IMPROVING CONDITIONAL COVERAGE IN TIME-SERIES FOUNDATION MODELS VIA DIRECT VOLATILITY MODELING

Jason B. Cho & David S. Matteson
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14850, USA
{bc454,dm484}@cornell.edu

ABSTRACT

Time-series foundation models (TSFMs) provide probabilistic forecasts and are often reported to be well calibrated when evaluated via marginal coverage. However, forecast errors in time series are typically serially dependent and heteroskedastic, so marginal calibration can mask substantial regime-dependent miscalibration. We show empirically that prediction intervals from several TSFMs systematically under-cover during high-volatility periods, despite achieving near-nominal coverage on average. We propose a post-hoc multiplicative volatility correction that uses the uncertainty quantification produced by the TSFM as a baseline scale and dynamically adjusts it through a GARCH model. We evaluate this proposed method against adaptive conformal prediction as a benchmark. Across four real-world datasets, the proposed correction yields markedly improved conditional calibration over both the native TSFM intervals and adaptive conformal prediction. These findings highlight the importance of treating predictive variance as a dynamic process in TSFM uncertainty quantification.

Track: Research

1 INTRODUCTION

Time-Series Foundation Models (TSFMs) are large-scale architectures trained on broad and diverse collections of time-series data, designed to support zero-shot forecasting across a wide range of domains (Bommasani et al., 2022). Prominent examples include Chronos (Ansari et al., 2025), TimesFM (Das et al., 2024), Lag-Llama (Rasul et al., 2024), and Moirai (Liu et al., 2025). These models have demonstrated near state-of-the-art predictive accuracy across a variety of forecasting tasks. Owing to their versatility, they are increasingly being evaluated in applications such as finance (Chen et al., 2025; Goel et al., 2025), energy forecasting (Ferdaus et al., 2026; Meyer et al., 2024), demand forecasting (Yang et al., 2025), and climate and Earth system modeling (Bodnar et al., 2024).

Despite strong predictive performance, the behavior of TSFM uncertainty quantification remains comparatively underexplored. TSFMs are inherently probabilistic: they produce predictive quantiles or allow sampling from a predictive distribution, thereby providing uncertainty estimates alongside their forecasts. However, existing evaluations typically focus either on predictive accuracy or on marginal coverage, that is, whether prediction intervals achieve nominal coverage on average across time. Time-series data, by contrast, often exhibit substantial temporal dependence not only in the conditional mean but also in its variance. A well-known manifestation of such second-order dependence is volatility clustering in financial time series (Mandelbrot, 1963), where periods of high volatility tend to be followed by further high volatility. As a result, marginal calibration can mask substantial regime-dependent miscalibration. This suggests that predictive variance should not be treated as a static byproduct of the forecasting mechanism, but rather as a dynamic quantity that may itself require explicit modeling. In this paper, we propose a simple post-hoc volatility correction that uses the uncertainty produced by a TSFM as a baseline and dynamically adjusts it via a Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model Bollerslev (1986).

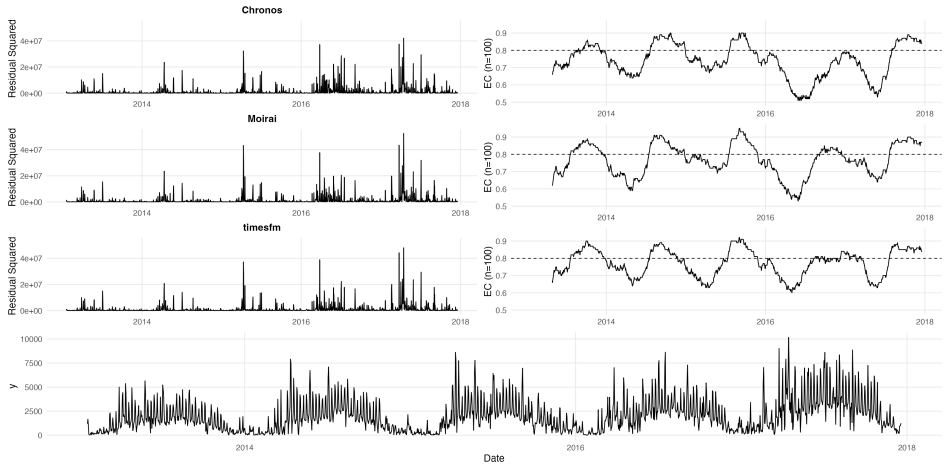

Figure 1: Daily bike rental counts from Capital Bikeshare (Bikeshare, 2024) (bottom), squared forecast residuals from three time-series foundation models: Chronos, Moirai 2, and TimesFM 2.5 (left column), and the empirical coverage of the models' prediction intervals, computed over 100 consecutive forecasts. The nominal 80% coverage level is indicated by the horizontal dashed line (right column).

The two main contributions of this paper are as follows:

1. **Conditional miscalibration of TSFMs:** We show that although several TSFMs achieve nominal marginal coverage on average, their forecast errors exhibit substantial second-order temporal dependence. As a result, prediction intervals can be severely miscalibrated conditionally, particularly during periods of elevated volatility.
2. **TSFM uncertainty correction with GARCH:** We propose a post-hoc multiplicative volatility correction that uses the uncertainty quantification already produced by the TSFM as a baseline scale and dynamically adjusts it through a GARCH model. We compare the proposed method with the uncertainty quantification produced by TSFMs and with adaptive conformal prediction following Gibbs & Candes (2021).

2 Related Work

Work that explicitly studies uncertainty quantification of TSFMs remains limited. The closest related study we are aware of is Achour et al. (2025), which combines TSFMs with conformal prediction to improve marginal coverage. While this approach enhances distribution-free coverage guarantees, evaluation is still performed in terms of coverage averaged over time and does not address temporal dependence in forecast errors. In parallel, TSFMs have been applied in finance for volatility forecasting. For example, Goel et al. (2025) use TSFMs with incremental fine-tuning to predict realized volatility of asset returns. However, these works focus on modeling the volatility of the underlying data-generating process. In contrast, our interest lies in modeling the variance of the model's forecast errors, that is, the conditional dispersion of prediction deviations.

3 Motivation: The Conditional Coverage Gap

To motivate our analysis, consider the bike rental data from Capital Bikeshare (Bikeshare, 2024), together with rolling 16-step-ahead forecasts and their associated 80% prediction intervals from three TSFMs, Chronos (Ansari et al., 2025), Moirai 2 (Liu et al., 2025), and TimesFM 2.5 (Das et al., 2024) (Figure 1). The observed series displays strong seasonal structure, with markedly higher volatility during the summer months when demand peaks. The left panels of Figure 1 show that squared forecast residuals cluster during these same periods, indicating that prediction errors increase systematically in high-demand regimes. In other words, forecast errors are not uniformly

distributed over time but are conditionally larger when the volatility of the series itself is elevated. As a consequence, empirical coverage is also serially dependent. The right panels of Figure 1 show the empirical coverage of the nominal 80% prediction intervals, which also exhibits a seasonal fluctuation and drops below 60% in certain periods. This example highlights the need to examine the conditional distribution of forecast errors rather than relying solely on marginal coverage.

4 METHODS

Our goal is to improve local adaptivity in predictive uncertainty quantification. Because TSFM forecast errors often exhibit time-varying second-order dependence, static interval adjustments based only on marginal coverage can be inadequate. Our proposed method is a multiplicative volatility correction based on GARCH (Bollerslev, 1986), which explicitly models residual scale dynamics while using the uncertainty quantification produced by the TSFM as a baseline scale. As a benchmark, we use adaptive conformal prediction (Gibbs & Candes, 2021), which recalibrates intervals using conformity scores derived from past forecast residuals. We evaluate both approaches against the native prediction intervals returned by the TSFMs.

To formalize this setup, consider the observation equation for a time series of length T, $\{y_t\}_{t=1}^{T}$

$$y_t = f(t) + v(t) r_t, \tag{1}$$

where $f(t)$ denotes the location component, $v(t)$ denotes the scale component, and r_t is a mean-zero residual process.

For any ordered index set $D \subseteq \{1, \ldots, T\}$, interpreted as a context window, the point forecast produced by a TSFM using $\{y_s : s \in D\}$ may be viewed as an estimator of the location component $f(t)$ at time t. We denote this estimator by $\hat{f}_D(t)$. Similarly, we let $\hat{v}_D(t)$ denote the standard deviation, or volatility, at time t implied by the model's uncertainty quantification. While some TSFMs provide the full predictive distribution, allowing $\hat{v}_D(t)$ to be estimated directly from predictive samples, many lightweight models return only a limited number of predictive quantiles. In such cases, $\hat{v}_D(t)$ can be approximated under a Gaussian assumption as

$$\hat{v}_D(t) = \frac{\hat{Q}^D_{1-\alpha/2}(t) - \hat{Q}^D_{\alpha/2}(t)}{2\Phi^{-1}(1-\alpha/2)}, \tag{2}$$

where $\Phi^{-1}(\cdot)$ is the standard normal quantile function, and $\hat{Q}^D_{\alpha}(t)$ denotes the predictive α-quantile returned by the TSFM at time t given context window D.

Multiplicative Volatility Correction with GARCH Conditional miscalibration of the prediction intervals produced by TSFMs, as shown in Section 3, suggests that the model does not fully capture predictive uncertainty. Even after accounting for the model-implied uncertainty quantification $\hat{v}_D(t)$, the residual process r_t may still exhibit time-varying second-order dependence. To model this remaining structure, we assume that r_t follows a GARCH(p, q) process (Bollerslev, 1986):

$$r_t = \sigma_t \varepsilon_t, \tag{3}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i r_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \tag{4}$$

where σ_t^2 denotes the conditional variance of r_t given past information. The parameter $\omega > 0$ determines the long-run variance level, $\alpha_i \geq 0$ captures the short-term impact of past shocks r_{t-i}^2, and $\beta_j \geq 0$ governs the persistence of volatility through past conditional variances. Large values of $\sum_i \alpha_i + \sum_j \beta_j$ indicate strong persistence in volatility. In our implementation, we use GARCH(1,1) because it provides a parsimonious yet empirically effective specification. In a large comparative study, Hansen & Lunde (2005) found that GARCH(1,1) was not significantly outperformed by more elaborate volatility models. More generally, the orders p and q may be selected using an information criterion such as AIC or BIC (Brooks & Burke, 2003).

The fitted GARCH model yields forecast quantiles for the residual process, which can then be mapped back to the original scale using the TSFM point forecast and implied volatility. Accordingly, the corrected $(1-\alpha)$-level prediction interval for y_{t+1} is defined as

$$\hat{C}^{\mathrm{G}}_D(t+1) = [\hat{f}_D(t+1) + \hat{v}_D(t+1)\, \hat{Q}^{r,D}_{\alpha/2}(t+1), \hat{f}_D(t+1) + \hat{v}_D(t+1)\, \hat{Q}^{r,D}_{1-\alpha/2}(t+1)], \tag{5}$$

where $\hat{Q}_{\alpha}^{r,D}(t+1)$ denotes the forecasted α-quantile of the residual process r_{t+1} implied by the fitted GARCH model under context window D. Unlike conformal score-based methods, which construct prediction intervals solely from past forecast residuals, our approach explicitly incorporates the uncertainty quantification already produced by the TSFM and uses it as a baseline scale for dynamic correction.

Adaptive Conformal Prediction As a benchmark for the proposed method, we consider adaptive conformal prediction (Gibbs & Candes, 2021). Adaptive conformal prediction extends split conformal prediction (Angelopoulos & Bates, 2022) to accommodate distributional shift. While classical split conformal uses a fixed miscoverage level α, adaptive conformal instead uses a time-varying level α_t that is updated online based on recent coverage performance.

Let $D_1, D_2 \subseteq \{1, \ldots, T\}$ denote the training and calibration index sets. Conformity scores are defined as

$$R_s = |y_s - \hat{f}_{D_1}(s)|, \quad s \in D_2.$$

Let $\hat{q}_t$ be the empirical $(1-\alpha_t)$-quantile of $\{R_s : s \in D_2\}$. The adaptive conformal prediction interval for y_{t+1} is

$$\hat{C}_{D_1}^{\mathrm{ACI}}(t+1) = \big[\hat{f}_{D_1}(t+1) - \hat{q}_t,\ \hat{f}_{D_1}(t+1) + \hat{q}_t\big].$$

After observing y_{t+1}, the miscoverage indicator is defined as

$$I_{t+1} = \mathbf{1}\{y_{t+1} \notin \hat{C}_{D_1}^{ACI}(t+1)\}.$$

The level α_t is then updated via

$$\alpha_{t+1} = \alpha_t + \gamma\big(\alpha - I_{t+1}\big),$$

where $\gamma > 0$ is a step-size parameter and α is the target miscoverage rate. Following Gibbs & Candes (2021), we set the step-size parameter to $\gamma = 0.005$, which was found to provide a reasonable balance between stability and adaptability under distribution shift. Intuitively, if recent coverage falls below the nominal level, α_t decreases, widening future intervals; if coverage is overly conservative, α_t increases, shrinking them. This online adjustment aims to maintain approximate marginal coverage under drifting distributions while preserving the distribution-free nature of conformal prediction.

5 RESULTS

	Chronos			Moirai			TimesFM		
Metric	Vanilla	A.Conformal	GARCH	Vanilla	A.Conformal	GARCH	Vanilla	A.Conformal	GARCH
Bike Rental Counts									
MC	0.743	0.780	**0.813**	0.773	**0.792**	0.811	0.775	**0.795**	0.816
RMSCE	0.118	0.135	**0.060**	0.102	0.127	**0.069**	0.087	0.114	**0.066**
MaxUC	0.311	0.367	**0.156**	0.278	0.322	**0.156**	0.211	0.289	**0.111**
Colorado River Streamflow									
MC	0.775	0.767	**0.787**	0.729	0.774	**0.800**	0.802	0.781	**0.799**
RMSCE	0.122	0.213	**0.106**	0.174	0.212	**0.127**	0.134	0.199	**0.112**
MaxUC	0.433	0.633	**0.378**	0.467	0.600	**0.344**	0.411	0.578	**0.300**
Electricity Demand									
MC	0.761	**0.790**	0.816	**0.800**	0.784	0.818	0.643	0.775	**0.808**
RMSCE	0.087	0.143	**0.083**	0.085	0.150	**0.080**	0.187	0.147	**0.072**
MaxUC	0.244	0.356	**0.189**	0.244	0.411	**0.178**	0.411	0.389	**0.189**
NO2 Concentration									
MC	0.816	0.754	**0.784**	0.820	0.757	**0.791**	0.783	0.748	**0.788**
RMSCE	0.069	0.171	**0.053**	0.070	0.168	**0.063**	0.077	0.179	**0.057**
MaxUC	**0.144**	0.400	0.167	0.222	0.411	**0.178**	0.222	0.422	**0.189**

Table 1: Marginal Coverage (MC), RMSCE (Root Mean Squared Conditional Error), and Maximum undercoverage (MaxUC) A.2 across four datasets A.1. Vanilla refers to the original prediction intervals and A.Conformal refers to the Adaptive Conformal Prediction by Gibbs & Candes (2021). Bold values indicate the best method for each model-dataset pair.

We consider four time-series datasets: bike rental counts in the Washington metropolitan area via Capital Bikeshare (Bikeshare, 2024), electricity demand in New York State from the New York Independent System Operator (NYISO)(EIA, 2024), NO_2 concentration in Pasadena, CA (EPA AQS site 06-037-2005)(U.S. Environmental Protection Agency, 2026), and streamflow (discharge, cubic feet per second) of the Colorado River near Cisco, UT (USGS NWIS site 09180500, parameter 00060) (U.S. Geological Survey, 2026). The corresponding time-series plots are provided in Appendix A.1, Figure 2.

Comparisons among uncertainty quantification methods are made across three metrics. Marginal Coverage (MC) is defined as the average proportion of observations falling within the nominal 80% prediction interval ($\alpha = 0.2$). RMSCE (Root Mean Squared Conditional Error) and MaxUC quantify conditional calibration by measuring, respectively, the average root mean squared deviation of rolling coverage from the nominal level and the maximum shortfall of rolling coverage below the nominal level. These latter metrics capture regime-dependent miscalibration that may be obscured by marginal aggregation. The precise mathematical definitions are provided in Appendix A.2.

Table 1 evaluates our proposed GARCH-based correction against both the native TSFM intervals and adaptive conformal prediction across four datasets. The Vanilla TSFM intervals exhibit reasonable MC overall, with coverage typically within 5% points of the nominal level. However, there are notable failures, most prominently for Electricity Demand under TimesFM, where MC drops to 64%. In addition, conditional calibration remains problematic. RMSCE ranges between 0.07 and 0.17, with smaller deviations observed for datasets exhibiting milder volatility dynamics, such as Electricity Demand and NO_2 concentration. MaxUC is consistently substantial, ranging from 0.144 to 0.47, indicating pronounced regime-dependent undercoverage even when marginal coverage appears adequate.

Adaptive conformal prediction generally achieves performance comparable to Vanilla on MC. However, its conditional calibration remains unstable. Both RMSCE and MaxUC remain relatively large across datasets and are typically worse than those obtained via GARCH adjustment. This suggests that recalibrating quantiles alone does not sufficiently address persistent second-order temporal dependence, particularly in datasets with pronounced volatility regime shifts.

In contrast, the GARCH-based adjustment consistently delivers the most stable performance. MC remains within 2% points of the nominal level across datasets. GARCH acheives RMSCE below 0.12 and clearly outperforms both Vanilla and adaptive conformal methods. MaxUC is substantially reduced in most datasets, with the exception of NO_2 concentration, where Vanilla (0.144) and GARCH (0.167) perform comparably. Overall, explicitly modeling predictive variance as a dynamic process yields markedly improved conditional calibration, particularly in datasets exhibiting strong and persistent volatility regimes.

6 Conclusion

Our results demonstrate that conditional miscalibration is a systematic and practically relevant phenomenon in time-series foundation models. Although TSFMs often achieve nominal marginal coverage, their prediction intervals can substantially underperform during high-volatility regimes, revealing strong regime-dependent behavior. Our analysis suggests that marginal evaluation alone therefore provides an incomplete picture of a model's uncertainty quantification.

We show that directly modeling the variance process leads to meaningful improvements in conditional calibration. Our method likely outperforms conformal prediction for two reasons. First, the forecast errors exhibit serial dependence in their second-order structure, making a GARCH-type model a useful approximation to the residual volatility dynamics. Second, our approach uses the implied volatility extracted from the TSFM prediction intervals, whereas conformal methods rely only on the point forecast and past residuals. This allows our method to exploit additional uncertainty information already contained in the TSFM output. These findings suggest that second-order dynamics are not fully captured by current TSFM uncertainty mechanisms. Future work may consider more flexible volatility models, including deep learning-based approaches, that can be integrated more closely with foundation forecasting architectures.

ACKNOWLEDGMENTS

Financial support from National Science Foundation grant DMS-2114143 is gratefully acknowledged.

REFERENCES

Sami Achour, Yassine Bouher, Duong Nguyen, and Nicolas Chesneau. Foundation models for time series forecasting: Application in conformal prediction, July 2025. URL http://arxiv.org/abs/2507.08858. arXiv:2507.08858 [cs] version: 1.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

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, 2025. URL https://arxiv.org/abs/2510.15821.

Capital Bikeshare. Capital bikeshare system data, 2024. URL https://capitalbikeshare.com/system-data. Accessed: 2024-06-16.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Vaughan, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A Foundation Model for the Earth System, November 2024. URL http://arxiv.org/abs/2405.13063. arXiv:2405.13063 [physics].

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(86)90063-1.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

Chris Brooks and Simon Burke. Information criteria for garch model selection. *The European Journal of Finance*, 9(6):557–580, None 2003. doi: 10.1080/1351847021000029188.

Liyuan Chen, Shuoling Liu, Jiangpeng Yan, Xiaoyu Wang, Henglin Liu, Chuang Li, Kecheng Jiao, Jixuan Ying, Yang Veronica Liu, Qiang Yang, and Xiu Li. Advancing Financial Engineering with Foundation Models: Progress, Applications, and Challenges. *Engineering*, pp.

S209580992500757X, December 2025. ISSN 20958099. doi: 10.1016/j.eng.2025.11.029. URL http://arxiv.org/abs/2507.18577. arXiv:2507.18577 [q-fin].

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

U.S. Energy Information Administration EIA. New York independent system operator electricity demand data, 2024. URL https://www.eia.gov/electricity/gridmonitor/dashboard/electric_overview/regional/REG-NY. Accessed: 2024-06-16.

Md Meftahul Ferdaus, Tanmoy Dam, Md Rasel Sarkar, Moslem Uddin, and Sreenatha G. Anavatti. Foundation models for clean energy forecasting: A comprehensive review. *Renewable and Sustainable Energy Reviews*, 226:116452, January 2026. ISSN 1364-0321. doi: 10.1016/j.rser.2025.116452. URL https://www.sciencedirect.com/science/article/pii/S1364032125011256.

Isaac Gibbs and Emmanuel Candes. Adaptive Conformal Inference Under Distribution Shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672. Curran Associates, Inc., 2021. URL https://papers.neurips.cc/paper_files/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html.

Anubha Goel, Puneet Pasricha, Martin Magris, and Juho Kanniainen. Foundation Time-Series AI Model for Realized Volatility Forecasting, May 2025. URL http://arxiv.org/abs/2505.11163. arXiv:2505.11163 [q-fin] version: 1.

Peter R. Hansen and Asger Lunde. A forecast comparison of volatility models: does anything beat a garch(1,1)? *Journal of Applied Econometrics*, 20(7):873–889, 2005. doi: https://doi.org/10.1002/jae.800.

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, 2025. URL https://arxiv.org/abs/2511.11698.

Benoit Mandelbrot. The Variation of Certain Speculative Prices. *The Journal of Business*, 36(4): 394–419, 1963. ISSN 0021-9398. URL https://www.jstor.org/stable/2350970.

Marcel Meyer, David Zapata, Sascha Kaltenpoth, and Oliver Müller. Benchmarking Time Series Foundation Models for Short-Term Household Electricity Load Forecasting, October 2024. URL https://arxiv.org/abs/2410.09487v3.

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

U.S. Environmental Protection Agency. Air quality system (aqs) data mart. https://aqs.epa.gov/aqsweb/airdata/download_files.html, 2026. Accessed February 2026.

U.S. Geological Survey. National water information system data available on the world wide web (usgs water data for the nation). https://waterdata.usgs.gov/, 2026. Accessed February 2026.

Wei Yang, Defu Cao, and Yan Liu. Foundation Models for Demand Forecasting via Dual-Strategy Ensembling, July 2025. URL https://arxiv.org/abs/2507.22053v1.

A APPENDIX

A.1 ADDITIONAL FIGURES

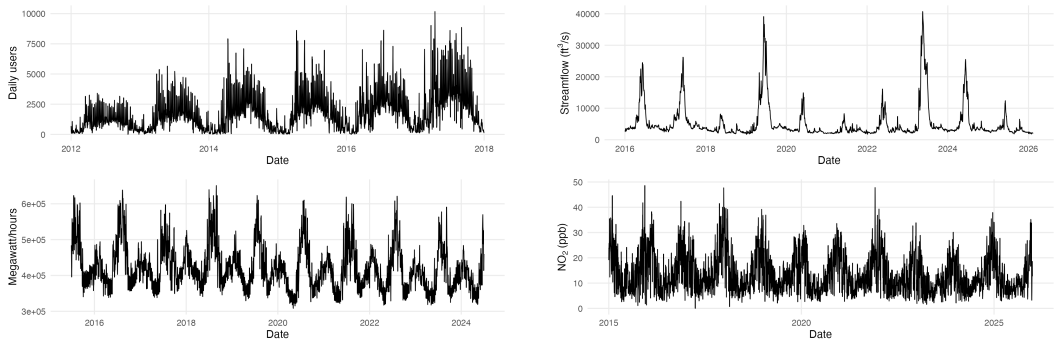

Figure 2: Four daily time-series datasets are considered: bike rental counts in the Washington metropolitan area via Capital Bikeshare (Bikeshare, 2024), electricity demand in New York State from the New York Independent System Operator (NYISO)(EIA, 2024), NO_2 concentration in Pasadena, CA (EPA AQS site 06-037-2005)(U.S. Environmental Protection Agency, 2026), and streamflow (discharge, cubic feet per second) of the Colorado River near Cisco, UT (USGS NWIS site 09180500, parameter 00060) (U.S. Geological Survey, 2026). A common feature across these datasets is pronounced time-varying volatility, with distinct regimes in which the variability of the underlying process changes substantially over time. Such heteroskedastic behavior makes certain periods intrinsically more difficult to predict than others.

A.2 EVALUATION METRICS

Let $\{(y_t, \hat{C}_t)\}_{t=1}^{T}$ denote the observed outcomes and associated prediction intervals, where $\hat{C}_t = [\hat{\ell}_t, \hat{u}_t]$ is the $(1-\alpha)$ prediction interval at time t. Define the coverage indicator: $I_t = \mathbf{1}\{y_t \in \hat{C}_t\}$.

Marginal Coverage (MC). Marginal coverage is defined as

$$\mathrm{MC} = \frac{1}{T}\sum_{t=1}^{T} I_t.$$

MC measures average coverage over the entire sample. A method is marginally calibrated if $\mathrm{MC} \approx 1-\alpha$. However, this metric does not account for temporal variation in coverage.

Root Mean Squared Conditional Error (RMSCE) To assess conditional calibration, we compute rolling coverage over windows of length w,

$$\widehat{\mathrm{MC}}_t^{(w)} = \frac{1}{w}\sum_{s=t-w+1}^{t} I_s.$$

We then define

$$\mathrm{RMSCE} = \sqrt{\frac{1}{T_w}\sum_{t}\left(\widehat{\mathrm{MC}}_t^{(w)} - (1-\alpha)\right)^2},$$

where T_w denotes the number of valid rolling windows. RMSCE measures the root mean squared deviation of local coverage from the nominal level. By squaring deviations before averaging, it penalizes larger conditional miscalibration more heavily and captures persistent under- or over-coverage across volatility regimes.

Maximum Under Coverage (MaxUC). To quantify worst-case conditional undercoverage, we define

$$\mathrm{MaxUC} = \max_t \left\{ (1-\alpha) - \widehat{\mathrm{MC}}_t^{(w)} \right\}.$$

MaxUC measures the largest shortfall of rolling coverage below the nominal level and highlights periods of severe conditional undercoverage.