Calibration Properties of Time Series Foundation Models: An Empirical Analysis

Anonymous Authors[1]

Abstract

The recent development of foundation models for time series data has generated considerable interest in using such models across a variety of applications. Although they achieve state-of-the-art predictive performance, the ability to produce well-calibrated probabilistic distributions is critical for practical applications and is relatively underexplored. In this paper we investigate the calibration-related properties of four recent time series foundation models and two competitive baselines. We perform systematic evaluations and identify significant variation in calibration performances across models.

1. Introduction

Time series modeling has applications across a broad range of fields including analysis and forecasting of data from climate science (Gharbi et al., 2011), consumer behavior (Makridakis et al., 2022; Adams, 1964), financial transactions (Krollner et al., 2010), and energy consumption (Godahewa et al., 2020; Peterson, 2017). Traditional statistical approaches such as linear auto-regressive (AR) models and associated variants are well-established in the field (Hamilton, 1994; Hyndman & Athanasopoulos, 2018). The toolbox of traditional models has been supplemented in recent years by a variety of machine learning approaches, e.g., deep learning time series models such as N-BEATS (Oreshkin et al., 2019) and Informer (Zhou et al., 2021).

A more recent trend, as an alternative to these earlier approaches, is the emergence of time series foundation models (Nie et al., 2022; Yeh et al., 2023), leveraging transformer-based representations. Unlike traditional statistical and machine learning approaches where models are fitted to a specific time series, foundation models are general-purpose models trained on a broad range of time series and are capable of zero-shot or few-shot forecasting on any time series in principle (Ye et al., 2024; Liang et al., 2024). This property makes them especially appealing to practitioners in that they require only a single global model (the foundation model) rather than retraining a new model for every time series (Benidis et al., 2022; Bommasani et al., 2021).

With this increase of interest in larger and more flexible foundation models for time series, it becomes important to understand and characterize the calibration properties of such models. In particular, for many applications, rather than just a single point forecast (e.g., the expected value of the time series at a future time), the availability of distributional information, such as conditional densities or conditional quantiles, can be essential for decision-making (Gneiting et al., 2007; Petropoulos et al., 2022), e.g., downstream tasks such as anomaly detection (Menon & Williamson, 2018).

Methods for evaluating the calibration properties of classification and regression models have received significant attention in machine learning in recent years (e.g., Guo et al. (2017); Song et al. (2019); Chung et al. (2021)). However, for time series foundation models there has been less attention to date on evaluating their probabilistic calibration properties. Common calibration metrics such as Weighted Quantile Loss (WQL), Continuous Ranked Probability Score (CRPS), and Mean Scaled Interval Score (MSIS) evaluate probabilistic forecasts as a combination of accuracy and calibration (e.g., see Chung et al. (2021)) rather than solely focusing on calibration properties directly. As a result, model performance on these metrics can be highly correlated with point-wise accuracy metrics such as Mean Absolute Scaled Error (MASE), potentially hiding weaknesses in a model's calibration performance.

To address this gap, in this paper we empirically evaluate time series foundation models and baselines with respect to calibration-specific metrics such as probabilistic calibration error (PCE): see Dheur & Taieb (2023), also used in Kuleshov et al. (2018); Chung et al. (2021). More specifically, we evaluate four state-of-the-art time series foundation models and two widely-used baseline methods, in terms of their zero-shot forecasts across a variety of univariate time series datasets. We investigate and compare

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

their calibration properties through a comprehensive set of metrics across, focusing on different aspects of conditional uncertainty in model predictions. Our main findings are:

- Time series foundation models generally have better calibration properties than baselines; however, some foundation models are poorly calibrated.
- Calibration properties of models tend to be consistent across datasets; models that are well (or poorly) calibrated on one dataset tend to also be well (or poorly) calibrated on other datasets.
- Calibration properties of foundation models tend to be robust across prediction horizons.

2. Methodology

2.1. Foundation Models and Baselines

Since we are primarily interested in exploring the calibration properties of time series foundation models, we identify four foundation models trained on time series data from scratch, either re-using large language model (LLM) architectures or proposing novel architectures, that are able to perform zero-shot forecasting and probabilistic predictions. To provide a better context, we also include two baseline models, AutoARIMA (Hyndman & Khandakar, 2008) and N-BEATS (Oreshkin et al., 2019) to represent parametric and neural baselines.

Although all four foundation models are transformer-based, Lag-Llama (Rasul et al., 2023) and Chronos (Ansari et al., 2024) adapted existing LLMs, while TimesFM (Das et al., 2024) and MOIRAI (Woo et al., 2024) proposed their transformer architectures. The main differences between these models that affect their calibration are (i) training objectives, (ii) sampling procedures, and (iii) how they compute quantiles that are then used to compute the metrics.

Specially, **TimesFM** builds pre-fixed quantile heads into training and jointly optimizes mean squared error (MSE) and quantile loss (Wen et al., 2017). TimesFM can produce joint predictions of multiple steps and perform forecasting on the same horizon with fewer auto-regressive steps than other models, but their quantile forecasting is limited to pre-determined quantile levels without further fine-tuning. **MOIRAI** and **Lag-Llama** maximizes the likelihood of a mixture of parametric distributions and a Student's t-distribution, respectively; both sample from the parametric distribution for one-step-ahead predictions. **Chronos** transforms real-valued time series into a fixed vocabulary via scaling and quantization, and is optimized for classifications of binned predictions. It samples one-step-ahead like other LLMs. Probabilistic quantities are computed by sampling multiple continuations of the trajectory. For baselines **N-BEATS** and **AutoARIMA**, we use the `neuralforecast` library (Olivares et al., 2022) that performs probabilistic forecasts analytically or samples from the distribution given the underlying assumption, depending on implementations.

2.2. Datasets

We selected datasets to capture a range of tasks differing in time-step granularity, seasonality, and forecasting difficulty, across a variety of domains. We believe that the datasets that we selected were not used in training of the foundation models used in our experiments (with the exception of the M5 data for the MOIRAI model).

We evaluate models on consumer behavior datasets with two temporal granularities: (i) a **Reviews** dataset consisting of hourly counts of Amazon product reviews (Hou et al., 2024) and Google Places reviews (Li et al., 2022), and (ii) a modified **M5** competition dataset (Makridakis et al., 2022) consisting of the daily number of products being sold at different locations.

Many of the datasets used in training time series foundation models are related to human behavior and natural phenomena that exhibit periodic trends (e.g., 24-hour effects). To analyze model performance on datasets that differ from the pre-training data, we included datasets for **Glucose Level** (Cho et al., 2023) and **Heart Rate** prediction (Peng et al., 1999). The former measures interstitial glucose concentration of 16 subjects every 5 minutes over the course of 10 days, and the latter records the heart rate of 14 volunteers every 4 seconds during a 10 minute metronomic breathing activity. As noted in Gu et al. (2025), the strong predictive accuracy of foundation models on many of the datasets used in machine learning evaluations does not necessarily translate into high predictive accuracy in applications such as vital sign forecasting in healthcare.

2.3. Notation and Metrics

In this section, we formalize the notation and define the comprehensive set of metrics for our analysis. Given a *context* of T observations for a time series, $y_{1:T}$, we are interested in evaluating the H-length forecasting performance of each model conditioned on the context, i.e., predictions related to $y_{T+1:T+H} \mid y_{1:T}$. We denote the median prediction as $\hat{y}_t^{0.5}$ at each $t \in \{T+1, ..., T+H\}$ for point estimates, and predicted quantiles $\hat{y}_t^q$ to assess uncertainty and calibration of these predictions, where $q \in Q := \{0.1, ..., 0.9\}$. The predicted quantiles are produced either directly or indirectly (e.g., via sampling) by each model. We consider all h-step ahead predictions for $h \in \{1, ..., H\}$ where $H = 48$ and average metrics over these horizons unless otherwise stated.

In addition to calibration-based metrics, we assess model accuracy using Mean Absolute Scaled Error (MASE) for the predicted median, scaling the Mean Absolute Error (MAE)

by a naive predictor (Hyndman & Athanasopoulos, 2018):

$$MASE = \frac{\frac{1}{H}\sum_{t=T+1}^{T+H} |\hat{y}_t^{0.5} - y_t|}{\frac{1}{H-1}\sum_{t=T+2}^{T+H} |y_t - y_{t-1}|}. \quad (1)$$

As for calibration, we compute metrics that are specifically designed for calibration, e.g.,Probabilistic Calibration Error (PCE) (Dheur & Taieb, 2023; Kuleshov et al., 2018):

$$PCE_p = \frac{1}{Q}\sum_{q\in Q}\left| q - \frac{1}{H}\sum_{t=T+1}^{H} \mathbf{1}[y_t \le \hat{y}_t^q]\right|^p. \quad (2)$$

We use PCE with $p = 1$ and denote PCE_1 as PCE for brevity. Intuitively, PCE measures the differences between empirical and predicted CDFs with lower values indicating better-calibrated models. PCE is lower-bounded by 0 and upper-bounded (in effect) by 0.5 for the case where predicted quantiles are always all above or below the observed value of y.

However, well-calibrated models are not sufficient to produce useful forecasts: one model could always predict the marginal distribution that does not depend on the inputs and, hence, is not informative. In this context a metric that specifically captures *sharpness* (the concentration of the predictive distributions) is also important in an overall evaluation of calibration (Gneiting et al., 2007; Kuleshov et al., 2018). A simple surrogate for sharpness is the width of a predicted confidence interval, e.g., where a 95%-confidence interval is the interval between the q_{low} =2.5% and q_{high} =97.5% quantile predictions. We refer to this as Scaled Interval Width (SIW) where s is the confidence associated with the interval (i.e., $s = 95\%$ in the preceeding example):

$$SIW_s = \frac{1}{H}\sum_{t=T+1}^{T+H} \frac{\hat{y}_t^{q_{\text{high}}} - \hat{y}_t^{q_{\text{low}}}}{y^{q_{\text{high}}} - y^{q_{\text{low}}}}. \quad (3)$$

Models with lower SIW values are more confident in their predictions, while models with larger SIW values are less confident, i.e., their predictions have more spread.

Finally, we are interested in whether a model is systematically miscalibrated in one direction or another. To quantify this, we define the metric Centered Calibration Error (CCE) that compares the amount of observed data in a predicted interval with the associated confidence s:

$$CCE = \frac{1}{S}\sum_{s\in S} s - \frac{1}{H}\sum_{t=T+1}^{T+H} \mathbf{1}\left[\hat{y}_t^{q_{\text{low}}} \le y_t \le \hat{y}_t^{q_{\text{high}}}\right]. \quad (4)$$

The direction of CCE for a model, combined with its SIW, can be used to identify if a model is being over- or under-confident. Positive CCE values indicate there is more observed data outside the predicted interval than expected by the confidence level; together with a low SIW value, we can infer that a model is over-confident. On the other hand, negative CCE values and larger SIW values imply that the model is under-confident.

3. Results

Based on the models, datasets, and metrics outlined above, in our experiments we found that foundation models typically exhibited competitive (but not necessarily better) point-forecasting performance compared to baselines, with TimesFM, Chronos, and MOIRAI often having a lower MASE than N-BEATS and AutoARIMA (see the x-axis of Figure 1). This did not hold for all foundation models: Lag-Llama typically had worse forecasts than the baselines. As visualized in Figure 3 in the Appendix, WQL and MSIS had high correlation with MASE, with the result that using WQL or MSIS alone could lead to missing key aspects of calibration performance for a model. For example, the WQL and MSIS did not show significant differences in calibration performance between the foundation models and the baselines in the M5 dataset (see Figure 6 in Appendix). In contrast, when evaluating calibration using the PCE metrics, we saw a clear difference in the foundation models' calibration compared to the baselines.

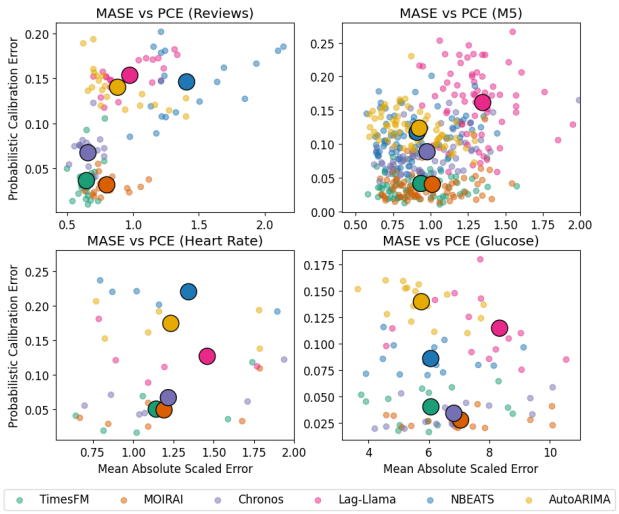

Figure 1. Probabilistic calibration error (PCE) versus point-forecast accuracy (MASE) across the four datasets. Each dot represents model performance on an individual time series; larger centroids being the average over all time series in a dataset.

3.1. Calibration Error

For point prediction (MASE), Figure 1 (x-axis) illustrates that the difference in prediction performance between foundation models and baselines can vary across datasets and is not always in favor of foundation models. In contrast, in terms of calibration (PCE), the variation in the y-axis in Figure 1 shows that the best foundation models tend to

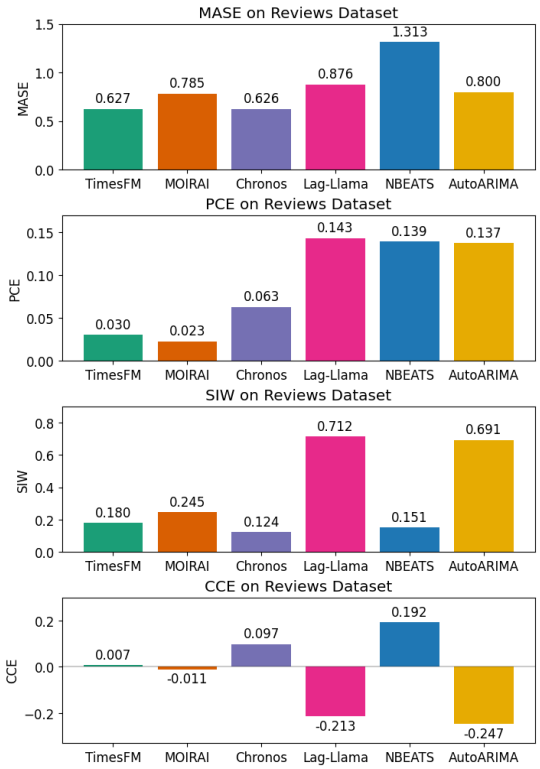

Figure 2. Model accuracy (top) and different calibration metrics (bottom 3 charts) for all models for the Reviews dataset.

have systematically better calibration performance. To provide a general sense of scale, PCE numbers below 0.05 can (loosely) be considered to be much better calibrated that numbers in the range of 0.15 to 0.2. TimesFM, MOIRAI, and Chronos had much lower PCE than the baselines (typically less than 0.05 across datasets) with Lag-Llama being more comparable to the baselines (between 0.1 and 0.2).

Although foundation model point-forecasting performance varied across datasets, the probabilistic calibration remained consistent. For example, despite all models performing poorly (in terms of point prediction/MASE) on the Glucose dataset (which is difficult to predict), their calibration performance was consistent, i.e., the well-calibrated models were still well-calibrated even if point predictions were inaccurate. If we had evaluated calibration using WQL or MSIS, we would have incorrectly concluded that N-BEATS and AutoARIMA were equally or better calibrated than the best foundation models on the Glucose dataset in Figure 6.

3.2. Confidence Calibration

Of the foundation models evaluated, Chronos had consistently the smallest mean SIW, indicative of narrower prediction intervals, i.e., higher confidence. Noting in addition that Chronos has large positive CCE on all the datasets, this indicates that Chronos is being consistently over-confident. Figure 2 provides the SIW and CCE values (as well as MASE and PCE) for the Reviews dataset for Chronos and the other 5 models. The overconfidence of Chronos was a systematic feature of the model across all 4 datasets (see Figure 4 in Appendix). TimesFM and MOIRAI were well-calibrated according to PCE, and this is reflected in the near-zero CCE values, often with a smaller SIW than the baselines. There was no clear trend with Lag-Llama's confidence calibration.

3.3. Forecast Horizon Length

For point-forecasting we found that MASE steadily increased with a larger prediction horizon (see Figure 7 in the Appendix). However, this trend does not hold up for calibration: Figure 8 (in Appendix) shows that the foundation models are robust across increasing prediction horizons, increasing only a couple of percent in PCE when comparing predictions from 1 time-step ahead to 48 steps. The Glucose dataset was particularly difficult to predict at longer forecast horizons (see Table 1), yet TimesFM and Lag-Llama only increased their PCE by 0.02 while MOIRAI and Chronos improved their PCE by 0.01 and 0.02 respectively. The baselines were less consistent in the change in calibration, with N-BEATS improving PCE by 0.17 and AutoARIMA a much smaller 0.02.

We include additional figures and findings in the Appendix, where we discuss the empirical correlation of WQL, MSIS, and MASE along with details on calibration error on the tail end of probabilistic forecasts.

4. Conclusions and Future Work

In summary, in this work we systematically evaluated the calibration properties of time series foundation models and found that the best-performing models, TimesFM and MOIRAI, were consistently well-calibrated across multiple datasets compared to baselines. However, not all foundation models were accurate or were well-calibrated, with Lag-Llama for example having both poor point-prediction accuracy and poor probabilistic calibration, and the Chronos model exhibiting significant over-confidence.

For the final workshop version of this paper, we plan a case study to assess the accuracy of all models (4 foundations models plus 2 baselines) on the downstream task of anomaly detection using the UCR Anomaly Detection Dataset (Wu & Keogh, 2021). Using model predictive uncertainty as an anomaly detection signal, we will investigate how calibration of the different models affects their ability to detect anomalies. We will also extend the results to additional datasets and additional models (both foundation models and baselines). Finally, we also plan to investigate the effect of training loss, decoupled from model architecture, on calibration performance, e.g., evaluating the calibration performance of baselines when trained with quantile loss.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and, as such, has the potential to have a positive impact on society. We are not aware of any significant negative societal consequences that would result from this work.

References

Adams, F. G. Consumer attitudes, buying plans, and purchases of durable goods: A principal components, time series approach. *The Review of Economics and Statistics*, pp. 347–355, 1964.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. Chronos: Learning the language of time series. *Transactions on Machine Learning*, November 2024.

Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., et al. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36, 2022.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Cho, P., Kim, J., Bent, B., and Dunn, J. Big ideas lab glycemic variability and wearable device data. *(version 1.1. 1). PhysioNet*, 2023.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:10971–10984, 2021.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Dheur, V. and Taieb, S. B. A large-scale study of probabilistic calibration in neural network regression. In *International Conference on Machine Learning*, pp. 7813–7836. PMLR, 2023.

Gharbi, M., Quenel, P., Gustave, J., Cassadou, S., Ruche, G. L., Girdary, L., and Marrama, L. Time series analysis of dengue incidence in guadeloupe, french west indies: forecasting models using climate variables as predictors. *BMC Infectious Diseases*, 11:1–13, 2011.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69 (2):243–268, 2007.

Godahewa, R., Bergmeir, C., Webb, G., Hyndman, R., and Montero-Manso, P. London smart meters dataset (without missing values), June 2020.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36: 19622–19635, 2023.

Gu, X., Liu, Y., Mohsin, Z., Bedford, J., Thakur, A., Watkinson, P., Clifton, L., Zhu, T., and Clifton, D. Are time series foundation models ready for vital sign forecasting in healthcare? In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 401–419. PMLR, 15–16 Dec 2025. URL https://proceedings.mlr.press/v259/gu25a.html.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Hamilton, J. D. *Time Series Analysis*. Princeton University Press, 1994.

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., and McAuley, J. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Hyndman, R. and Athanasopoulos, G. *Forecasting: Principles and Practice*. OTexts, Australia, 2nd edition, 2018.

Hyndman, R. J. and Khandakar, Y. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 27:1–22, 2008.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Krollner, B., Vanstone, B., and Finnie, G. Financial time series forecasting with machine learning techniques: A

survey. In *European Symposium on Artificial Neural Networks: Computational Intelligence and Machine Learning*, pp. 25–30, 2010.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804. PMLR, 2018.

Li, J., Shang, J., and McAuley, J. Uctopic: Unsupervised contrastive learning for phrase representations and topic mining. *arXiv preprint arXiv:2202.13469*, 2022.

Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., Pan, S., and Wen, Q. 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, 2024.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4): 1325–1336, 2022.

Menon, A. K. and Williamson, R. C. A loss framework for calibrated anomaly detection. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1494–1504, 2018.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Olivares, K. G., Challú, C., Garza, A., Canseco, M. M., and Dubrawski, A. NeuralForecast: User friendly state-of-the-art neural forecasting models. PyCon Salt Lake City, Utah, US 2022, 2022. URL https://github.com/Nixtla/neuralforecast.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2019.

Peng, C.-K., Mietus, J. E., Liu, Y., Khalsa, G., Douglas, P. S., Benson, H., and Goldberger, A. L. Exaggerated heart rate oscillations during two meditation techniques. *International Journal of Cardiology*, 70(2):101–107, 1999.

Peterson, M. *An Introduction to Decision Theory*. Cambridge University Press, 2017.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., et al. Forecasting: theory and practice. *International Journal of Forecasting*, 38(3): 705–871, 2022.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020.

Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Bayazi, M. J. D., Adamopoulos, G., Riachi, R., Hassen, N., et al. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Song, H., Diethe, T., Kull, M., and Flach, P. Distribution calibration for regression. In *International Conference on Machine Learning*, pp. 5897–5906. PMLR, 2019.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wen, R., Torkkola, K., Narayanaswamy, B., and Madeka, D. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2017.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Wu, R. and Keogh, E. J. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2421–2429, 2021.

Ye, J., Zhang, W., Yi, K., Yu, Y., Li, Z., Li, J., and Tsung, F. A survey of time series foundation models: Generalizing time series representation with large language model. *arXiv preprint arXiv:2405.02358*, 2024.

Yeh, C.-C. M., Dai, X., Chen, H., Zheng, Y., Fan, Y., Der, A., Lai, V., Zhuang, Z., Wang, J., Wang, L., et al. Toward a foundation model for time series data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4400–4404, 2023.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

A. Experimental Setup

We report the default model parameters used in the main paper results in the following table:

	prediction length	context size	seasonality
Reviews	48	512	24
M5	48	128	7
Glucose	48	128	1
Heart-Rate	48	256	1

The exceptions to the default model parameters are as follows: (1) AutoARIMA had a max lag of 64 timesteps and (2) N-BEATS used a context length of 128 for the Heart-Rate dataset due to limited dataset size.

A.1. Foundation Models

Time series foundation models are generally classified into three categories: (i) adapt pre-trained LLMs (Gruver et al., 2023; Jin et al., 2023), (ii) re-use LLM architectures while trained on time series data from scratch (Ansari et al., 2024; Rasul et al., 2023), and (iii) novel architectures tailored for time series data (Das et al., 2024; Woo et al., 2024; Goswami et al., 2024). In addition to AutoARIMA (Hyndman & Khandakar, 2008) and N-BEATS (Oreshkin et al., 2019) as parametric and neural baselines, we identify four foundation models representing the latter two categories that come with probabilistic predictions out of the box, without requiring post-hoc conformal prediction.

TimesFM TimesFM (Das et al., 2024) uses a decoder-only stacked transformer architecture, and provides probabilistic predictions at pre-trained fixed quantile heads. The training objective combines mean squared error (MSE) and quantile loss (Hyndman & Athanasopoulos, 2018; Gneiting et al., 2007). The model is trained with larger output patches than inputs, and thus able to make joint predictions of forecast quantiles over horizons $h > 1$ with fewer auto-regressive steps.

MOIRAI MOIRAI (Woo et al., 2024) is an encoder-based transformer where the time series prediction is a mixture of several parametric distributions, including Student's t-distribution, negative binomial distribution, and log-normal distribution. Model parameters are optimized to maximize the log-likelihood, and predictive samples are drawn from conditional mixtures of parametric distributions.

Lag-Llama Lag-Llama (Rasul et al., 2023) is another decoder-only transformer model reusing the architecture of Meta's Llama LLM (Touvron et al., 2023). The network is trained on a corpus of real-world time series data, where input patching is modified to fit time series data and an output head is added to predict a conditional Student's t-distribution.

Chronos Chronos (Ansari et al., 2024) is a foundation model based on the T5 family of encoder-decoder language models (Raffel et al., 2020) trained on time series data, where they tokenize real-valued time series into fixed vocabularies via scaling and quantization. The training objective is cross-entropy, and probabilistic forecasting is obtained by sampling different trajectories through repeating one-step-ahead predictions of the next token.

A.2. Datasets

For the Reviews ensemble dataset, we aggregated reviews by product and location category and sampled 10 of the most abundant categories from each dataset. In total the ensemble consisted of 20 time series binned at hourly granularity. We evaluated the models on the period between 02-01-2021 and 03-01-2021.

We aggregated the M5 dataset to reduce sparsity by binning each product by their product department and store id, totaling 70 time series.

The Glucose dataset remained unchanged with no preprocessing steps required. The meditative Heart-Rate dataset is recorded as seconds since start of meditation session. However, foundation models require and use timestamps for forecasting so we mapped the Heart-Rate dataset as if it started at 01-01-2000 and lasted 10 minutes.

Table 1. Calibration evaluation results. Best results are highlighted in bold, and second best results are underlined.

		Foundation Model				Baseline	
		TimesFM	MOIRAI	Chronos	Lag-Llama	N-BEATS	AutoARIMA
MASE	Reviews	0.627	0.785	**0.626**	0.876	1.313	0.800
	M5	0.852	0.968	0.915	1.324	**0.811**	0.867
	Glucose	6.006	7.012	6.783	8.219	6.015	**5.675**
	Heart-Rate	**1.060**	1.106	1.120	1.275	1.212	1.147
PCE	Reviews	0.030	**0.022**	0.063	0.143	0.139	0.137
	M5	0.024	**0.020**	0.075	0.148	0.095	0.116
	Glucose	0.030	**0.019**	0.023	0.109	0.075	0.132
	Heart-Rate	**0.020**	0.023	0.049	0.107	0.219	0.171
CCE	Reviews	**0.007**	-0.011	0.097	-0.213	0.192	-0.247
	M5	**-0.009**	-0.022	0.124	0.087	0.151	-0.203
	Glucose	-0.026	**0.005**	0.032	0.094	-0.093	-0.230
	Heart-Rate	**-0.011**	-0.022	0.051	-0.174	-0.370	-0.306
SIW	Reviews	0.180	0.245	0.124	0.712	0.151	0.691
	M5	0.250	0.310	0.166	0.351	0.119	0.558
	Glucose	0.921	0.976	0.873	0.953	1.082	1.786
	Heart-Rate	0.582	0.650	0.514	1.310	2.711	1.963
WQL	Reviews	**39.109**	49.857	40.752	72.120	97.598	68.954
	M5	**45.662**	51.980	50.308	79.341	51.109	55.722
	Glucose	**44.348**	51.463	49.783	63.292	46.350	47.378
	Heart-Rate	**13.690**	14.379	14.698	18.677	29.783	23.110

B. Additional Results and Figures

B.1. Tail Forecasting

In downstream tasks such as Anomaly Detection, calibration at the tail ends of probabilistic predictions is more important than at the body of a predictive distribution. We evaluated the tailed calibration with a modified PCE that only considers the 0.1 and 0.9 quantile predictions, and found that while some models are better calibrated at the tail end of the probabilistic predictions, foundation models do not have a significant change of calibration at the tails (see Figure 5. Unlike the baselines whose PCE varied considerably when comparing the calibration of the tail and body of the probabilistic forecast, TimesFM and MOIRAI saw a consistent PCE delta of less than 0.02. Chronos and Lag-Llama were less consistent in the delta, but to a lesser extent than the baselines.

B.2. Additional Metrics

As mentioned in the main paper, WQL, MSIS, and CRPS are common metrics for evaluating model calibration. However, as described in Chung et al. (2021), CRPS and Weighted Quantile Loss (WQL), which rely on Pinball loss, measures a combination of probabilistic calibration and sharpness (Gneiting et al., 2007). This arbitrary combination leads to an imbalance often skewing to prioritize predictive sharpness (Chung et al., 2021). MSIS has the same limitations being a measure of interval size with a penalty term for observed values outside the interval (Gneiting et al., 2007; Hyndman & Athanasopoulos, 2018; Gneiting & Raftery, 2007). We found that these metrics were highly correlated to MASE in Figure 3. Therefore, when using these metrics to evaluate calibration, their values resulted in a measure of sharpness and accuracy that diverged from a measurement of calibration (see Figure 6).

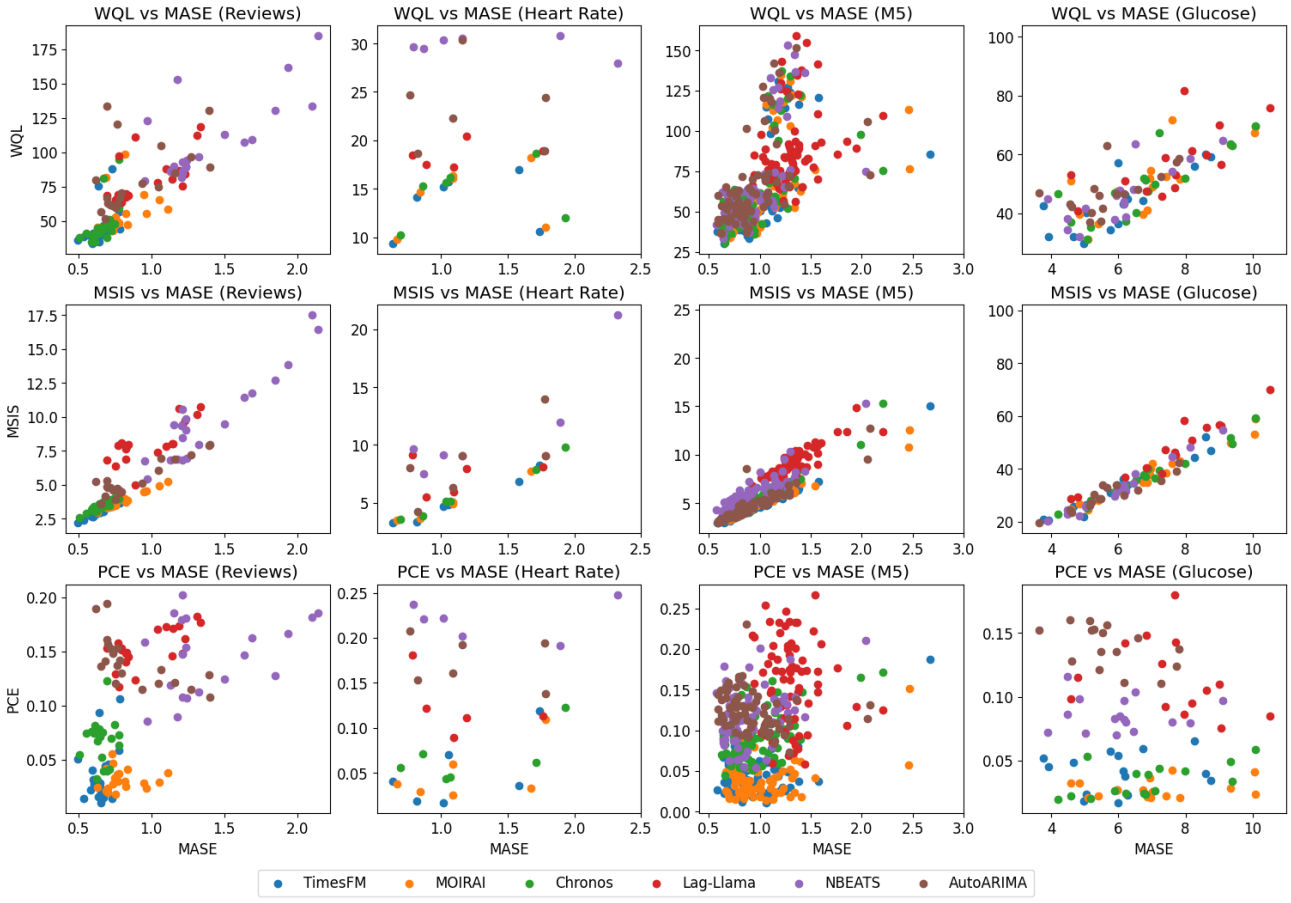

Figure 3. WQL (top), MSIS (middle), and PCE (bottom) compared to MASE where each dot is the performance of an individual time series.

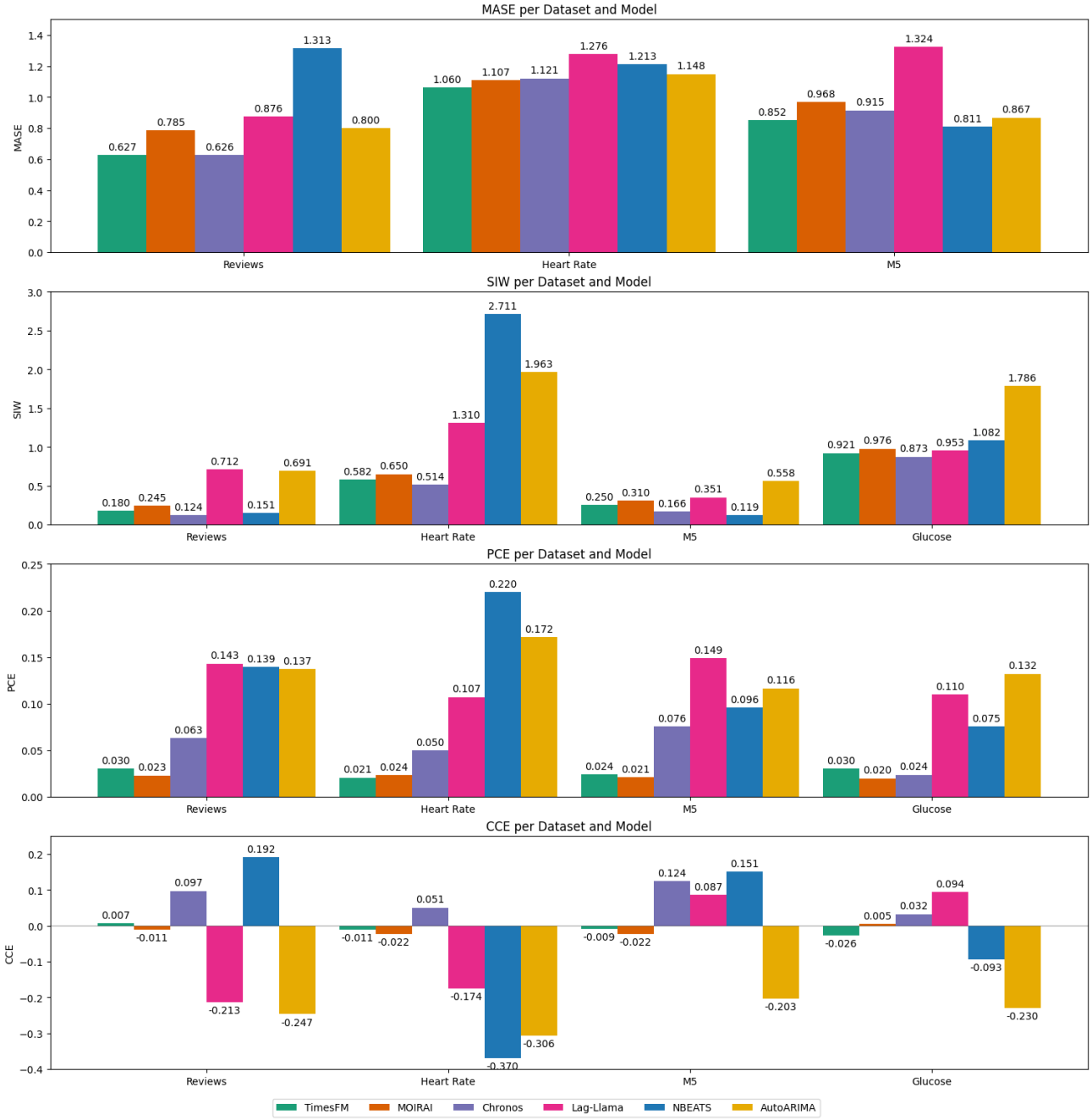

Figure 4. Comparison of MASE, SIW, PCE, and CCE across all models and datasets.

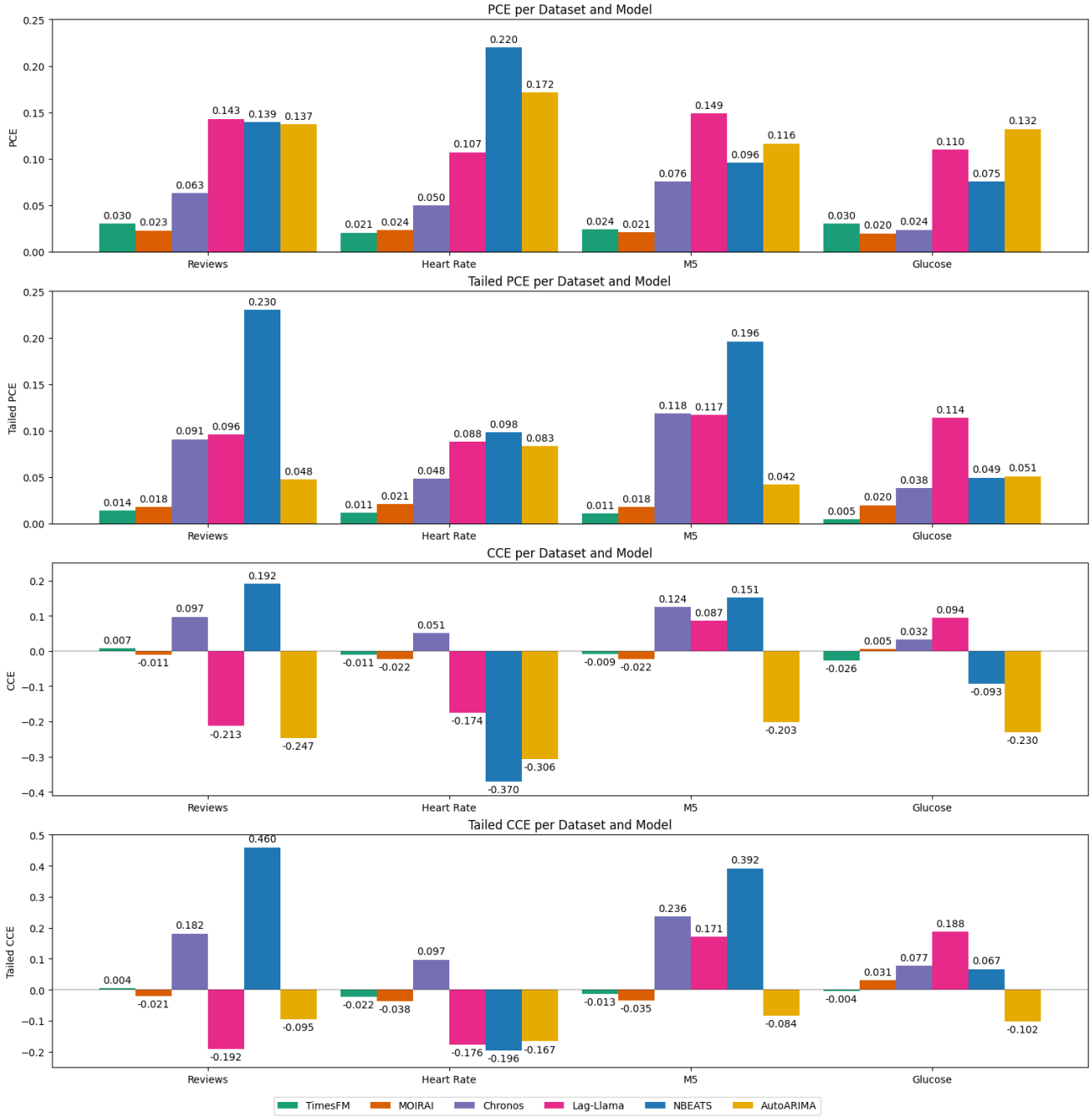

Figure 5. Comparison of model calibration on the entire probabilistic distribution versus calibration of the tail-ends of the distribution.

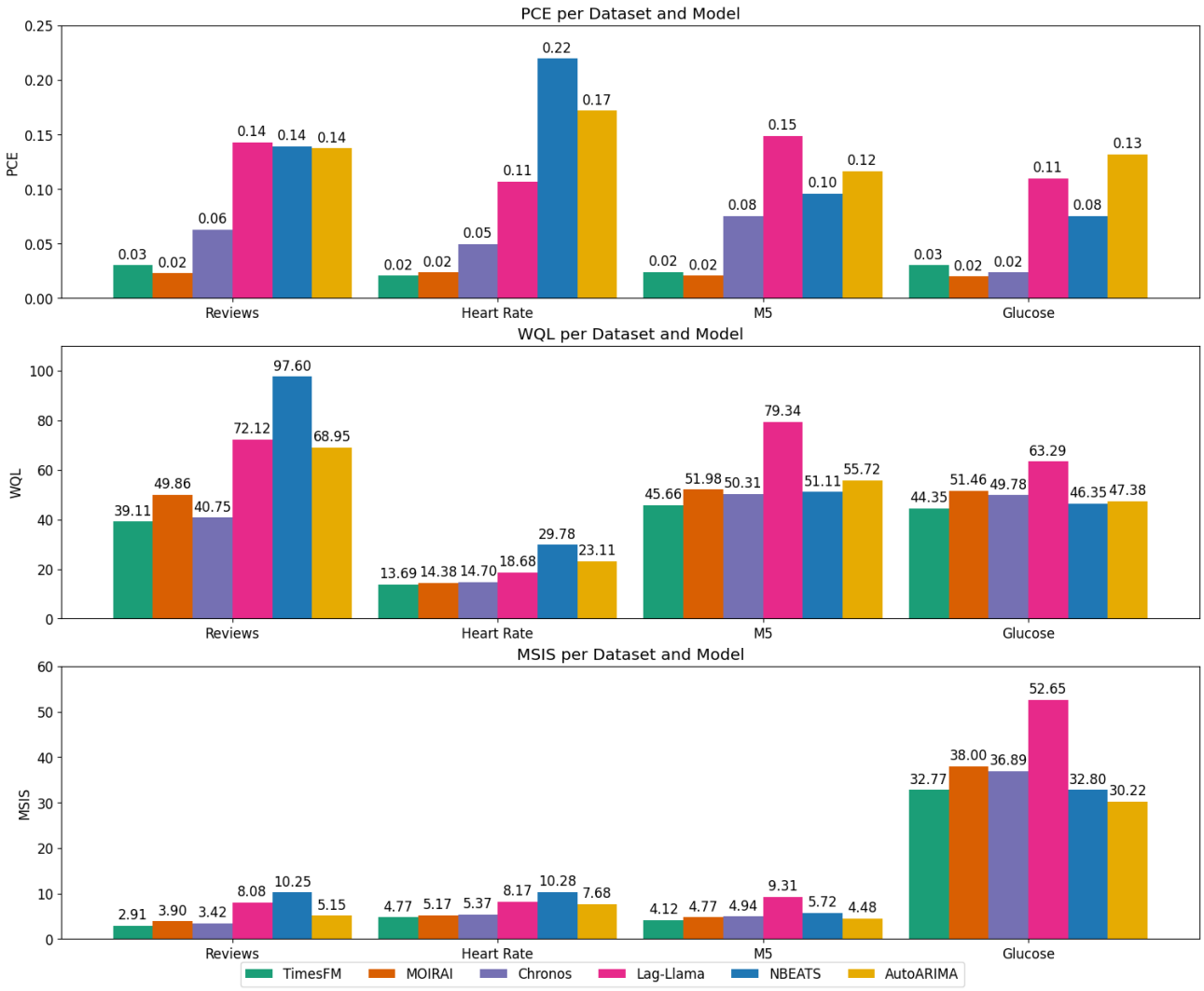

Figure 6. Comparison of PCE to WQL and MSIS across all models and datasets.

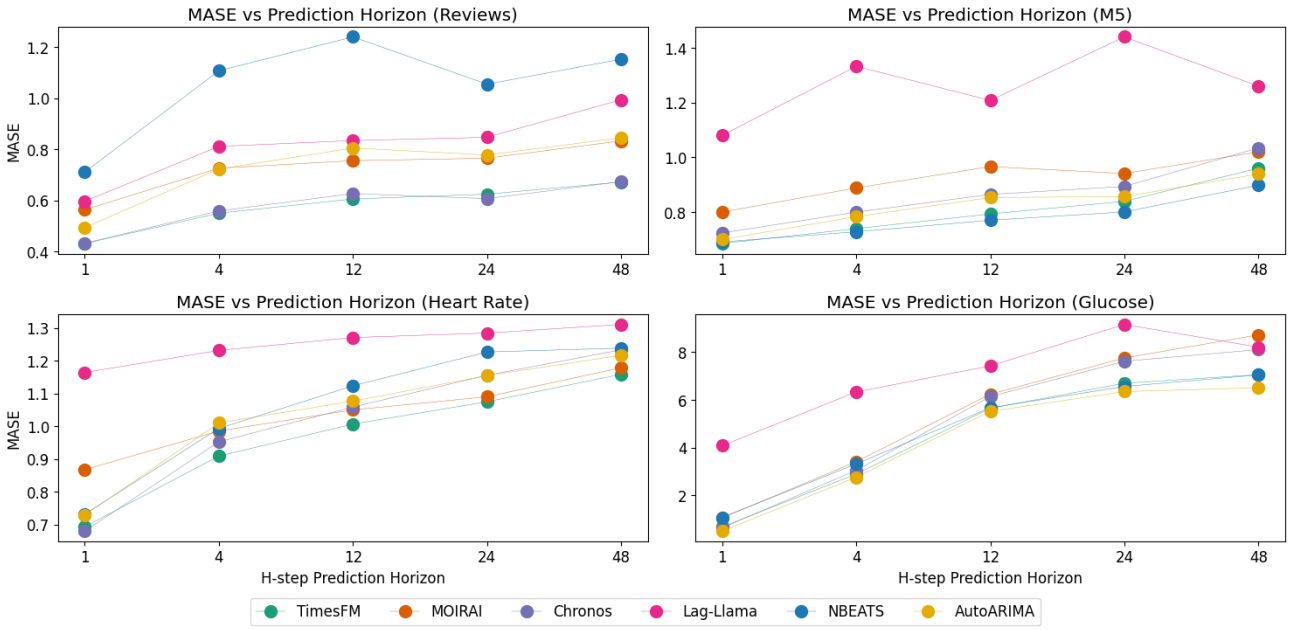

Figure 7. MASE accuracy on all datasets across increasing forecast prediction horizons.

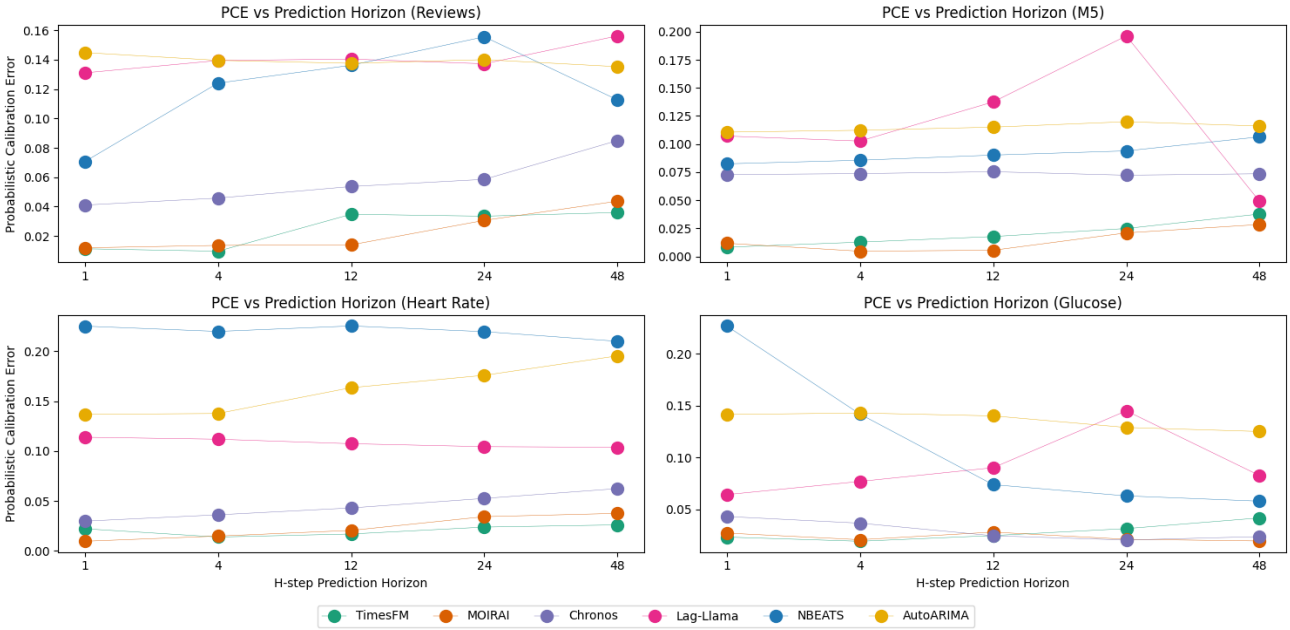

Figure 8. PCE on all datasets across increasing forecast horizon lengths.