Calibration Properties of Time Series Foundation Models: An Empirical Analysis

Coen Adler [1] **Yuxin Chang** [1] **Samar Abdi** [2] **Padhraic Smyth** [1 3]

Abstract

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 five 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 climate science (Mudelsee, 2014), energy forecasting (Deb et al., 2017), healthcare (Crabtree et al., 1990), consumer behavior modeling (Goel et al., 2010), and financial forecasting (Tsay, 2005). 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, including 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 (TSFMs) (Nie et al., 2022; Yeh et al., 2023). Unlike traditional statistical and machine learning approaches where models are fitted to a specific time series, TSFMs 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 is appealing to practitioners in that only a single global model (the foundation model) is required rather than retraining a new model for every time series (Bommasani et al., 2021; Benidis et al., 2022).

With this increase of interest in TSFMs, it becomes important to understand and characterize the calibration properties of such models. 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) and for downstream tasks such as anomaly detection (Menon & Williamson, 2018).

While the calibration properties of classification and regression models have received considerable attention in machine learning in recent years (e.g., Guo et al. (2017); Song et al. (2019); Chung et al. (2021)), there has been significantly less attention on investigating the calibration properties of TSFMs. To address this gap, we empirically evaluate TSFMs and baselines with respect to calibration-specific metrics. More specifically, we evaluate five state-of-the-art TSFMs 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 their calibration properties through a comprehensive set of metrics, such as probabilistic calibration error (Dheur & Taieb, 2023; Kuleshov et al., 2018; Chung et al., 2021), and investigate a variety of different aspects of conditional uncertainty in model predictions. Our main findings[1] for the models and datasets we used in our evaluations are:

- TSFMs generally have better calibration properties than baselines, although some TSFMs can be poorly calibrated.
- Calibration properties of models tend to be consistent across datasets; models that are well (or poorly) cali-

[1]Department of Computer Science, University of California, Irvine, Irvine California, United States [2]Google, Irvine, United States [3]Department of Statistics, University of California, Irvine, Irvine California, United States. Correspondence to: Coen Adler <ctadler@uci.edu>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

[1]Code to reproduce our experiments is available at: https://github.com/Coaster41/TSFM-Calibration.

brated on one dataset tend to also be well (or poorly) calibrated on other datasets.

- Calibration properties of models tend to be robust across varying prediction horizons.

2. Methodology

2.1. Foundation Models and Baselines

We focus on the calibration properties of TSFMs in terms of their zero-shot forecasts. We first identify five foundation models, including one variant, that are trained on time series data from scratch and can perform probabilistic forecasts: Chronos (Ansari et al., 2024a), Chronos-Bolt (Ansari et al., 2024b), TimesFM (Das et al., 2024), MOIRAI (Woo et al., 2024), and Lag-Llama (Rasul et al., 2023). These models are either LLM-based or have their own specific architectures. We also include two baseline models, AutoARIMA (Hyndman & Khandakar, 2008) and N-BEATS (Oreshkin et al., 2019) to represent parametric and neural baselines. Details on pre-trained model checkpoints and hyperparameter settings are provided in Appendix A.1.

Although all five foundation models are transformer-based, Lag-Llama, Chronos, and Chronos-Bolt adapted existing LLMs, while TimesFM and MOIRAI proposed their own transformer architectures. The main differences between these models that affect their calibration are (i) training objectives, (ii) sampling procedures, and (iii) quantile computations.

Specifically, **TimesFM** uses pre-determined quantile heads during training and jointly optimizes mean squared error (MSE) and quantile loss (Wen et al., 2017). **MOIRAI** and **Lag-Llama** maximize the likelihood of a mixture distribution and a Student's t-distribution, respectively. Both **Chronos** and **Chronos-Bolt** transform real-valued time series into a fixed vocabulary via scaling and quantization, and are optimized for classification of binned predictions.

MOIRAI, Lag-Llama, and Chronos autoregressively sample one-step-ahead predictions. These forecasts are then incorporated into the context to predict the next time step. TimesFM and Chronos-Bolt differ in that they can produce joint forecasts over multiple steps. For baselines **N-BEATS** and **AutoARIMA**, we use the `neuralforecast` and `statsforecast` libraries (Olivares et al., 2022) that perform probabilistic forecasts via quantile loss and residual variance respectively. For each time series, the first half (approximately) is used for training and model selection. We evaluate all model forecasts (TSFMs and baselines) on the remaining test data. Unlike the TSFMs which perform zero-shot forecasting, we do a hyper-parameter search for the baseline models using the training set on each time-series. We then use all the data in the training split to re-train N-BEATS while AutoARIMA is refit at each timestep during evaluation.

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. To the best of our knowledge, these datasets 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 three human behavior datasets: (i) a **Reviews** dataset consisting of hourly counts of Amazon product reviews (Hou et al., 2024) and Google Places reviews (Li et al., 2022), (ii) a modified **Shopping (M5)** dataset (Makridakis et al., 2022) consisting of the daily number of products being sold at different locations, and (iii) an NYC **Crime** report dataset (New York City Police Department, 2025) aggregating daily crime occurrences.

Many datasets used in training TSFMs 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 in this respect 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.

To evaluate models on coarser time granularities, we also use the **Patents** dataset (Marco et al., 2015) which counts the number of US patents filed per month between 1981 and 2014.

2.3. Notation and Metrics

In this section, we formalize the notation and define a comprehensive set of metrics for our analysis. Given a *context* of T observations for a time series, $y_{1:T}$, we evaluate 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.

We assess point accuracy using **Mean Absolute Scaled Error (MASE)** for the predicted median, scaling the Mean Absolute Error (MAE) by a naive predictor (Hyndman & Athanasopoulos, 2018):

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

Commonly-used calibration metrics for regression, 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 MASE, potentially hiding weaknesses in a model's calibration performance. Therefore, we use metrics specifically designed for calibration, such as **Probabilistic Calibration Error (PCE)** (Dheur & Taieb, 2023; Kuleshov et al., 2018):

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

with $p = 1$ in our analysis. 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 y_t.

However, well-calibrated models are not sufficient to produce useful forecasts: a model could always predict the marginal distribution, independent of the inputs. 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 $q_{\text{low}} = 2.5\%$ and $q_{\text{high}} = 97.5\%$ 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 preceding 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:

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

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. For both CCE and SIW, we average over $s = \{0.2, 0.4, 0.6, 0.8\}$ in our analyses.

3. Results

Based on the models, datasets, and metrics outlined above, we found that foundation models exhibited competitive (but not necessarily better) point-forecasting performance compared to baselines, with TimesFM, Chronos, Chronos-Bolt, 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. Calibration performance across models showed clear differences with foundation models (except for Lag-Llama) having significantly lower PCE than the baseline models (see the y-axis of Figure 1). Below, we summarize the main results. In the Appendix, we include additional figures and findings, where we discuss the empirical correlation of WQL, MSIS, and MASE, along with details on the calibration of tail probabilities.

3.1. Calibration Error

Figure 1 compares calibration performances (PCE) against point predictions (MASE). The 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, the variation in the y-axis shows that the best foundation models tend to 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 than numbers in the range of 0.15 to 0.2. TimesFM, MOIRAI, and both Chronos models have 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 point-forecasting performance for foundation models varies across datasets, their probabilistic calibration remains relatively consistent. For example, despite all mod-

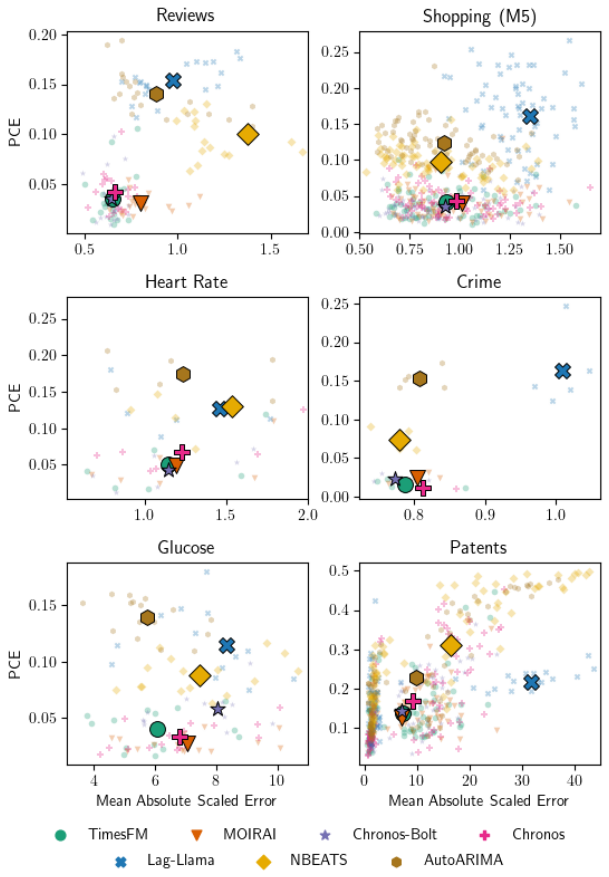

Figure 1. Probabilistic calibration error (PCE) versus point-forecast accuracy (MASE) across the six datasets. Each dot represents model performance on an individual time series; larger centroids being the average over all time series in a dataset.

els performing poorly (in terms of point prediction/MASE) on the Glucose dataset (which is difficult to predict), the well-calibrated models are still well-calibrated even if their point predictions are inaccurate. In general, all of the models have the poorest calibration performance on the Patents dataset; but the TSFMs (except for Lag-Llama) are still better calibrated than the baselines.

3.2. Confidence Calibration

Figure 2 shows foundation model confidence calibration on the Reviews and Crime datasets (see Figure 3 in the Appendix for all datasets). The interval width (SIW) and centered calibration (CCE) tend to have a negative correlation: models with smaller intervals tend to have a larger CCE or were overconfident. Chronos (with top_k as vocabulary size) generally had smaller SIW and greater positive CCE, indicating it being slightly more overconfident than the other foundation models. With the exception of Lag-Llama often having larger (both positive and negative) CCE, the remaining foundation models had low CCE.

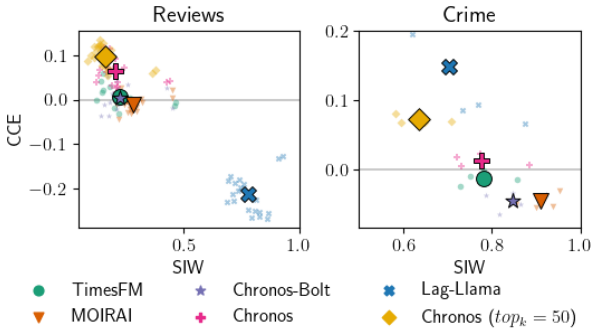

Figure 2. Centered calibration error (CCE) versus scaled interval width (SIW) for foundation models on Reviews and Crime datasets.

We also compared Chronos sampling using only the top 50 tokens ($\text{top}_k = 50$, which is the default setting for Chronos) versus setting top_k to be the vocabulary size. While changing to $\text{top}_k = 50$ does not affect point prediction accuracy (see Figure 10 in the Appendix), the model's confidence significantly changes, now producing more overconfident forecasts (lower SIW and higher positive CCE), likely due to a lack of sampling of low-probability tokens.

3.3. Forecast Horizon Length

For point-forecasting, we found that MASE increased significantly as the prediction horizon increased (see Figure 8 in the Appendix). However, in terms of calibration, the foundation models tend to be more robust across increasing prediction horizons, increasing only a couple of percent in PCE when comparing predictions from 1 time-step ahead to 48 steps (Figure 9 in the Appendix).

4. Conclusions and Future Work

In summary, we systematically evaluated the calibration properties of TSFMs on six sets of time series datasets and found that the best-performing models, TimesFM, MOIRAI, and both Chronos models, were consistently well-calibrated relative to baselines. However, not all foundation models were accurate or well-calibrated, with Lag-Llama for example having both poor point-prediction accuracy and poor probabilistic calibration.

Potential directions for future work include more in-depth investigations of sensitivity of calibration metrics to training objectives and prediction hyperparameters, broader ranges of datasets for evaluation, development of methods for improving calibration (e.g., in an online manner), and measuring calibration performance in the context of few-shot inference (via fine-tuning). Additionally, it would be of interest to assess how calibration performance affects the performance of models in downstream tasks such as anomaly detection based on model predictive uncertainty.

Acknowledgments

We thank the workshop reviewers for numerous useful suggestions that improved the quality of the paper. This work was supported in part by the Hasso Plattner Institute (HPI) Research Center in Machine Learning and Data Science at the University of California, Irvine, by a Google faculty award, and by the National Institutes of Health under award 1R01CA297869-01.

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

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 2024a.

Ansari, A. F., Turkmen, C., Shchur, O., and Stella, L. Fast and accurate zero-shot forecasting with Chronos-Bolt and AutoGluon, Dec 2024b. URL https://aws.amazon.com/blogs/machine-learning/fast-and-accurate-zero-shot-forecasting-with-chronos-bolt-and-autogluon/.

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.

Crabtree, B. F., Ray, S. C., Schmidt, P. M., O'Connor, P. T., and Schmidt, D. D. The individual over time: time series applications in health care research. *Journal of Clinical Epidemiology*, 43(3):241–260, 1990.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Deb, C., Zhang, F., Yang, J., Lee, S. E., and Shah, K. W. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017.

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.

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.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, 2010.

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.

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.

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.

Marco, A. C., Carley, M., Jackson, S., and Myers, A. The USPTO historical patent data files two centuries of innovation, 2015.

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.

Mudelsee, M. *Climate Time Series Analysis: Classical Statistical and Bootstrap Methods*, volume 51. Springer, 2014.

New York City Police Department. NYPD Complaint Data Historic. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i/about$_data$, 2025.

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.

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.

Tsay, R. S. *Analysis of Financial Time Series*. John Wiley & Sons, 2005.

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.

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 experiment parameters used in the main paper results in the following table:

Table 1. Experimental setup details per dataset.

	prediction length	context size	seasonality	train/test split
Reviews	48	512	24	2021-01-31 23:00
Shopping (M5)	48	128	7	2015-04-23
Glucose	48	128	1	2020-02-16 12:43
Heart-Rate	48	256	1	2000-01-01 00:04:59
Crime	48	128	7	2015-01-01
Patents	48	128	1	2004-01-01

Prediction length is fixed at 48 timesteps across all datasets. Seasonality is only used in initializing AutoARIMA. The exceptions to the default model parameters are as follows: (i) AutoARIMA has a max lag of 64 timesteps and (ii) N-BEATS uses a context length of 128 for the Heart-Rate dataset due to limited dataset size.

Baseline models (N-BEATS and AutoARIMA) use the training data (dates before train/test split column) for parameter selection and training. Models forecast the first 48 time steps of the test set using the end of the training set as the context. The forecast date is then shifted by a single time step, forecasting an additional 48 time steps. The models forecast at each timestep in the time series.

A.1. Models

TSFMs 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., 2024a; 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 five 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 than repeating one-step-ahead predictions. We use the `timesfm-2.0-500m-pytorch` version for our experiments.[2]

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 single-step autoregressive samples are drawn from conditional mixtures of parametric distributions. We use the `moirai-1.1-R-small` version for our experiments with default 100 samples. [3]

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) for single-step autoregressive forecasting. 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 (use default 100 samples). [4]

Chronos/Chronos-Bolt Chronos (Ansari et al., 2024a) 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

[2]TimesFM checkpoint on Hugging Face: https://huggingface.co/google/timesfm-2.0-500m-pytorch.

[3]MOIRAI checkpoint on Hugging Face: https://huggingface.co/Salesforce/moirai-1.1-R-small

[4]Lag-LLama checkpoint on Hugging Face: https://huggingface.co/time-series-foundation-models/Lag-Llama

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. The `chronos-t5-small`[5] version was used in our experiments with default parameters except setting the top_k parameter to the size of the vocabulary (4096). We use the default num_samples: 20. We also tested with the Chronos-Bolt model which uses direct multi-step joint forecasting similar to TimesFM for significantly faster inference times (Ansari et al., 2024b). [6]

N-BEATS N-BEATS (Oreshkin et al., 2019) is a deep neural architecture that has been designed for the purposes of time-series predictions. Similarly to TimesFM, N-BEATS jointly forecasts the entire prediction horizon in a single forward pass. Due to the sensitivity of the model results to hyper-parameters, we perform a grid search on a variety of parameters (see Table 2) using the held-out training data. Quantile loss outperforms normal distribution loss on all datasets. When forecasting with distribution loss, we use the default 100 samples.

Table 2. N-BEATS hyper-parameter grid-search space.

Parameter	Search Space
Epochs	{100, 1000}
Learning Rate	{0.001, 0.0001}
Early Stop Patience Steps	{-1, 2}
Number of Blocks	{(1,1,1), (3,3,3)}
Validation Check Steps	{10, 100}
Loss Function	{Normal Distribution Loss, Quantile Loss}

AutoARIMA AutoARIMA (Hyndman & Khandakar, 2008) is a variation of the statistical ARIMA model. We use Nixtla's StatsForecast implementation of AutoARIMA that automatically selects the optimal ARIMA parameters for each timeseries on the training set. The model is then refit on all earlier data before each forecast on the evaluation set. The ARIMA implementation uses Kalman filters to recursively predict the mean and variance. Quantiles are computed by fitting a normal distribution to the forecasts and using the inverse CDF (PPF) with the appropriate z-scores.

A.2. Datasets

For the **Reviews** ensemble dataset, we aggregated Amazon (Hou et al., 2024) and Google (Li et al., 2022) 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 trimmed the dataset to consist of data from 2021-01-04 00:00 to 2021-03-01 23:00.

We aggregated the **Shopping (M5)** (Makridakis et al., 2022) dataset to reduce sparsity by binning each product by their product department and store ID, totaling 70 time series.

The **Glucose** (Cho et al., 2023) dataset remained unchanged with no preprocessing steps required. We used the "Dexcom" measurement values.

The meditative **Heart-Rate** dataset (Peng et al., 1999) 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 to start at 2000-01-01 00:00:00 and last 10 minutes.

We cutoff the NYC **Crime** reports dataset (New York City Police Department, 2025) to only include reports between 2006 and 2023, and we split the dataset into timeseries based on borough.

For the US **Patents** dataset (Marco et al., 2015), we removed sparse time series with a minimum value of less than 100.

[5]Chronos checkpoint on Hugging Face: `https://huggingface.co/amazon/chronos-t5-small`
[6]Chronos-Bolt checkpoint on Hugging Face: `https://huggingface.co/amazon/chronos-bolt-small`

Table 3. Dataset information breakdown

	# Time Series	Time Granularity	# Time Steps
Reviews	20	Hourly	1368
Shopping (M5)	70	Daily	1912
Glucose	16	5 Min	1686
Heart-Rate	6	Second	744
Crime	5	Daily	6574
Patents	83	Monthly	408

B. Additional Results and Figures

Table 4. Calibration evaluation results. Best results are highlighted in bold, and second best results are underlined.

		Foundation Models					Baselines	
		TimesFM	MOIRAI	Chronos-Bolt	Chronos	Lag-Llama	N-BEATS	AutoARIMA
MASE	Reviews	0.627	0.785	0.631	**0.626**	0.876	1.247	0.800
	M5	0.852	0.968	0.846	0.914	1.324	**0.802**	0.867
	Glucose	6.007	7.012	8.022	6.764	8.220	7.381	**5.676**
	Heart-Rate	**1.060**	1.107	1.074	1.122	1.276	1.404	1.148
	Crime	0.789	0.810	**0.774**	0.814	1.009	0.776	0.812
	Patents	7.451	7.037	**6.675**	9.816	33.554	18.738	11.566
PCE	Reviews	0.030	**0.023**	0.031	0.035	0.143	0.077	0.137
	M5	0.024	0.021	**0.008**	0.026	0.149	0.083	0.116
	Glucose	0.030	**0.020**	0.057	0.022	0.110	0.066	0.132
	Heart-Rate	0.021	0.024	**0.016**	0.051	0.107	0.087	0.172
	Crime	0.014	0.023	0.020	**0.010**	0.160	0.065	0.152
	Patents	0.050	**0.039**	0.070	0.117	0.101	0.211	0.166
CCE	Reviews	0.007	-0.011	**0.004**	0.065	-0.213	0.126	-0.247
	M5	-0.009	-0.022	**-0.009**	0.041	0.087	-0.181	-0.203
	Glucose	-0.026	**0.005**	0.034	0.036	0.094	0.075	-0.230
	Heart-Rate	**-0.011**	-0.022	-0.027	0.062	-0.174	-0.170	-0.306
	Crime	-0.014	-0.046	-0.045	**0.013**	0.149	-0.127	-0.340
	Patents	0.082	-0.042	0.011	0.132	-0.005	0.274	**0.001**
SIW	Reviews	0.180	0.245	0.184	0.157	0.712	0.297	0.691
	M5	0.250	0.310	0.239	0.243	0.351	0.419	0.558
	Glucose	0.921	0.976	1.010	0.878	0.953	0.869	1.786
	Heart-Rate	0.582	0.650	0.615	0.519	1.310	1.305	1.963
	Crime	0.289	0.341	0.315	0.287	0.276	0.477	0.847
	Patents	0.048	0.098	0.067	0.048	0.597	0.020	0.066
WQL	Reviews	**39.109**	49.857	39.343	39.870	72.120	86.890	68.954
	M5	45.662	51.980	**45.561**	49.168	79.341	48.126	55.722
	Glucose	**44.348**	51.463	58.524	49.627	63.292	55.980	47.378
	Heart-Rate	**13.690**	14.379	13.887	14.686	18.677	21.530	23.110
	Crime	31.322	32.678	**31.053**	32.605	42.045	32.235	40.334
	Patents	10.559	9.983	**9.328**	14.593	55.518	37.193	19.772

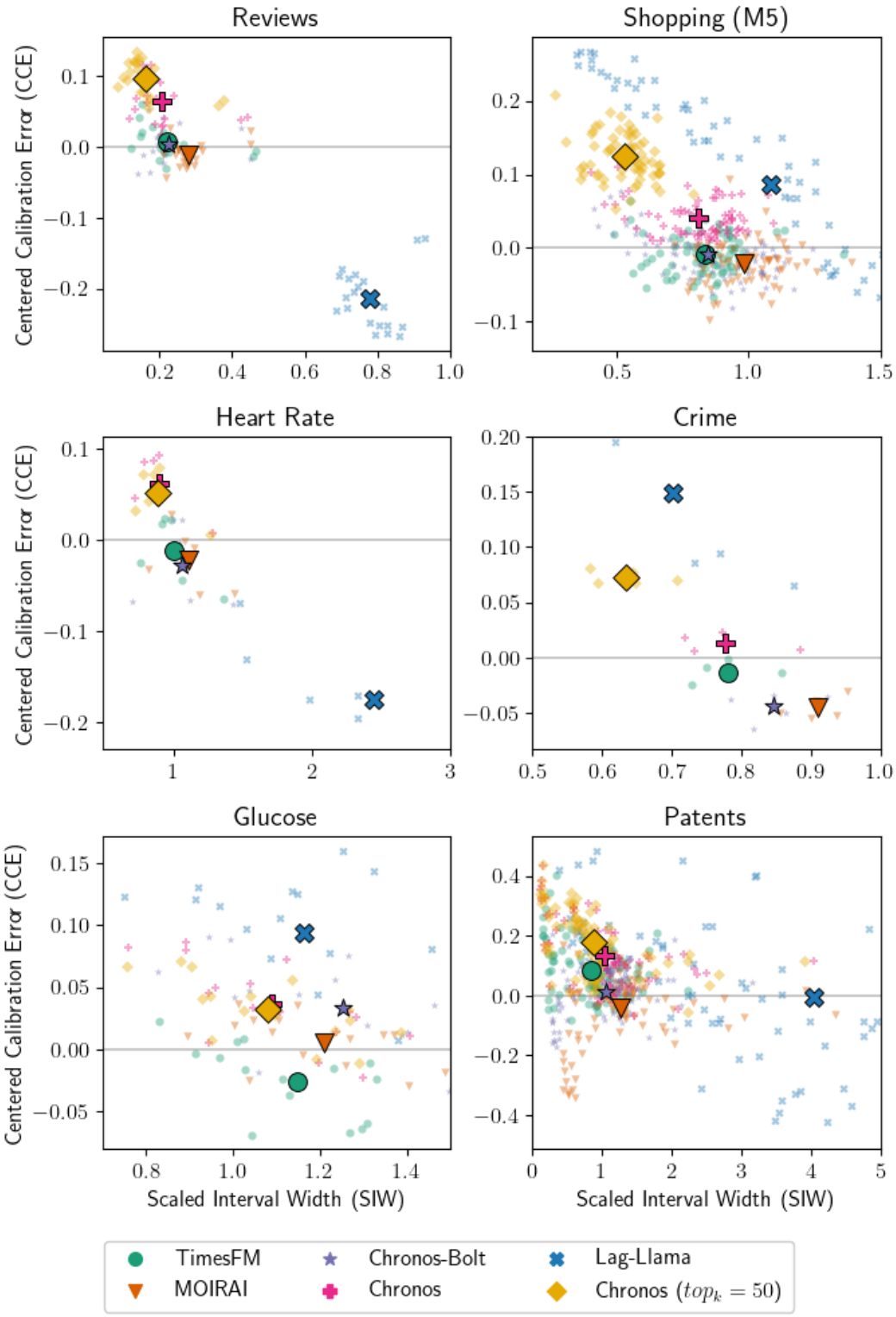

Figure 3. Centered calibration error (CCE) versus scaled interval width (SIW) for foundation models. Chronos ($top_k = 50$) is typically more overconfident compared to the other foundation models.

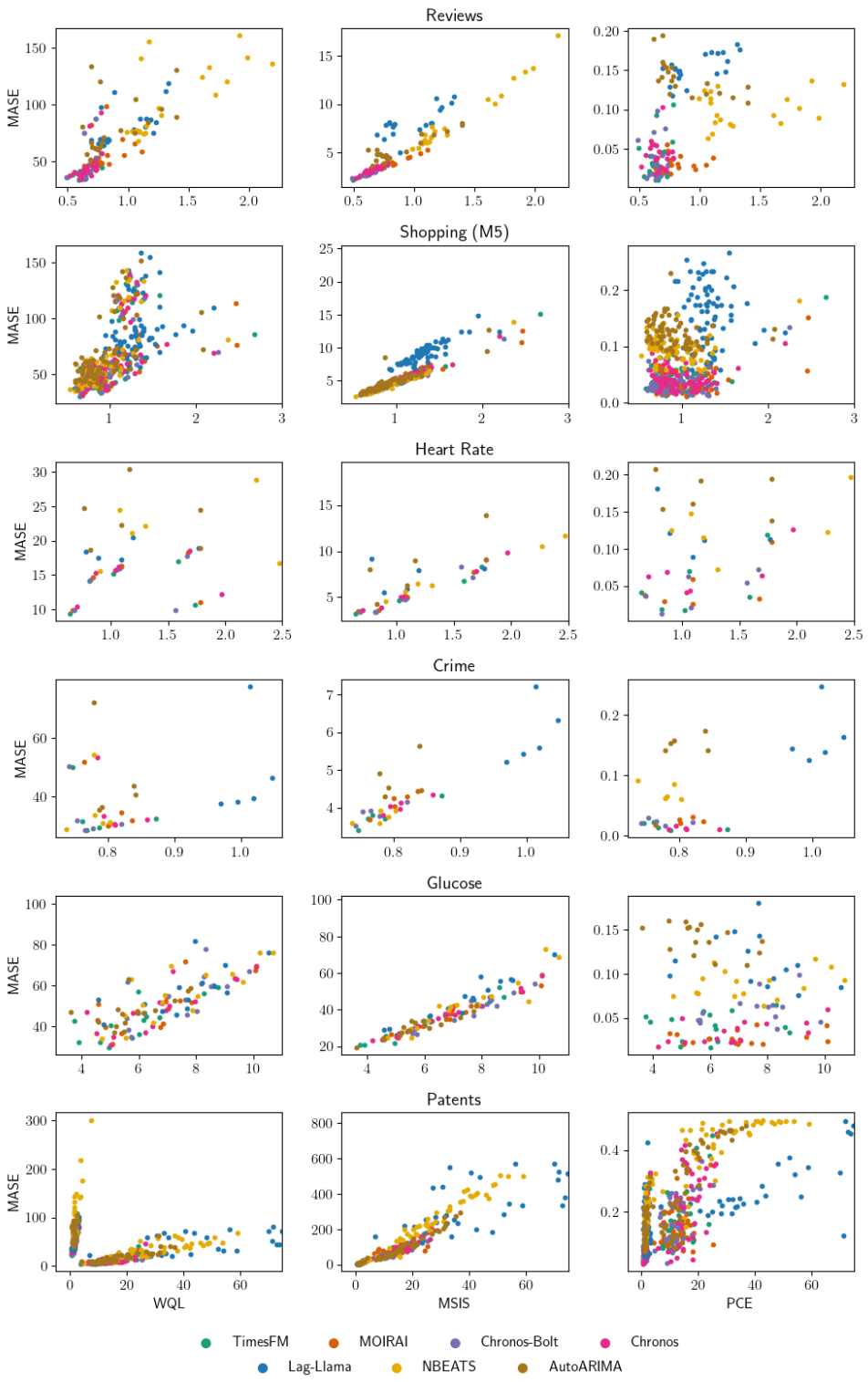

Figure 4. WQL (left), MSIS (middle), and PCE (right) compared to MASE where each dot is the performance of an individual time series. WQL and MSIS show (strong) positive correlation with MASE while PCE and MASE are not correlated.

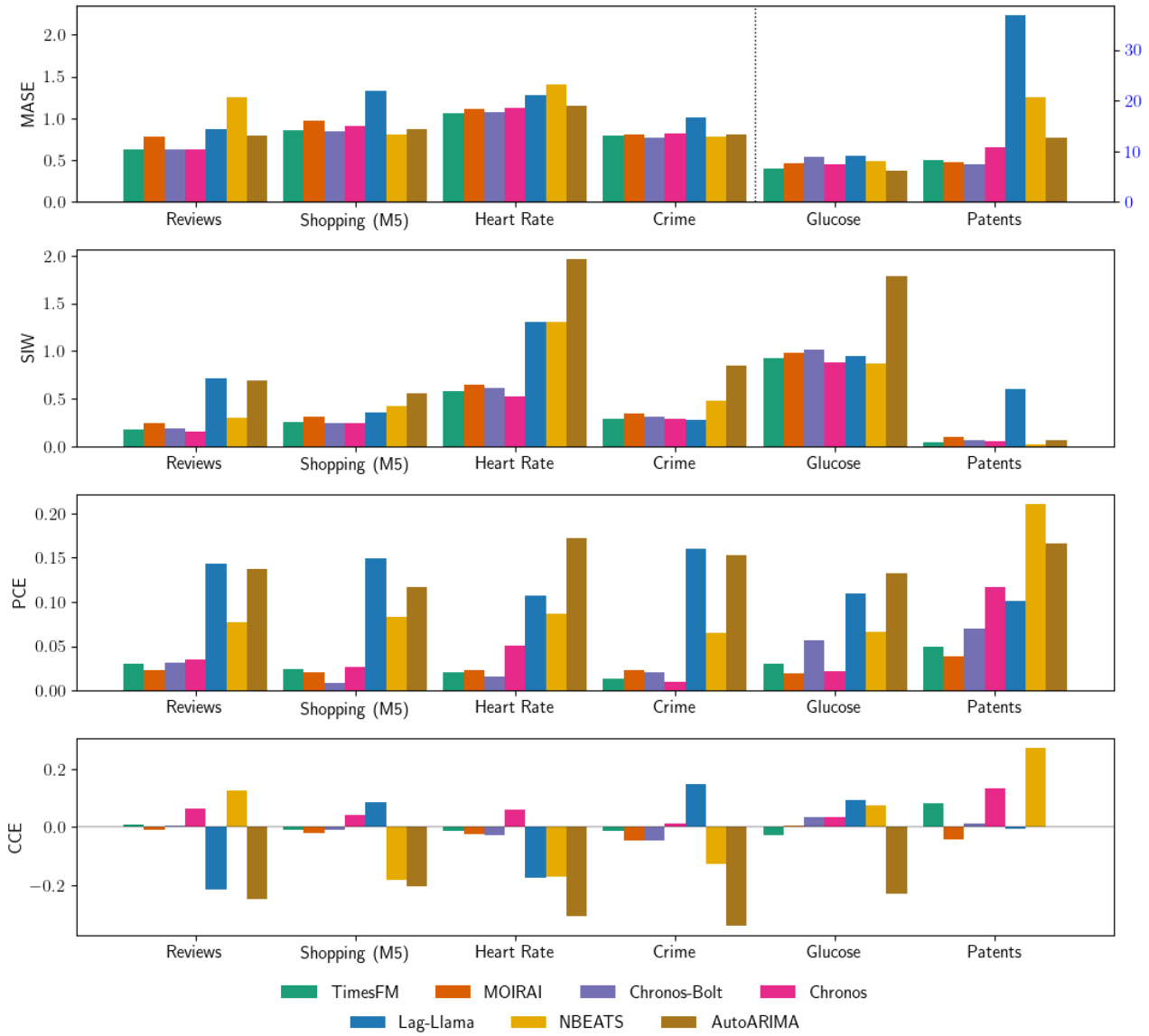

Figure 5. Comparison of MASE, SIW, PCE, and CCE across all models and datasets. MASE y-axis for Glucose and Patents are on a separate scale from the other datasets.

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 evaluate the tailed calibration with a modified PCE that only considers the 0.1 and 0.9 quantile predictions, and find 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 6. Unlike the baselines whose PCE varies considerably when comparing the calibration of the tail and body of the probabilistic forecast, TimesFM and MOIRAI see a consistent PCE delta of less than 0.02. Chronos and Lag-Llama are less consistent in the delta, but to a lesser extent than the baselines.

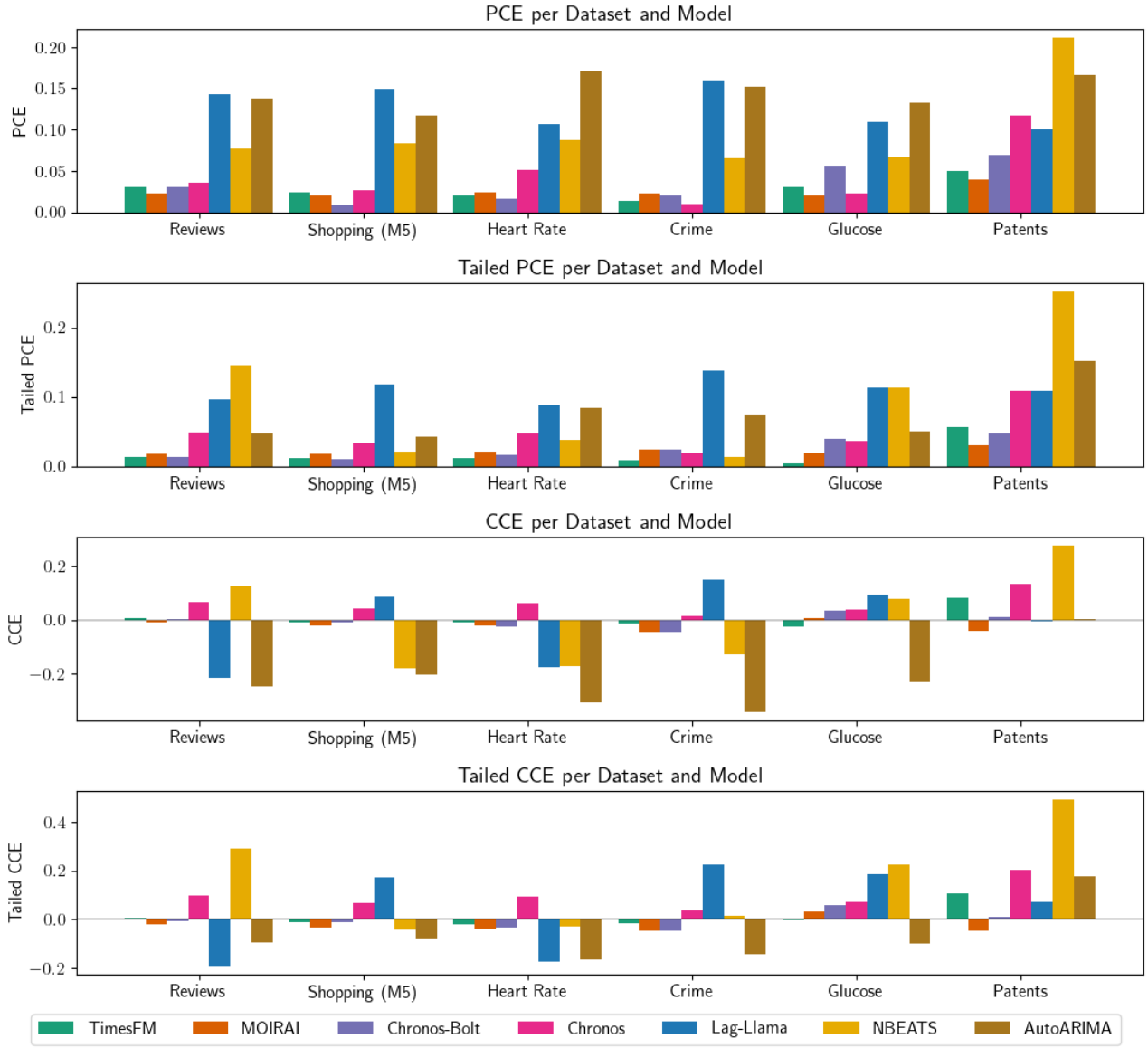

Figure 6. Comparison of model calibration on the entire probabilistic distribution versus calibration of the tail-ends of the distribution. Tailed calibration errors are computed using only the 0.1 and 0.9 quantiles.

B.2. Additional Metrics

As mentioned in the main paper, WQL, MSIS, and CRPS are common metrics for evaluating model calibration. Weighted Quantile Loss (WQL) is an approximation of Continuous Ranked Probability Score (CRPS) defined as the pinball (or quantile) loss p_q scaled by the absolute sum of the true values:

$$p_q(y_t, \hat{y}_t^q) = \begin{cases} 2 \cdot (1-q) \cdot (\hat{y}_t^q - y_t) & \text{if } \hat{y}_t^q \geq y_t \\ 2 \cdot q \cdot (y_t - \hat{y}_t^q) & \text{if } \hat{y}_t^q < y_t \end{cases} \tag{5}$$

$$WQL = \frac{1}{\sum_{t=T+1}^{T+H} |y_t|} \sum_{t=T+1}^{T+H} \sum_q p_q(y_t, \hat{y}_t^q) \tag{6}$$

However, as described in Chung et al. (2021), CRPS and WQL, measure 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).

Mean scaled interval score (MSIS) is a scaled version of mean interval score (MIS) which is the mean of the difference in upper and lower bound prediction summed with the error when outside the bounds:

$$
\begin{aligned}
MSIS = \frac{1}{MAE_n} \frac{1}{H} \sum_{t=T+1}^{H} & (U_t^s - L_t^s) \\
& + \frac{2}{1-s}(L_t^s - y_t)\mathbf{1}[y_t < L_t^s] \\
& + \frac{2}{1-s}(y_t - U_t^s)\mathbf{1}[y_t > U_t^s]
\end{aligned}
\tag{7}
$$

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 find that these metrics were highly correlated to MASE in Figure 4. Therefore, when using these metrics to evaluate calibration, their values result in a measure of sharpness and accuracy that diverge from a measurement of calibration (see Figure 7). For example, if we evaluate calibration using WQL or MSIS, we would incorrectly conclude that AutoARIMA is equally or better calibrated than the best foundation models on the Glucose dataset.

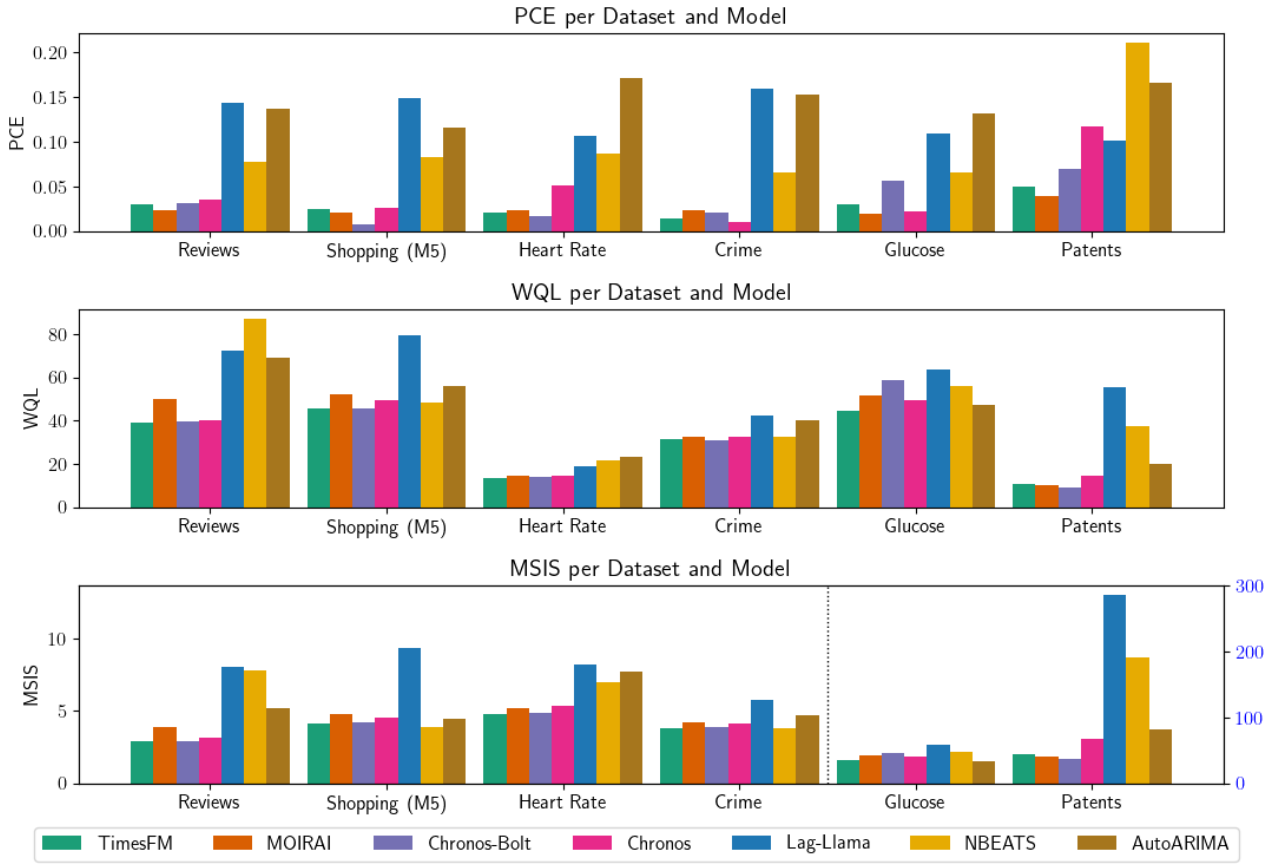

Figure 7. Comparison of PCE to WQL and MSIS across all models and datasets. MSIS y-axis for Glucose and Patents are on a separate scale from the other datasets. Models with the lowest PCE do not always have the lowest WQL or MSIS.

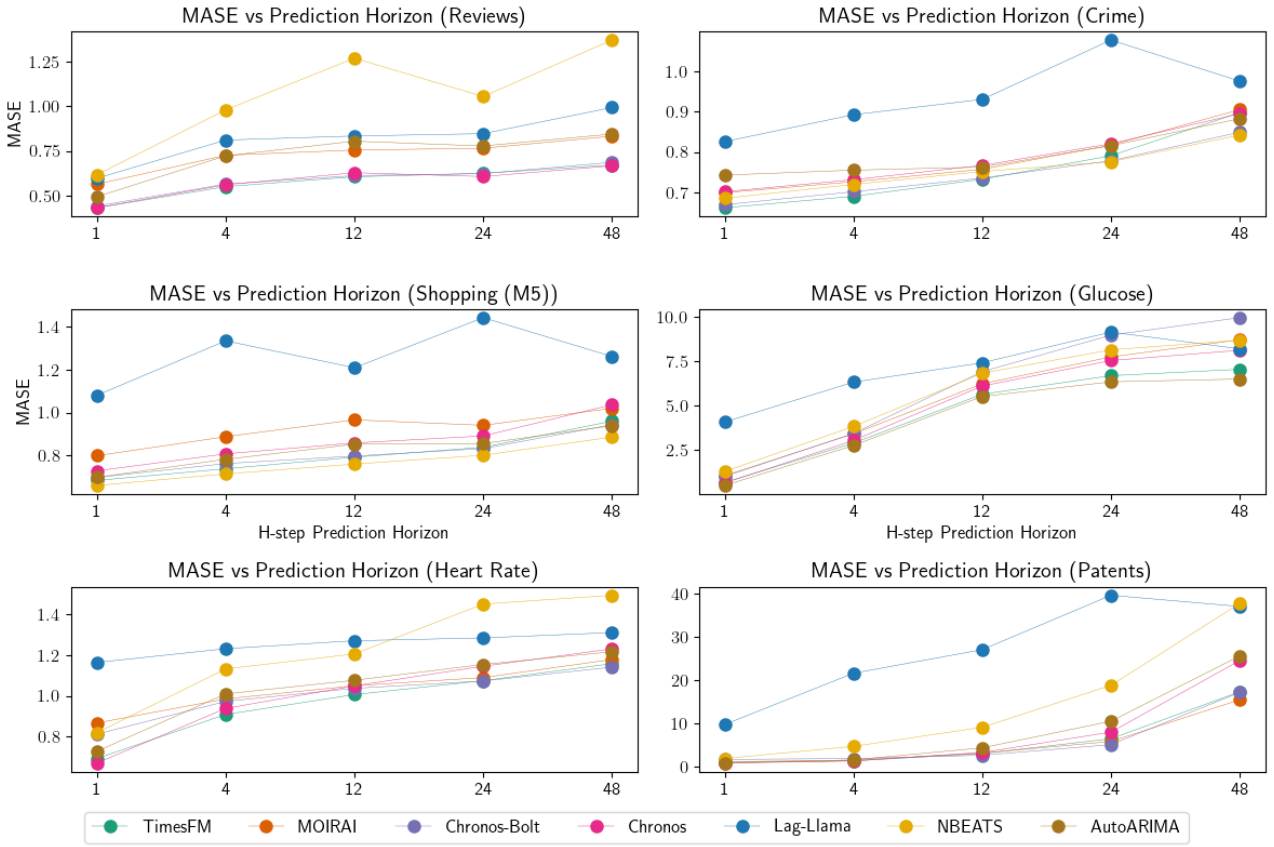

Figure 8. MASE accuracy on all datasets across increasing forecast prediction horizons. Point forecasting error tends to increase with a further prediction horizon.

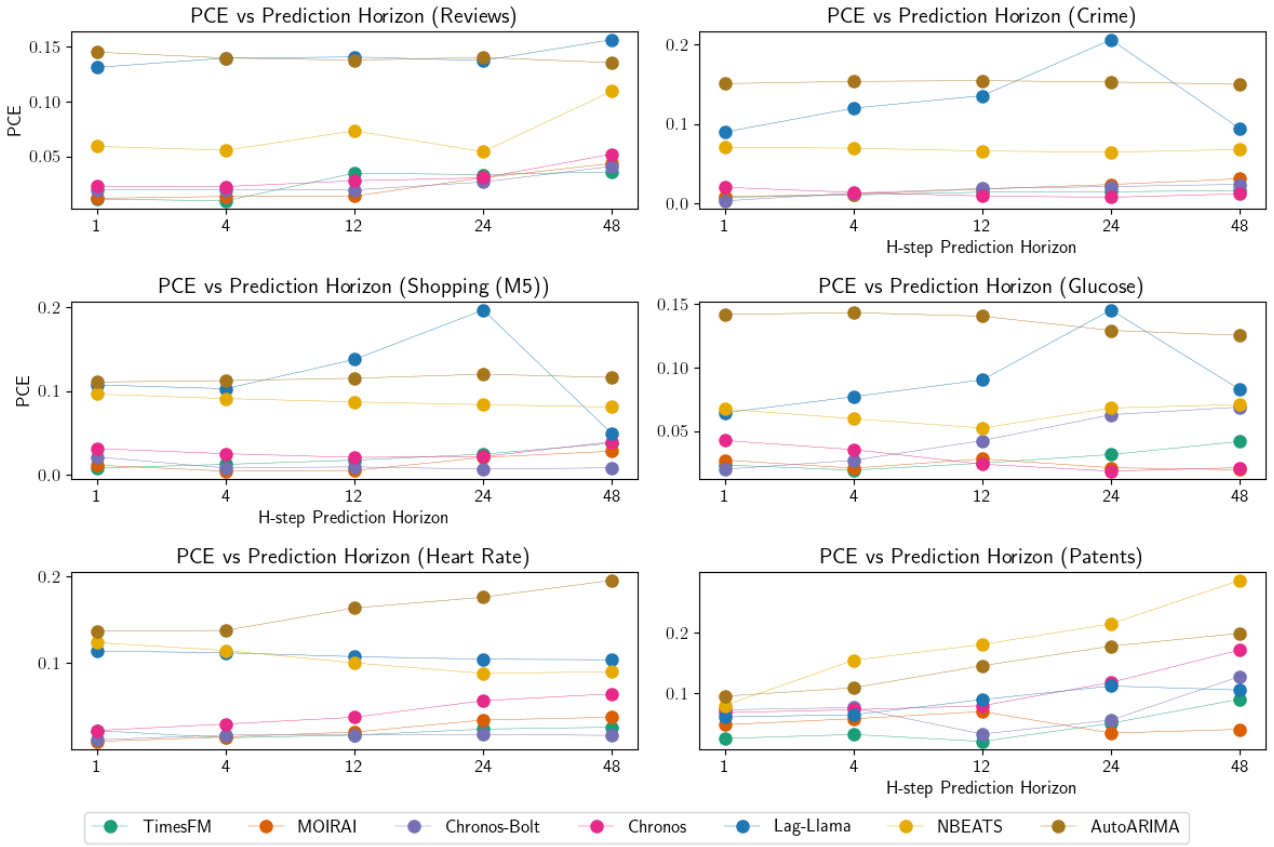

Figure 9. PCE on all datasets across increasing forecast horizon lengths. Model accuracy does not consistently increase with a further prediction horizon.

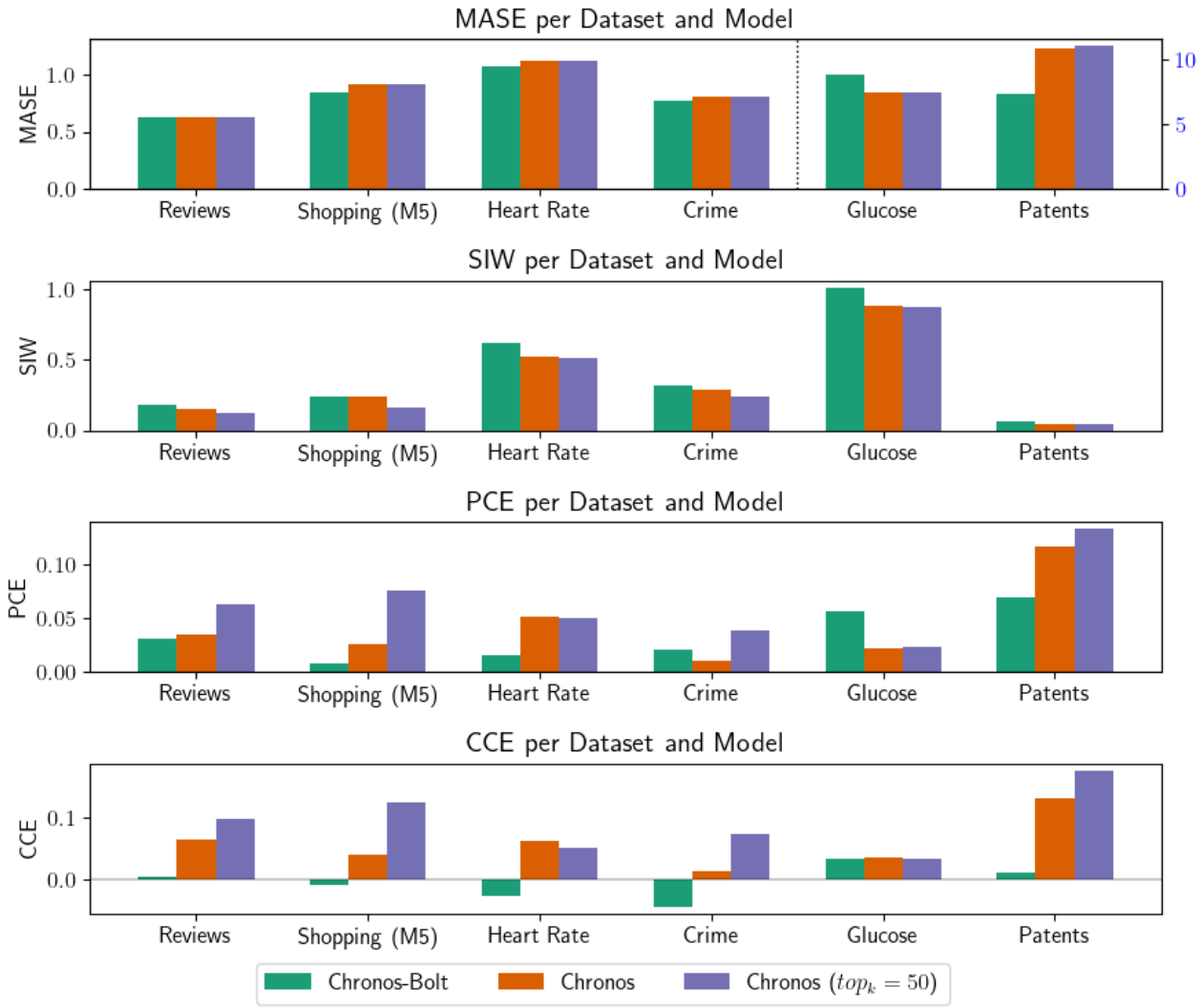

Figure 10. Comparison between Chronos models and when using $top_k = 50$ compared to the entire vocabulary.

B.3. Visualizations on Model Forecasts

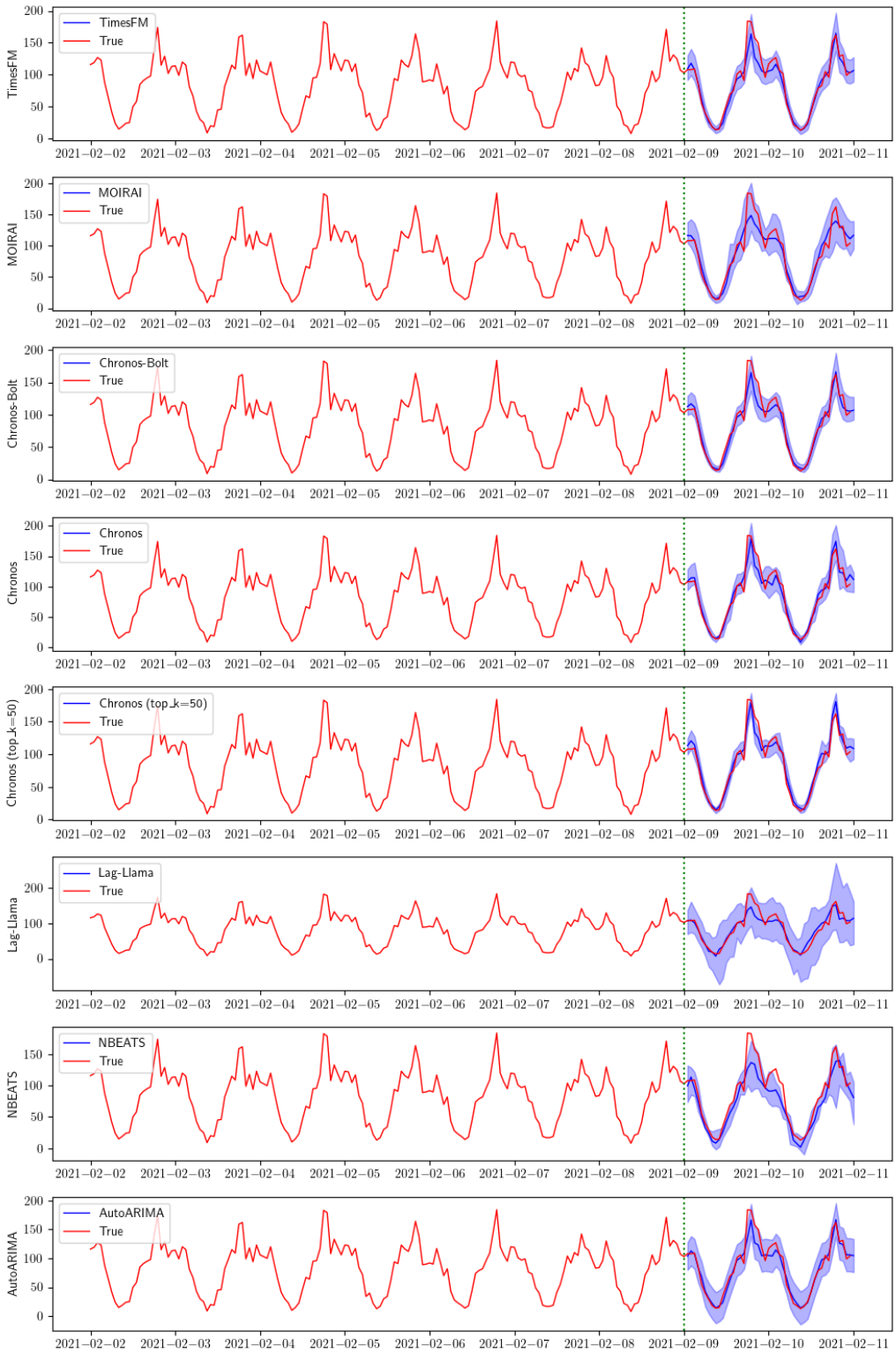

Figure 11. Model forecasts on Reviews dataset. Prediction timestamp randomly selected.

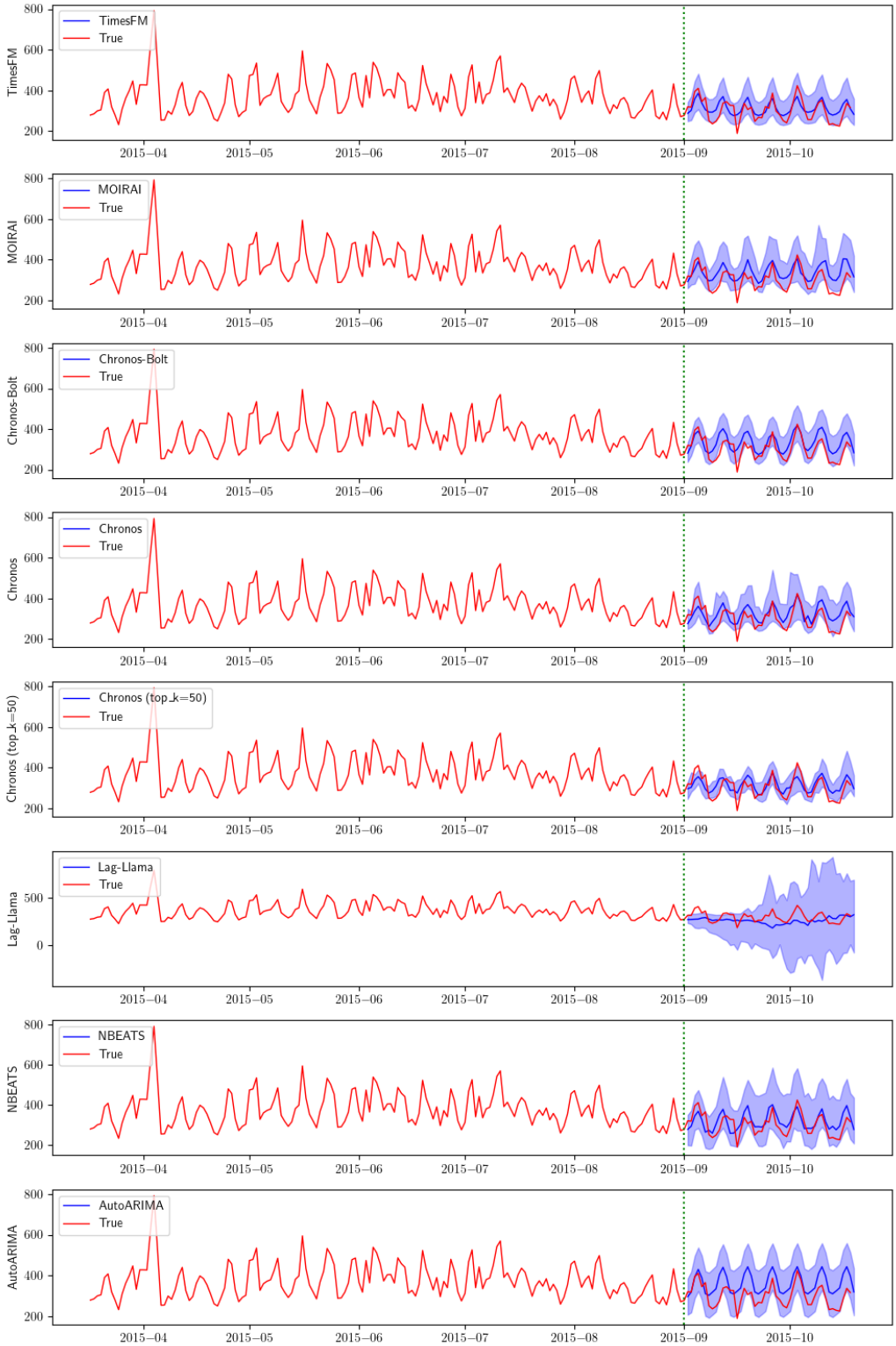

Figure 12. Model forecasts on M5 dataset. Prediction timestamp randomly selected.

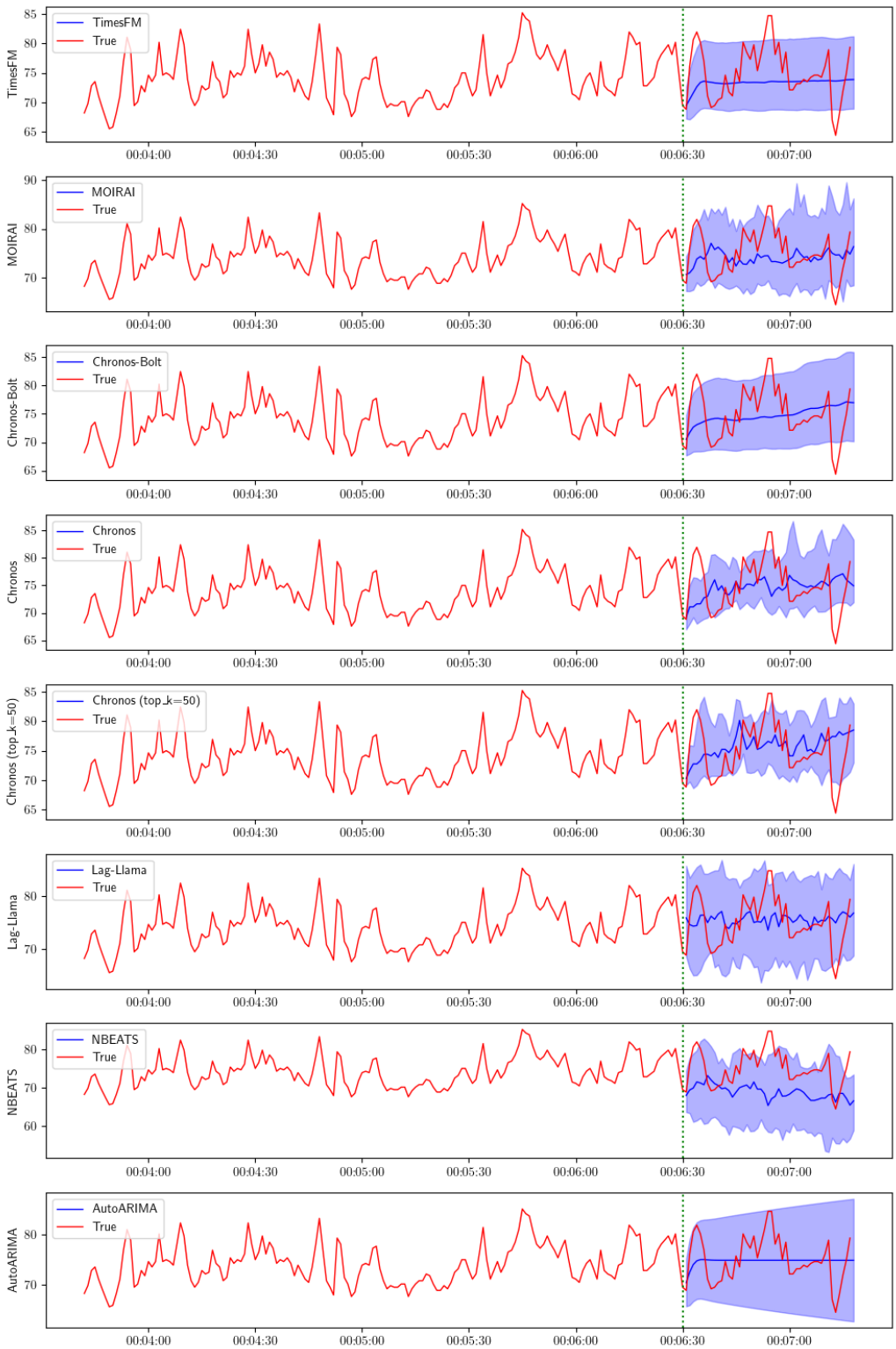

Figure 13. Model forecasts on Heart-Rate dataset. Prediction timestamp randomly selected.

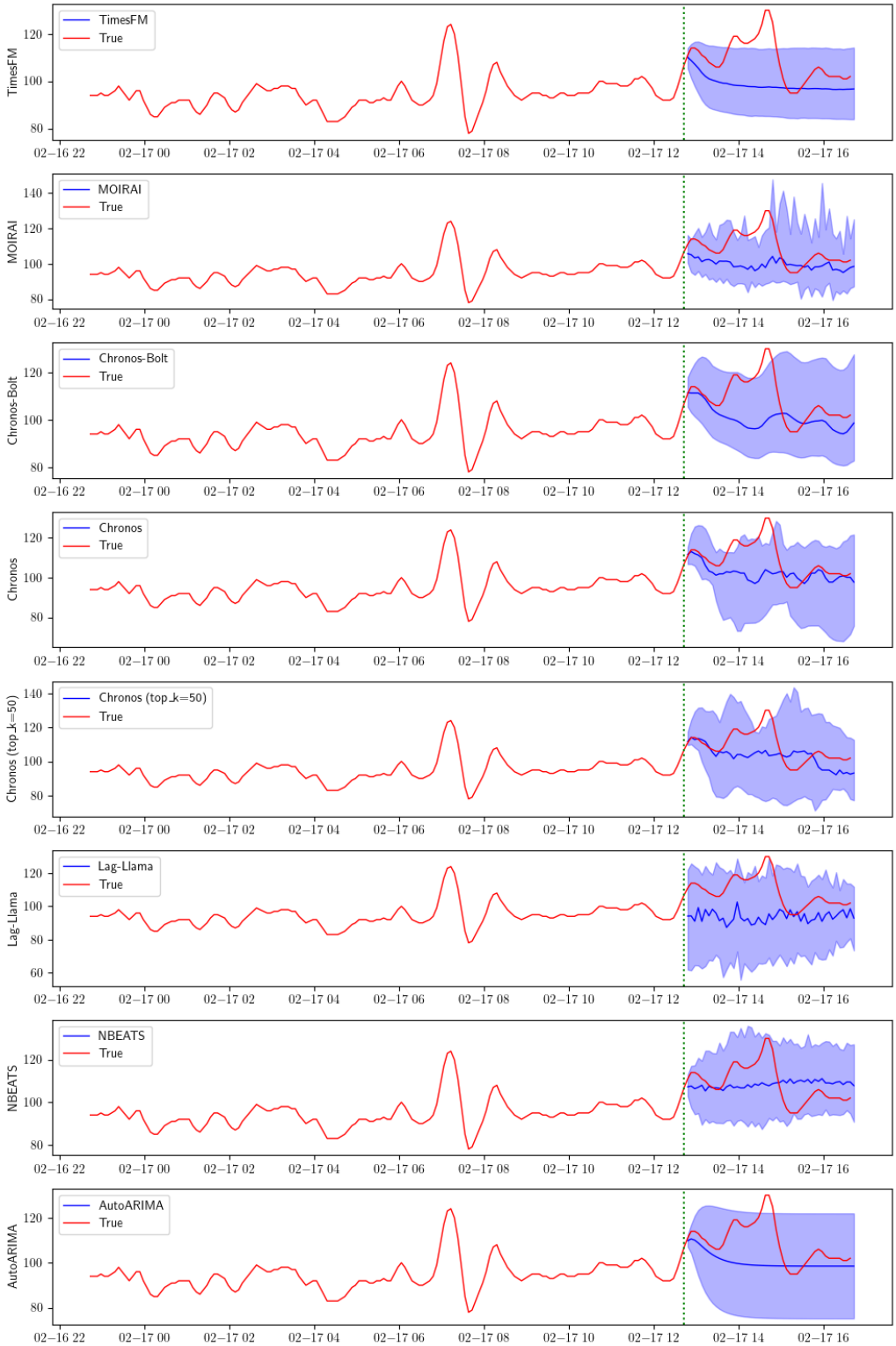

Figure 14. Model forecasts on Glucose dataset. Prediction timestamp randomly selected.

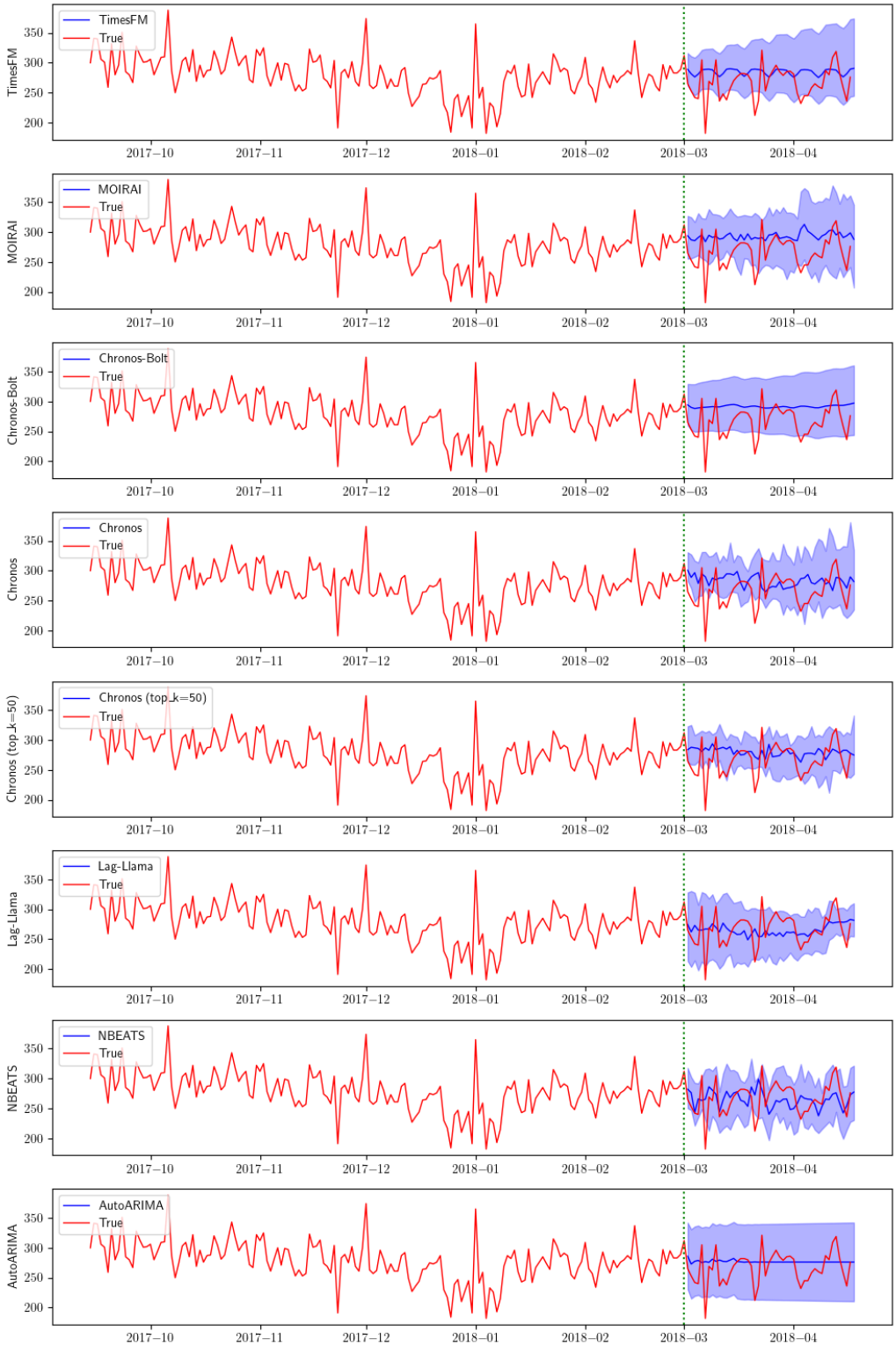

Figure 15. Model forecasts on Crime dataset. Prediction timestamp randomly selected.

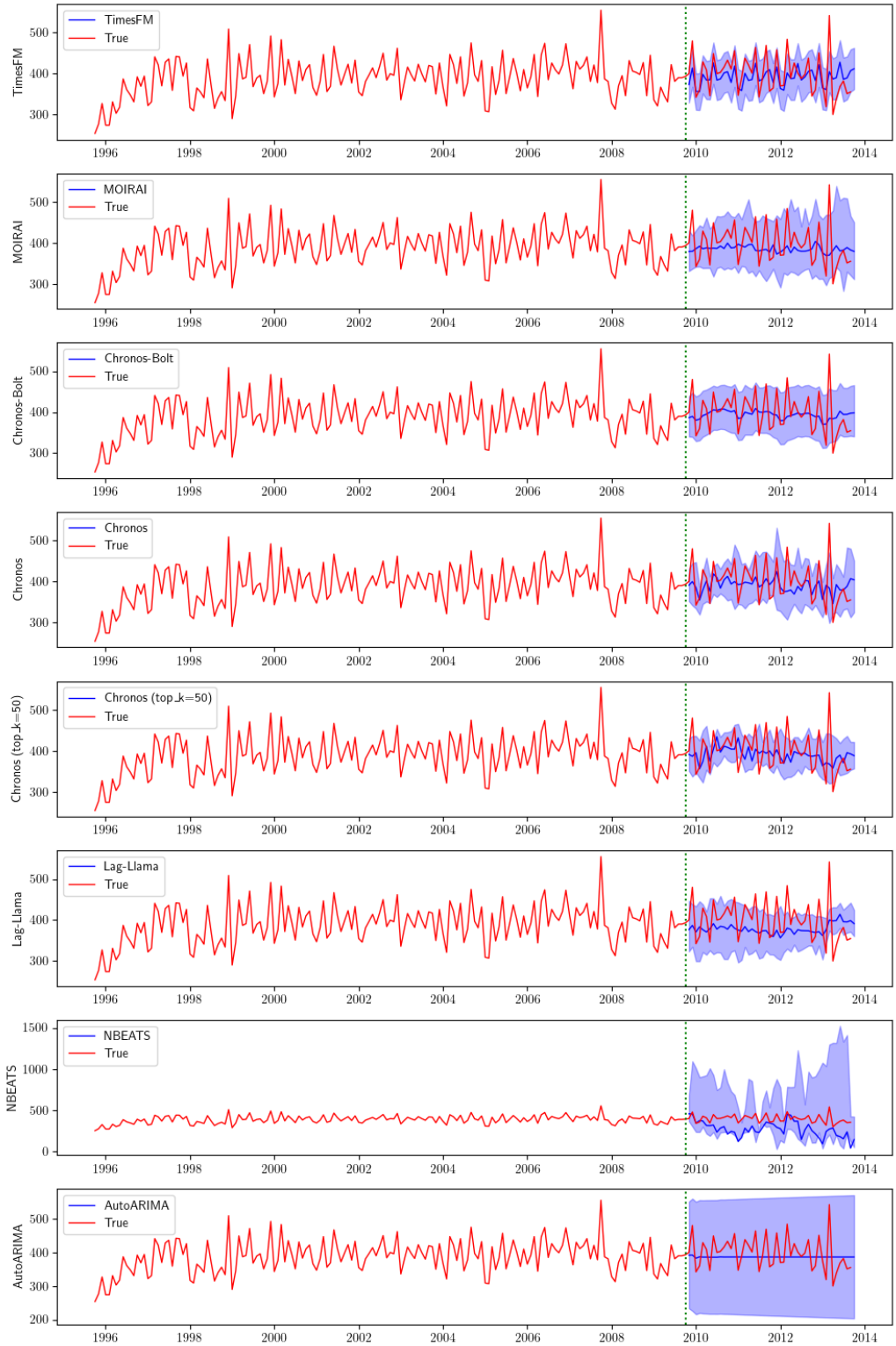

Figure 16. Model forecasts on Patents dataset. Prediction timestamp randomly selected.