025), energy (Meyer et al., 2025b), and operations (Klee & Xia, 2025). The field is undergoing a paradigm shift toward Time Series Foundation Models (TSFMs): large neural networks pre-trained on broad, heterogeneous collections of time series that can forecast in a zero-shot manner (Liang et al., 2024), following scaling laws analogous to those observed in LLMs (Edwards et al., 2024; Kaplan et al., 2020; Yao et al., 2025). These models demonstrate zero-shot performance that is competitive with or superior to models requiring retraining (Aksu et al., 2024; Li et al., 2025).

The emergence of TSFMs necessitates a rethink of evaluation metrics, specifically to address the dual risks of direct and indirect information leakage. First, **direct information leakage** (test data contamination) occurs when training datasets inadvertently include benchmark test data. This phenomenon is well-documented in LLMs (Liao & Xiao, 2023) and now increasingly observed in TSFMs (Aksu et al., 2024; Meyer et al., 2025a). Second, **indirect temporal information leakage** is specific to time series: it arises when TSFMs are evaluated on held-out test series whose forecast horizons *temporally* overlap with training series from *correlated* time series (e.g., bus usage counts as training data and metro usage counts as zero-shot evaluation from the same period (Rodrigo & Ortiz, 2024)), allowing implicit access to future information through shared temporal structure (Meyer et al., 2025a). Both types of leakage can substantially inflate reported TSFM performance, leading to unreliable conclusions (Aksu et al., 2024; Meyer et al., 2025a; Rodrigo & Ortiz, 2024).

Existing approaches cannot fully resolve these concerns. Introducing fixed pre-training and test splits (Aksu et al., 2024) constrains the flexibility needed to scale TSFMs with diverse data (Edwards

et al., 2024; Yao et al., 2025). Benchmarks based on historical data (Aksu et al., 2024; Goktas et al., 2025; Li et al., 2025; Xu et al., 2025), regardless of how carefully curated, remain vulnerable once published test samples are used for training or when models train on correlated series from the same time period. More fundamentally, any fixed dataset that does not evolve over time can eventually lead to information leakage (Meyer et al., 2025a).

To address these inherent challenges, we propose TS-Arena, which leverages the unique generative nature of time series data: observations are continuously produced in real time. Our core contribution is a *Forecast Pre-Registration Protocol* (FPRP) that enforces the submission of forecasts *before* the ground-truth data exists. By evaluating pre-registered forecasts only after the ground truth materializes, this protocol makes both direct and indirect information leakage impossible by design. TS-Arena implements this protocol as a live benchmarking platform with continuous evaluation rounds, providing a forward-only, leakage-free evaluation infrastructure.

2 Forecast Pre-Registration Protocol

Building on the concept of *living* benchmarks (Erickson et al., 2025), TS-Arena adapts continuous maintenance and evolving leaderboards to the specific characteristics of time series forecasting. Here, *live* implies not merely a continuously updated ranking, but a dynamic test environment where evaluation targets are generated in real time by the unfolding future.

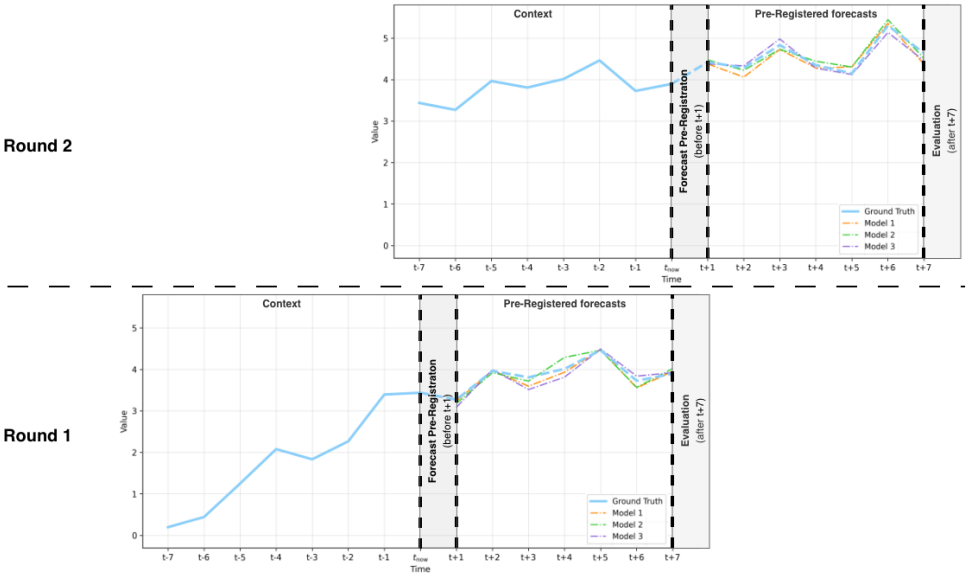

Figure 1: The Forecast Pre-Registration Protocol (FPRP). Models submit predictions within a registration window that closes before the next ground-truth value exists. Consecutive rounds enable continuous, leakage-free evaluation.

While current living benchmarks rely on retrospective train-test splits of static datasets, we propose a novel Forecast Pre-Registration Protocol (FPRP). It uses the evaluation time t as a strict temporal split point, denoted as t_{now}, that advances continuously with real time. Models have full access to historical ground-truth data up to t_{now} and must commit their forecasts for a fixed future horizon H before the first future ground-truth value becomes available. The protocol is applied continuously in *rounds*, which are fast-paced competitions (Bojer & Meldgaard, 2021; Makridakis et al., 2024), and proceeds as follows:

1. **Round Initialization:** At a predefined time, a new round is initialized for a set of time series with common characteristics.
2. **Context Acquisition:** The historical ground truth $[X_{t_{\text{now}}-C}, \ldots, X_{t_{\text{now}}}]$ with context length C is provided to participants.

3. **Inference and Registration:** Participants generate forecasts for the horizon $[t_{\text{now}} + 1, \ldots, t_{\text{now}} + H]$ and upload them within the registration window.
4. **Window Closure:** The registration window closes *before* the next ground-truth value $X_{t_{\text{now}}+1}$ is observed. Late submissions are rejected.
5. **Evaluation:** Once ground-truth values for the full horizon have materialized, error metrics are computed.

Figure 1 illustrates the protocol across two consecutive rounds. By strictly closing the submission window before any future data point comes into existence, the test set remains logically inaccessible during inference. This prevents any form of information leakage and look-ahead bias by design. The iterative repetition of rounds enables robust longitudinal evaluation analogous to time series cross-validation (Hyndman & Athanasopoulos, 2021).

3 TS-ARENA PLATFORM

To implement the FPRP, TS-Arena employs a modular microservice architecture orchestrated via Docker Compose, comprising: (i) a **Database** and (ii) a **Data Portal** that continuously ingests real-time data from external APIs, normalizes heterogeneous formats, and stores data into the database using a Slowly Changing Dimension Type 2 (SCD2) archiving strategy (Kimball, 2013) to enable exact reconstruction of information states at any historical t_{now}; (iii) an **API Portal** that manages user authentication, creates challenge rounds, validates submissions against active registration windows, and triggers evaluation; (iv) a **Reference Model Service** hosting containerized TSFMs and statistical baselines that participate autonomously under identical constraints; and (v) a **Front-end** based on (vi) a **Dashboard API** providing real-time leaderboards and challenge visualizations. The Front-end is shown in Figure 2.

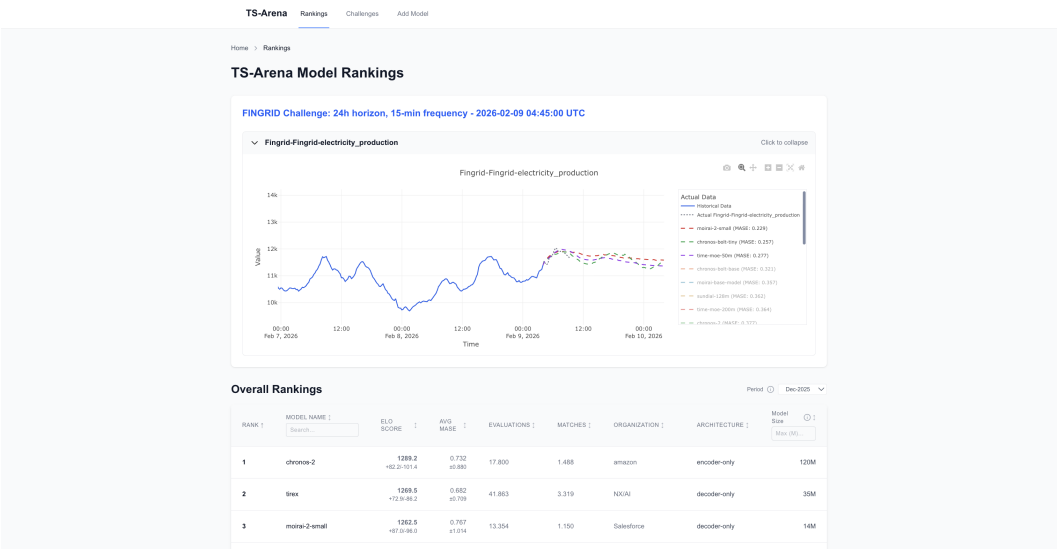

Figure 2: TS-Arena Challenge and Model View

Challenges and Evaluation. Time series are grouped into *challenges* by domain, frequency, and forecast horizon. Each challenge runs iterative rounds following the FPRP schedule. For evaluation, we compute the Mean Absolute Scaled Error (MASE) (Hyndman & Koehler, 2006) per round and report Elo ratings with bootstrap confidence intervals ($B = 500$) via randomized replay (Erickson et al., 2025). This pairwise ranking system naturally handles asynchronous model entry: established models accumulate narrow confidence intervals reflecting statistical reliability, while newcomers exhibit wider intervals that reflect their provisional status.

4 PRELIMINARY RESULTS

We populated TS-Arena with 186 live energy time series from three international sources (SMARD, Gridstatus, FINGRID), organized into 14 challenges at 15-minute and hourly frequencies with 1-day and 3-day horizons. To initialize the platform, we backtested the year 2025 with 13 TSFMs and statistical baselines, yielding over 5,000 challenge rounds and 3.9 million test data points. Reference TSFMs participated only from their respective release dates onward, consistent with the leakage-free evaluation principle.

The results confirm the discriminative power of live evaluation: all top-ranked models are TSFMs, with recent model generations (e.g., Chronos-2, Moirai 2) outperforming their predecessors, and larger model variants consistently ranking higher within families. Statistical baselines such as seasonal naive are clearly separated from TSFMs. Importantly, confidence intervals of top models often overlap, indicating the need for sustained evaluation over time to establish reliable rankings—precisely the capability that the continuous nature of TS-Arena provides. Detailed results of the year 2025 are available in Appendix B.

5 RELATED WORK

While earlier TSFMs usually had no choice than defining their own evaluation data (Ansari et al., 2024a; Das et al., 2024), recent benchmarks have substantially improved TSFM evaluation. TSFM-Bench (Li et al., 2025) expands model coverage across zero-shot, few-shot, and full-shot regimes with diverse datasets, but neither addresses nor discusses information leakage. Fev-Bench measures benchmark leakage by percentage of the test set included in pre-training (Shchur et al., 2025). However, the actual impact of information leakage on the inflation factor remains an open question. GIFT-Eval (Aksu et al., 2024) introduces a fixed test set intended for exclusion from future training; still, this constrains TSFM scaling flexibility and cannot prevent future contamination. Tempus-Bench (Goktas et al., 2025), BOOM (Cohen et al., 2025), and Fidel-TS (Xu et al., 2025) mitigate contamination through less common, private or API-protected datasets, yet these cannot guarantee that their data will not eventually enter TSFM training sets in the future. TabArena (Erickson et al., 2025) pioneered the living benchmark concept for tabular data, which we adapt and extend to exploit the unique temporal and generative properties of time series data.

In contrast to all existing approaches, TS-Arena eliminates information leakage *by design* through the FPRP: since ground-truth data does not physically exist at inference time, no form of contamination—whether direct sample overlap or indirect temporal correlation—is possible.

6 DISCUSSION AND CONCLUSION

We presented TS-Arena, a live benchmark platform for TSFMs that enforces a Forecast Pre-Registration Protocol: forecasts must be submitted before the corresponding ground truth exists. This simple temporal commitment eliminates benchmark leakage by design and enables continuous, transparent model comparison on evolving real-world data[1]. While the current scope is focused on univariate energy time series, TS-Arena is inherently extensible: as a living benchmark, it continuously incorporates new data sources, frequencies, horizons, and challenge types. A trade-off of our design is that while we impose no restrictions on pre-training data, current reliance on public APIs means we cannot guarantee true zero-shot evaluation for future models. As such, it complements existing static benchmarks that offer breadth and controlled analysis with a forward-only, leakage-free evaluation infrastructure that maintains its integrity as models and data evolve.

Future extensions include additional domains, multivariate forecasting with covariates, private data API's, and probabilistic evaluation metrics to further broaden the platform's evaluation capabilities.

[1] Links to the code are in Appendix A.1

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia under Grant No. EFRE-20801517 (iHeatPlan - Robust Planning of Grid-Based Heat Supply for Cities and Districts), and we gratefully acknowledge their support.

REFERENCES

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, November 2024. URL http://arxiv.org/abs/2410.10393. arXiv:2410.10393 [cs].

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

Abdul Fatir Ansari, Caner Turkmen, Oleksandr Shchur, and Lorenzo Stella. Fast and accurate zero-shot forecasting with Chronos-Bolt and AutoGluon, December 2024b. URL https://aws.amazon.com/blogs/machine-learning/fast-and-accurate-zero-shot-forecasting-with-chronos-bolt-and-autogluon/. tex.howpublished: AWS Machine Learning Blog.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From Univariate to Universal Forecasting, October 2025. URL http://arxiv.org/abs/2510.15821. arXiv:2510.15821 [cs].

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-Shot Forecasting Across Long and Short Horizons with Enhanced In-Context Learning. *arXiv preprint arXiv:2505.23719*, 2025.

Casper Solheim Bojer and Jens Peder Meldgaard. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2):587–603, April 2021. ISSN 01692070. doi: 10.1016/j.ijforecast.2020.07.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207020301114.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, and Afshin Rostamizadeh. This Time is Different: An Observability Perspective on Time Series Foundation Models. *arXiv preprint arXiv:2505.14766*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. 2024.

Thomas D. P. Edwards, James Alvey, Justin Alsing, Nam H. Nguyen, and Benjamin D. Wandelt. Scaling-laws for Large Time-series Models, 2024. URL https://arxiv.org/abs/2405.13867. Version Number: 2.

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.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A Living Benchmark for Machine Learning on Tabular Data, 2025. URL https://arxiv.org/abs/2506.16791. Version Number: 4.

Denizalp Goktas, Amy Greenwald, Gerardo Riano-Briceno, Alexandra Magnusson, Alif Abdullah, and Beatriz de Lucio. TempusBench: An evaluation framework for time-series forecasting. In *Recent advances in time series foundation models have we reached the 'BERT moment'?*, 2025. URL https://openreview.net/forum?id=3fMa060Ag5.

Lars Graf, Thomas Ortner, StanisĹ WoĹşniak, and Angeliki Pantazi. Flowstate: Sampling rate invariant time series forecasting. *arXiv preprint arXiv:2508.05287*, 2025.

Tao Hong, Pierre Pinson, Yi Wang, Rafal Weron, Dazhi Yang, and Hamidreza Zareipour. Energy Forecasting: A Review and Outlook. *IEEE Open Access Journal of Power and Energy*, 7:376–388, 2020. ISSN 2687-7910. doi: 10.1109/OAJPE.2020.3029979. URL https://ieeexplore.ieee.org/document/9218967/.

Rob Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 3rd edition, 2021.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, October 2006. ISSN 01692070. doi: 10.1016/j.ijforecast.2006.03.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207006000239.

Max Kanter. gridstatus: Extract data from ISOs and other energy grid sources, 2025. URL https://github.com/gridstatus/gridstatus.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, 2020. URL https://arxiv.org/abs/2001.08361. Version Number: 1.

Ralph Kimball. *The data warehouse toolkit: the definitive guide to dimensional modeling*. J. Wiley & Sons, Erscheinungsort nicht ermittelbar, 3rd ed edition, 2013. ISBN 978-1-118-53080-1.

Steven Klee and Yuntian Xia. Measuring time series forecast stability for demand planning. In *KDD 2025 workshop on AI for supply chain: Today and future*, 2025. URL https://openreview.net/forum?id=26zedugY8W.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, July 2021. ISSN 03062619. doi: 10.1016/j.apenergy.2021.116983. URL https://linkinghub.elsevier.com/retrieve/pii/S0306261921004529.

Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, and Bin Yang. TSFM-Bench: A Comprehensive and Unified Benchmark of Foundation Models for Time Series Forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 5595–5606, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 979-8-4007-1454-2. doi: 10.1145/3711896.3737442. URL https://doi.org/10.1145/3711896.3737442.

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*, pp. 6555–6565, Barcelona Spain, August 2024. ACM. ISBN 979-8-4007-0490-1. doi: 10.1145/3637528.3671451. URL https://dl.acm.org/doi/10.1145/3637528.3671451.

Q. Vera Liao and Ziang Xiao. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap, 2023. URL https://arxiv.org/abs/2306.03100. Version Number: 4.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025b.

Spyros Makridakis, Evangelos Spiliotis, Ross Hollyman, Fotios Petropoulos, Norman Swanson, and Anil Gaba. The M6 forecasting competition: Bridging the gap between forecasting and investment decisions. *International Journal of Forecasting*, pp. S0169207024001079, November 2024. ISSN 01692070. doi: 10.1016/j.ijforecast.2024.11.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207024001079.

Marcel Meyer, Sascha Kaltenpoth, Kevin Zalipski, and Oliver Müller. Time Series Foundation Models: Benchmarking Challenges and Requirements, 2025a. URL https://arxiv.org/abs/2510.13654. Version Number: 1.

Marcel Meyer, David Zapata Gonzalez, Sascha Kaltenpoth, and Oliver Müller. Benchmarking Time Series Foundation Models for Short-Term Household Electricity Load Forecasting. *IEEE Access*, 13:218141–218153, 2025b. ISSN 2169-3536. doi: 10.1109/ACCESS.2025.3648056. URL https://ieeexplore.ieee.org/document/11314515/.

Joaquín Amat Rodrigo and Javier Escobar Ortiz. Data leakage in pre-trained forecasting models, 2024. URL https://cienciadedatos.net/documentos/py63-data-leakage-pre-trained-forecasting-models.html.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A Realistic Benchmark for Time Series Forecasting, 2025. URL https://arxiv.org/abs/2509.26468. Version Number: 2.

Lefei Shen, Mouxiang Chen, Xu Liu, Han Fu, Xiaoxue Ren, Jianling Sun, Zhuo Li, and Chenghao Liu. VisionTS++: Cross-Modal Time Series Foundation Model with Continual Pre-trained Vision Backbones. *arXiv preprint arXiv:2508.04379*, 2025.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*, 2024.

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

Zhijian Xu, Wanxu Cai, Xilin Dai, Zhaorong Deng, and Qiang Xu. Fidel-TS: A High-Fidelity Multimodal Benchmark for Time Series Forecasting, 2025. URL https://arxiv.org/abs/2509.24789. Version Number: 3.

Qingren Yao, Chao-Han Huck Yang, Renhe Jiang, Yuxuan Liang, Ming Jin, and Shirui Pan. Towards neural scaling laws for time series foundation models. In *The thirteenth international conference on learning representations*, 2025. URL https://openreview.net/forum?id=uCqxDfLYrB.

Xu Zhang, Zhengang Huang, Yunzhi Wu, Xun Lu, Erpeng Qi, Yunkai Chen, Zhongya Xue, Qitong Wang, Peng Wang, and Wei Wang. Multi-period Learning for Financial Time Series Forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 2848–2859, Toronto ON Canada, July 2025. ACM. ISBN 979-8-4007-1245-6. doi: 10.1145/3690624.3709422. URL https://dl.acm.org/doi/10.1145/3690624.3709422.

A TS-ARENA MICROSERVICE ARCHITECTURE

A.1 CODE AVAILABILITY

Overview (Project Description, Participation Example):
`https://github.com/DAG-UPB/ts-arena`

Backend code (API-Portal, Data-Portal, Database, Dashboard-API):
`https://github.com/DAG-UPB/ts-arena-backend/`

Reference models:
`https://github.com/DAG-UPB/ts-arena-models`

Frontend:
`https://github.com/DAG-UPB/ts-arena-frontend`

Website:
`https://ts-arena.live/`

A.2 OVERALL ARCHITECTURE

TS-Arena employs a modular microservice architecture using Docker Compose, illustrated in Figure 3.

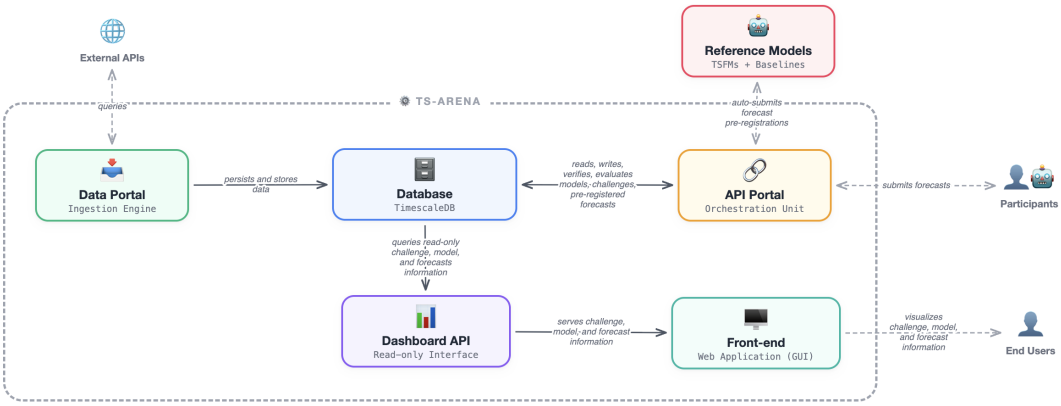

Figure 3: The Microservice Architecture

Time series are bundled into challenges based on shared characteristics such as domain, frequency, context length, or forecast horizon. Challenges are a construct for organizing participation at the TS-Arena and to distribute compute load across the day. Each challenge is executing iterative rounds following the FPRP protocol. The primary services of the architecture are:

Database: A TimescaleDB[2] instance manages storage and retrieval of high-frequency time series data, aggregates, forecast submissions, and user metadata.

Data Portal: Ingestion engine that queries external APIs, normalizes heterogeneous data formats, converts timestamps to UTC, and persists data using the Slowly Changing Dimension Type 2 (SCD2) archiving strategy (Kimball, 2013).

API Portal: Managing user authentication, challenge creation, forecast validation against registration windows, and evaluation triggering upon ground truth availability.

Dashboard API: A read-only API supplying the front-end with necessary information.

Front-end: A web application providing visualizations of challenges, model information, and live leaderboards.

Reference Model Service: Hosts containerized TSFMs and statistical baselines that autonomously participate in every challenge round.

[2]https://github.com/timescale/timescaledb

A.3 Database and Data Portal

TS-Arena initially focuses on energy domain challenges, which provide high-frequency data availability (Kanter, 2025) and complex, dynamic patterns such as wind and solar generation, human behavior in load patterns, and economic mechanisms in pricing (Hong et al., 2020; Lago et al., 2021; Meyer et al., 2025b). Our challenges cover load, generation, and price forecasting across various frequencies, horizons, and aggregation levels from market zones to trading hubs.

A.4 API Portal

Following the FPRP, challenges are conducted in *rounds*, where each round represents a single FPRP execution (Section 2). Rounds initialize at pre-defined, publicly visible times according to challenge-specific schedules.

We calculate the Mean Absolute Scaled Error ($MASE$) with naive baseline for each forecast (Hyndman & Koehler, 2006). Similar to TabArena (Erickson et al., 2025), we report ELO scores with confidence intervals through pairwise model comparisons based on $MASE$ scores per round. In difference to TabArena, where models are compared per dataset, in TS-Arena models are competing per challenge round. For models m_i and m_j, comparison in round r occurs only if both participated (MASEs $P_{i,r}$ and $P_{j,r}$ are valid). The outcome $S_{i,j,r}$ is:

$$S_{i,j,r} = \begin{cases} 1.0 & \text{if } P_{i,r} < P_{j,r} \quad \text{(Win)} \\ 0.0 & \text{if } P_{i,r} > P_{j,r} \quad \text{(Loss)} \end{cases} \tag{1}$$

Since ELO ratings exhibit path dependency, we employ *Randomized Replay Bootstrap* ($B = 500$) following TabArena (Erickson et al., 2025). We permute the chronological order of rounds across B iterations, recalculating the ELO trajectory from scratch each time. The 95% confidence interval is derived from the 2.5th and 97.5th percentiles of these bootstrap iterations. ELO scores are reported by challenge, frequency, and horizon, alongside mean $MASE$ with standard deviation. The current implementation focuses on point estimates, but the architecture supports future integration of probabilistic metrics such as $CRPS$.

A.5 Dashboard API & Front-End

The Dashboard API provides detailed and aggregated information for all time series and challenges, feeding the front-end as already shown in Figure 2. The interface includes aggregated leaderboards, live forecast views, and specific pages for challenges and models, with filtering capabilities enabling researchers to identify relevant models for their domain.

A.6 Reference Model Service

The Reference Model Service ensures consistent baseline participation in every challenge round, interacting with the public API Portal under identical constraints as external participants. We implemented the models like a practitioner by following official implementation suggestions:

- **Official Codebases:** We utilize official repositories provided by model authors.
- **Zero-Shot Configurations:** TSFMs are executed using author-recommended default settings.

We select TSFMs from peer-reviewed literature with available checkpoints and/or demonstrated strong performance in other benchmarks, including recent and successor models. Still, this initial set will be extended by integration of new models and hopefully external participants.

B Snapshot 2025-12-31 - Detailed Results

Table 2 summarizes the global ELO ratings and $MASE$ at the end of 2025 for all evaluated TSFMs and baseline methods, reporting both overall results and performance across specific frequencies and horizons.

Model	Architecture	Review	Published	Source
TTMs[*1] R1	Mixed	1	08.01.24	(Ekambaram et al., 2024)
Moirai	Encoder-only	1	04.02.24	(Woo et al., 2024)
Time-MoE	Decoder-only	1	24.09.24	(Shi et al., 2024)
TTMs[*1] R2	Mixed	1	08.10.24	(Ekambaram et al., 2024)
Chronos Bolt	Encoder-decoder	0	26.11.24	(Ansari et al., 2024b)
TimesFM 2.0	Decoder-only	0	20.12.24	(Das et al., 2024)
Sundial	Decoder-only	1	02.02.25	(Liu et al., 2025b)
TiRex	Encoder-only	1	29.05.25	(Auer et al., 2025)
VisionTS++	Vision	0	06.08.25	(Shen et al., 2025)
Flowstate	Encoder-decoder	0	07.08.25	(Graf et al., 2025)
TimesFM 2.5	Decoder-only	0	15.09.25	(Das et al., 2024)
Chronos-2	Encoder-only	0	17.10.25	(Ansari et al., 2025)
Moirai 2	Encoder-only	0	12.11.25	(Liu et al., 2025a)

Table 1: Reference TSFMs included in TS-Arena

[*1] TTMs = TinyTimeMixers

		Global		**15 min / 1 day**		**1 h / 3 days**	
Model	Rounds	ELO$^{(CI)}$	MASE$_{(\pm std)}$	ELO$^{(CI)}$	MASE$_{(\pm std)}$	ELO$^{(CI)}$	MASE$_{(\pm std)}$
chronos-2	1488	$\mathbf{1289^{+82}_{-101}}$	$0.732_{\pm 0.88}$	$\mathbf{1288^{+82}_{-92}}$	$0.706_{\pm 1.17}$	$\mathbf{1295^{+90}_{-99}}$	$0.751_{\pm 0.58}$
tirex	3319	1270^{+73}_{-86}	$\mathbf{0.682}_{\pm 0.71}$	1277^{+76}_{-72}	$\mathbf{0.685}_{\pm 0.89}$	1273^{+86}_{-92}	$\mathbf{0.680}_{\pm 0.48}$
moirai-2-small	1150	1263^{+87}_{-96}	$0.767_{\pm 1.01}$	1261^{+89}_{-97}	$0.734_{\pm 1.46}$	1270^{+78}_{-92}	$0.788_{\pm 0.58}$
timesfm-2.5-200m	1902	1240^{+86}_{-102}	$0.747_{\pm 0.92}$	1282^{+79}_{-87}	$0.721_{\pm 1.20}$	1216^{+88}_{-96}	$0.768_{\pm 0.61}$
chronos-bolt-base	5090	1214^{+94}_{-88}	$0.723_{\pm 0.98}$	1262^{+70}_{-95}	$0.727_{\pm 1.19}$	1180^{+97}_{-87}	$0.718_{\pm 0.72}$
flowstate	2411	1214^{+89}_{-102}	$0.710_{\pm 0.92}$	1241^{+93}_{-103}	$0.696_{\pm 1.20}$	1203^{+98}_{-96}	$0.722_{\pm 0.58}$
chronos-bolt-small	5090	1206^{+86}_{-97}	$0.726_{\pm 0.97}$	1233^{+86}_{-86}	$0.739_{\pm 1.20}$	1183^{+87}_{-89}	$0.713_{\pm 0.67}$
chronos-bolt-mini	5090	1200^{+82}_{-106}	$0.724_{\pm 0.92}$	1212^{+86}_{-92}	$0.742_{\pm 1.12}$	1187^{+71}_{-86}	$0.705_{\pm 0.67}$
chronos-bolt-tiny	5090	1194^{+74}_{-91}	$0.722_{\pm 0.89}$	1197^{+96}_{-86}	$0.746_{\pm 1.10}$	1192^{+78}_{-98}	$0.699_{\pm 0.60}$
timesfm-2.0-500m	5088	1178^{+95}_{-89}	$0.737_{\pm 1.00}$	1168^{+86}_{-82}	$0.779_{\pm 1.31}$	1197^{+78}_{-98}	$0.694_{\pm 0.54}$
sundial-128m	4828	1154^{+90}_{-80}	$0.736_{\pm 0.90}$	1195^{+78}_{-72}	$0.753_{\pm 1.14}$	1118^{+83}_{-85}	$0.720_{\pm 0.57}$
moirai-base-model	5090	1113^{+73}_{-84}	$0.753_{\pm 0.56}$	1066^{+96}_{-93}	$0.796_{\pm 0.61}$	1162^{+79}_{-87}	$0.709_{\pm 0.51}$
moirai-large	5090	1099^{+88}_{-92}	$0.774_{\pm 0.72}$	1021^{+107}_{-89}	$0.838_{\pm 0.88}$	1167^{+75}_{-91}	$0.710_{\pm 0.52}$
tinytimemixer-r2-1024-96	4890	1046^{+99}_{-77}	$0.790_{\pm 1.08}$	1078^{+89}_{-84}	$0.819_{\pm 1.35}$	1014^{+99}_{-85}	$0.760_{\pm 0.70}$
tinytimemixer-r1-1024-96	5088	1042^{+87}_{-87}	$0.785_{\pm 1.00}$	1053^{+92}_{-78}	$0.817_{\pm 1.26}$	1034^{+95}_{-89}	$0.753_{\pm 0.64}$
time-moe-50m	5089	1035^{+97}_{-90}	$0.823_{\pm 1.67}$	1011^{+81}_{-79}	$0.896_{\pm 2.25}$	1058^{+100}_{-97}	$0.749_{\pm 0.72}$
time-moe-200m	5088	1033^{+96}_{-85}	$0.824_{\pm 1.62}$	1003^{+80}_{-83}	$0.896_{\pm 2.16}$	1063^{+93}_{-89}	$0.752_{\pm 0.75}$
moirai-small	5090	1031^{+98}_{-83}	$0.788_{\pm 0.61}$	1005^{+93}_{-96}	$0.837_{\pm 0.71}$	1054^{+94}_{-84}	$0.740_{\pm 0.47}$
seasonal-naive	5105	922^{+117}_{-112}	$0.975_{\pm 1.30}$	913^{+118}_{-108}	$1.046_{\pm 1.70}$	924^{+128}_{-126}	$0.903_{\pm 0.70}$
seasonal-average	5101	866^{+114}_{-108}	$1.086_{\pm 2.56}$	822^{+119}_{-100}	$1.245_{\pm 3.49}$	903^{+112}_{-110}	$0.927_{\pm 0.93}$
visiontspp-base	2422	862^{+103}_{-84}	$1.026_{\pm 1.27}$	893^{+89}_{-84}	$0.978_{\pm 1.66}$	820^{+111}_{-83}	$1.068_{\pm 0.77}$
visiontspp-large	2422	797^{+99}_{-79}	$1.101_{\pm 1.58}$	825^{+91}_{-78}	$1.096_{\pm 2.10}$	765^{+99}_{-77}	$1.105_{\pm 0.93}$
simple-moving-average	5089	609^{+96}_{-74}	$1.431_{\pm 4.96}$	594^{+101}_{-73}	$1.606_{\pm 6.88}$	610^{+103}_{-83}	$1.256_{\pm 1.32}$

Table 2: Reference TSFM and Statistical Baseline MASEs and Elo-Scores in the Backtesting Year 2025

`chronos-2` leads with a global ELO of 1289^{+82}_{-101}, followed by `tirex` (1270^{+73}_{-86}) and `moirai-2-small` (1263^{+87}_{-96}). Differences in confidence intervals reflect the number of completed rounds: `tirex` (3319 rounds) shows tighter bounds ($+73/-86$) than `chronos-2` (1488 rounds, $+82/-101$). Since the top intervals overlap, their performances are statistically similar and further evaluations are required.

Rankings vary slightly by resolution. In the high-frequency setting (15 min / 1 day), `chronos-2` leads (1288), ahead of `timesfm-2.5-200m` (1282) and `tirex` (1277). In the lower-frequency setting (1 h / 3 days), `chronos-2` again ranks first (1295), followed by `tirex` (1273) and `moirai-2-small` (1270).

Scaling effects are evident within model families. In the `chronos-bolt` series, ELO increases monotonically with size: `tiny` (1194) < `mini` (1200) < `small` (1206) < `base` (1214). Newer versions also outperform earlier ones: `moirai-2-small` (1263) surpasses

`moirai-base-model` (1113) and `moirai-large` (1099), and `timesfm-2.5-200m` (1240) exceeds `timesfm-2.0-500m` (1178).

ELO rankings broadly align with $MASE$, though not perfectly. `tirex` attains the lowest global MASE (0.682) despite ranking second in ELO, suggesting more consistent accuracy. Larger models tend to win more often but also exhibit greater variability. `sundial-128m` performs competitively (ELO 1154, MASE 0.736), outperforming `moirai-base-model` (ELO 1113, MASE 0.753).