Evaluating Long-Range Temporal Structure in Foundation Model-Based Forecasts of Heartbeat Dynamics

Adrian Serapio [1 2] **Bharath Ramsundar** [3] **Sandya Subramanian** [1 2]

Abstract

We examine the long-range temporal structure of forecasts produced by Time-Series Foundation Models (TSFMs) on heartbeat dynamics using the MIT-BIH Normal Sinus Rhythm Database (NSRDB). Our findings indicate that these models do not adequately capture long-range dependencies, as reflected in growing errors in RR-interval predictions over longer forecast horizons. Code is available at https://github.com/SubramanianLab/ecg-tsfm-benchmark.

1. Introduction

Time-Series Foundation Models (TSFMs) have demonstrated remarkable capabilities in forecasting diverse temporal data (Liang et al., 2024). By leveraging large-scale pretraining on various time-series datasets, TSFMs are capable of forecasting without task-specific training. This in-context learning capability offers a promising solution to regimes where there is a scarcity of available data, particularly in the healthcare domain, where high-quality longitudinal data is often difficult to acquire.

Although TSFMs demonstrate the potential to greatly advance physiological modeling, it remains unclear whether such models are actually capable of modeling the long-range structure of physiological data. Particularly, we chose to focus on evaluating heartbeat dynamics and investigate whether TSFMs merely reproduce local beat-to-beat patterns or capture physiologically plausible heart rate dynamics over longer time scales.

We evaluate the ability of TSFMs to model the long-range structure of heartbeat dynamics by forecasting Electrocardiogram (ECG) data. First, ECG is a very wide-spread modality given its importance in cardiac monitoring (Siontis et al., 2021). Long-range ECG forecasting enables longitudinal adverse cardiac event prediction including earlier detection of sudden cardiac death, ventricular arrhythmias, and myocardial infarction which can greatly reduce cardiac disease-related morbidity and improve survival outcomes (Hernesniemi et al., 2026; DiMarco & Philbrick, 1990; Siontis et al., 2021).

Second, ECG data is characterized by a complex multi-scale temporal hierarchy, where fine-grained morphological features are nested within broader physiological trends (Draghici & Taylor, 2016). The focus of our work goes beyond the local waveform morphology (milliseconds) and more to the long-range RR interval structure of ECG data (minutes to hours). Measuring TSFMs' ability to model RR interval structure is critical, as autonomic regulation signals emerge over longer time scales. RR interval forecasting provides a more direct probe of autonomic regulation and the body's physiological control system.

Although ECG-specific foundation models have been proposed (Li et al., 2025a;b; Tang et al., 2025), the lack of publicly available weights and their predominant focus on classification rather than forecasting led us to focus our analysis on general-purpose time-series foundation models.

Motivated by the above, we made the following contributions:

1. **Long-Range Physiological Signal Evaluation**: Our analysis focuses on evaluating the capacity of TSFMs to forecast RR intervals, representing long-range temporal structure in the minutes to hours range than simply local waveform-level (milliseconds)

2. **Physiological Distributional-Level Evaluation**: We evaluate TSFMs beyond the typical metrics used typically limited to MAE and RMSE. We also include Standard Deviation of Normal-to-Normal Intervals (SDNN) and Root Mean Square of Successive Differences (RMSSD) in order to evaluate if the TSFMs evaluate the fluctuations of the RR interval signal. Moreover, we also include a Kolmogorov-Smirnov (KS) goodness-of-fit metric as a metric to measure align-

[1]University of California, Berkeley [2]University of California, San Francisco [3]Deep Forest Sciences. Correspondence to: Sandya Subramanian <sandyas@berkeley.edu>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

ment with the underlying statistical generative process governing heartbeat dynamics (Subramanian & Ramsundar, 2025).

Our findings underscore the necessity of evaluating TSFMs through the lens of physiological validity before they can be reliably deployed for clinical forecasting.

2. Materials and Methods

2.1. Dataset

We use the MIT-BIH normal sinus rhythm dataset for benchmarking the TSFMs (The Beth Israel Deaconess Medical Center, 1990). This dataset contains 18 ECG recordings of subjects referred to the Arrhythmia Laboratory at Boston's Beth Israel Hospital (now the Beth Israel Deaconess Medical Center).

2.2. Time-Series Foundation Models

Although several foundation models have been developed for ECG and PPG data (Li et al., 2025a;b; Tang et al., 2025), most focus on classification or event detection rather than forecasting, and many lack publicly available weights, precluding inclusion in this study. While specialized time-series models exist (Nie et al., 2023), we instead evaluate the zero-shot forecasting capabilities of general-purpose Time-Series Foundation Models (TSFMs). This is motivated by the fact that modern TSFMs already incorporate patch-based transformer architectures (e.g., PatchTST), enabling modeling of long-range temporal dependencies. For comparison, we additionally include an LSTM as a non-foundation model baseline to contextualize the performance of the TSFMs.

We evaluate the following TSFMs (Table 1):

1. **TimesFM 2.5 (Google):** A decoder-only transformer using patching to efficiently capture long-range dependencies, pre-trained on large-scale synthetic and real-world time-series data (Das et al., 2024).
2. **Chronos-2 (Amazon):** Frames forecasting as a language modeling problem by quantizing continuous values into discrete tokens and predicting their distribution autoregressively (Ansari et al., 2025).
3. **Moirai 2.0 (Salesforce):** A lightweight decoder-only TSFM with single-patch input representation and quantile loss for probabilistic forecasting (Liu et al., 2026).

2.3. Evaluation Metrics

To assess the accuracy and distributional fidelity of the forecasted ECG signals, we employed several primary metrics. Let y_t denote the ground-truth signal and $\hat{y}_t$ the corresponding model prediction at time index t, where $t = 1, \ldots, H$ and H is the forecast horizon (number of predicted time points).

- **Mean Absolute Error (MAE):** Measures the average absolute magnitude of the forecast errors.

$$\text{MAE} = \frac{1}{H}\sum_{t=1}^{H}|y_t - \hat{y}_t| \tag{1}$$

- **Root Mean Squared Error (RMSE):** Provides a metric that disproportionately penalizes large outliers, which is critical for identifying significant deviations in cardiac cycle timing.

$$\text{RMSE} = \sqrt{\frac{1}{H}\sum_{t=1}^{H}(y_t - \hat{y}_t)^2} \tag{2}$$

- **Standard Deviation of NN intervals (SDNN):** Measures the overall variability of the RR intervals (normal-to-normal intervals), reflecting both short- and long-term components of heart rate variability.

$$\text{SDNN} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(RR_i - \bar{RR})^2} \tag{3}$$

  where RR_i denotes the i-th RR interval and $\bar{RR}$ is the mean RR interval. SDNN captures global variability and is sensitive to long-range temporal structure in the heartbeat dynamics.

- **Root Mean Square of Successive Differences (RMSSD):** Quantifies short-term variability by measuring the root mean square of differences between successive RR intervals.

$$\text{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2} \tag{4}$$

  RMSSD is primarily sensitive to high-frequency variations and reflects parasympathetic (vagal) activity, making it useful for assessing short-term temporal fidelity of the forecasted signals.

- **Kolmogorov-Smirnov (KS) Score:** We use the KS score as a Goodness-of-Fit (GOF) metric based on the Time Rescaling Theorem (Subramanian & Ramsundar, 2025). If a model correctly captures the heartbeat generative process, the transformed inter-event intervals z_i should be independent and identically distributed uniform random variables on the interval $(0, 1)$.

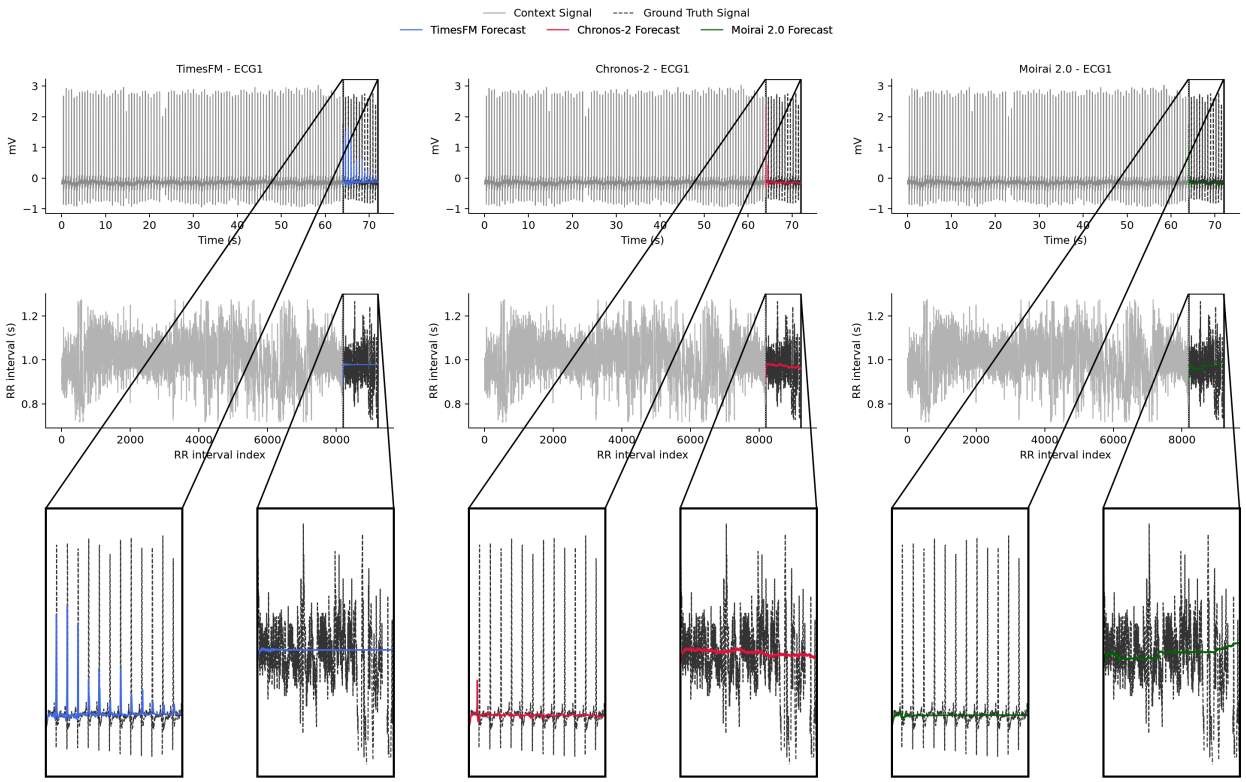

Figure 1. Time-Series Foundation Model (TSFM) forecasts of a single-subject's ECG data MIT-BIH NSRDB. Each panel compares the observed context window with the forecast horizon for ECG waveform and RR-interval prediction. TSFMs were given 8,192 samples of context and were tasked to forecast 1,024 future waveform or RR interval times.

Let τ_i denote the i-th observed RR interval and $\mathcal{H}_{i-1}$ the history prior to event i. If $F(\cdot|\mathcal{H}_{i-1})$ is the model's conditional cumulative distribution function for the next RR interval, then the time-rescaled variables are

$$z_i = F(\tau_i \mid \mathcal{H}_{i-1}).$$

The KS statistic is defined as the maximum deviation between the empirical cumulative distribution of the z_i values and the CDF of a uniform distribution:

$$KS = \max_{1 \le i \le N} \left| \frac{i - 0.5}{N} - z_{(i)} \right| \tag{5}$$

where $z_{(i)}$ are the sorted transformed intervals. A well-fit model yields a KS score that falls within the 95% significance cutoff, defined as:

$$\text{Cutoff}_{95\%} \approx \frac{1.36}{\sqrt{N}} \tag{6}$$

Table 1. Characteristics of Time-Series Foundation Models Benchmarked

	TimesFM-2.5	Chronos-2	Moirai 2.0
Number of Parameters	200M	120M	11.4M
Context Length	16,000	8,192	—
Maximum Pred. Length	—	1,024	—

3. Experimental Setup and Results

Figure 1 illustrates an example forecast for a particular patient. Qualitatively, we observe that the models cannot capture the long-range stochasticity that is reflected by an RR interval temporal signal, as opposed to the waveform where qualitatively, the TSFMs are capable of emulating.

For our first experiment, we evaluated RR interval forecasting performance across varying ECG context lengths and prediction horizons. This experiment tests whether longer historical context improves forecasting accuracy and whether errors accumulate as models forecast further into the future. We report subject-averaged MAE, RMSE, and KS statistics together with 95% bootstrap confidence inter-

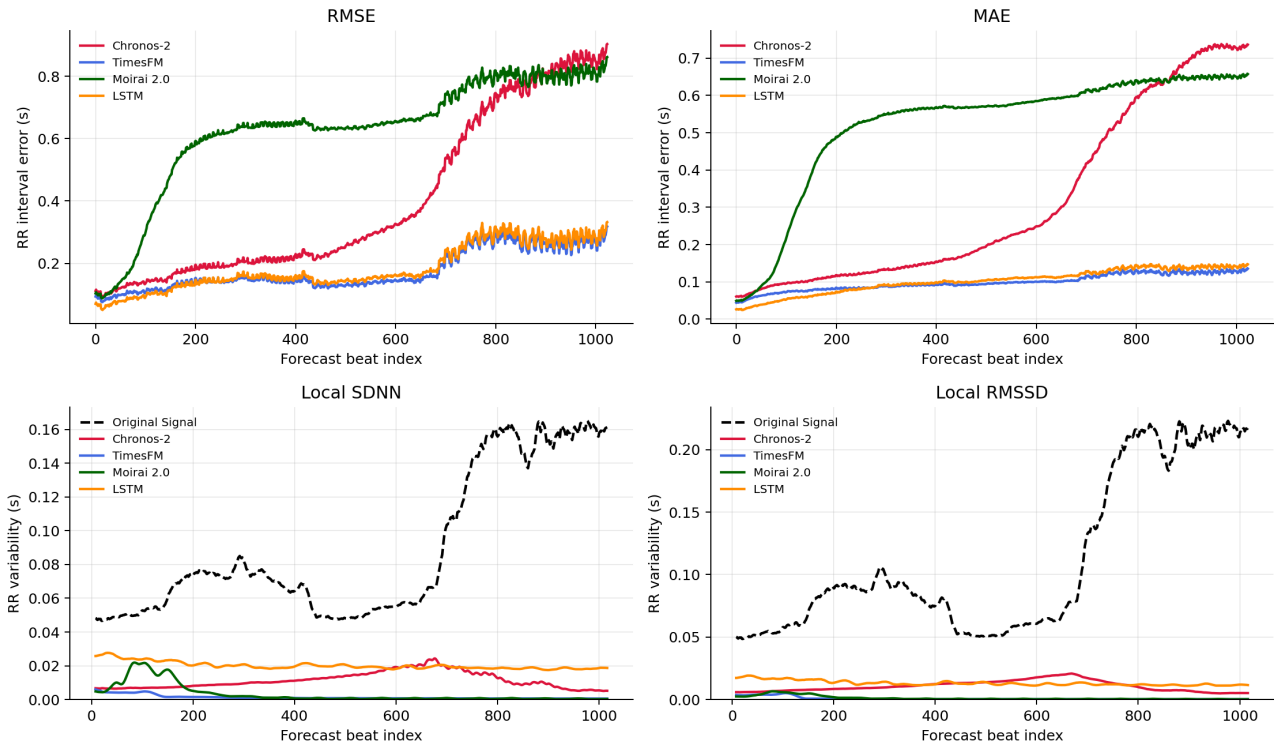

Figure 2. Per-beat evaluation of RR interval forecasting error and variability, averaged across the 18-subject cohort, for forecasts conditioned on the longest evaluated context window (8,192 samples). The RMSE and MAE are plotted as a function of forecast beat index. Local variability metrics (SDNN and RMSSD) are compared to the ground truth RR interval signal. (Note: Table 2 evaluates the model at different context windows, while Figure 2 evaluates at a single context window)

vals across the 18-subject cohort.

As shown in Table 2 (Appendix A), forecasting performance generally worsened as the prediction horizon increased. For example, Chronos-2 direct RR forecasting RMSE increased from 0.17 s (256-beat horizon, 8,192 context) to 0.59 s (1,024-beat horizon, 8,192 context), while TimesFM increased from 0.15 s to 0.34 s over the same horizons, indicating substantial error accumulation with longer forecasts. This degradation was most apparent in RMSE and MAE, particularly for Chronos-2 and Moirai 2.0, suggesting that pointwise prediction errors accumulate over longer horizons. In contrast, increasing the context length from 2,048 to 4,096 or 8,192 samples did not consistently improve performance across models or metrics. The conventional LSTM baseline achieved comparable RMSE and MAE across many settings, indicating that low pointwise forecasting error alone is insufficient to characterize physiologically meaningful long-range forecasting. These results suggest that simply providing additional historical ECG context is insufficient for reliably improving long-range RR interval forecasting.

For our next experiment, we compared two approaches for RR interval forecasting: direct prediction and extraction from model-generated ECG waveforms. Extracted RR intervals often achieved lower RMSE and MAE, indicating better pointwise accuracy. However, this advantage did not consistently extend to the KS metric. In several cases, direct prediction yielded lower KS despite higher RMSE/MAE, while in others the reverse held, suggesting a tradeoff between pointwise accuracy and distributional fidelity.

Lastly, we further analyzed error as a function of forecast beat index to characterize temporal degradation. As shown in Figure 2, all models exhibit increasing error, but with distinct dynamics. TimesFM shows gradual error growth, whereas Moirai 2.0 degrades rapidly and saturates early, and Chronos-2 exhibits delayed but sharp escalation at longer horizons. Critically, both TimesFM and Moirai 2.0 collapse in local variability (SDNN, RMSSD), producing near-zero RR variability relative to the ground truth. Moreover, all three models fail to adequately capture physiological variability of the ECG signal. This indicates that despite reasonable short-horizon accuracy, these models fail to preserve

physiologically meaningful dynamics.

4. Discussion

The results highlight both the promise and limitations of Time-Series Foundation Models (TSFMs) for physiological data. While these models generate realistic short-term ECG waveforms, TSFM forecasting capabilities does not necessarily extend to long-range temporal scales. Analysis of RR interval dynamics reveals that performance degrades when modeling long-range structure.

A central question is whether TSFMs capture physiologically meaningful dynamics or primarily reproduce local statistical patterns. We observe that the models appear to rely on beat-level morphology rather than learning long-range cardiac behavior. This distinction is critical, as many clinically relevant processes such as autonomic regulation emerge over longer time scales. RR intervals provide a more direct probe of this limitation. Unlike waveform morphology, which can be matched locally, RR interval structure reflects integrated physiological control. Failure to model these dynamics indicates limited ability to capture meaningful temporal hierarchies.

These results have important implications for clinical deployment. Although TSFMs are attractive in low-data settings, caution is warranted for tasks requiring long-range physiological consistency.

Future work should prioritize evaluation frameworks that explicitly test long-range temporal modeling, such as RR interval forecasting. Moreover, other metric such as the Fréchet distance (Tang et al., 2025) could further evaluate the statistical validity and physiological fidelity of model forecasts. Improving performance may require architectural biases, extended context windows, and training objectives that enforce multi-scale temporal structure in physiological signals.

References

Ansari, A. F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S. S., Shen, H., Stella, L., Zhang, X., Goswami, M., Kapoor, S., Maddix, D. C., Guerron, P., Hu, T., Yin, J., Erickson, N., Desai, P. M., Wang, H., Rangwala, H., Karypis, G., Wang, Y., and Bohlke-Schneider, M. Chronos-2: from univariate to universal forecasting, October 2025. URL http://arxiv.org/abs/2510.15821. arXiv:2510.15821.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting, April 2024. URL http://arxiv.org/abs/2310.10688. arXiv:2310.10688.

DiMarco, J. P. and Philbrick, J. T. Use of ambulatory electrocardiographic (Holter) monitoring. *Annals of Internal Medicine*, 113(1):53–68, July 1990. ISSN 0003-4819. doi: 10.7326/0003-4819-113-1-53.

Draghici, A. E. and Taylor, J. A. The physiological basis and measurement of heart rate variability in humans. *Journal of Physiological Anthropology*, 35(1): 22, September 2016. ISSN 1880-6805. doi: 10.1186/s40101-016-0113-7.

Hernesniemi, J. A., Pukkila, T., Rankinen, J., Kallonen, A., Uimonen, M., Lyytikäinen, L.-P., Nikus, K., Räsänen, E., and Tynkkynen, J. Performance of the 12-lead ECG in predicting short- and long-term risk of sudden cardiac death. *npj Digital Medicine*, 9(1): 317, March 2026. ISSN 2398-6352. doi: 10.1038/s41746-026-02456-1. URL https://www.nature.com/articles/s41746-026-02456-1.

Li, H., Deng, B., Xu, C., Feng, Z., Schlegel, V., Huang, Y.-H., Sun, Y., Sun, J., Yang, K., Yu, Y., and Bian, J. Mira: medical time series foundation model for real-world health data. October 2025a. URL https://openreview.net/forum?id=Auy2DmlJBO.

Li, J., Aguirre, A. D., Junior, V. M., Jin, J., Liu, C., Zhong, L., Sun, C., Clifford, G., Brandon Westover, M., and Hong, S. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7), June 2025b. ISSN 2836-9386. doi: 10.1056/AIoa2401033. URL https://ai.nejm.org/doi/10.1056/AIoa2401033.

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, Barcelona Spain, August 2024. ACM. ISBN 9798400704901. doi: 10.1145/3637528.3671451. URL https://dl.acm.org/doi/10.1145/3637528.3671451.

Liu, C., Aksu, T., Liu, J., Liu, X., Yan, H., Pham, Q., Savarese, S., Sahoo, D., Xiong, C., and Li, J. Moirai 2. 0: when less is more for time series forecasting, February 2026. URL http://arxiv.org/abs/2511.11698. arXiv:2511.11698.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: long-term forecasting with transformers, March 2023. URL http://arxiv.org/abs/2211.14730. arXiv:2211.14730.

Siontis, K. C., Noseworthy, P. A., Attia, Z. I., and Friedman, P. A. Artificial intelligence-enhanced

electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, July 2021. ISSN 1759-5010. doi: 10.1038/s41569-020-00503-2. URL https://www.nature.com/articles/s41569-020-00503-2.

Subramanian, S. and Ramsundar, B. Density-based neural temporal point processes for heartbeat dynamics, November 2025. URL http://arxiv.org/abs/2511.22096. arXiv:2511.22096.

Tang, Z., Qi, J., Zheng, Y., and Huang, J. A comprehensive benchmark for electrocardiogram time-series. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6490–6499, Dublin Ireland, October 2025. ACM. ISBN 9798400720352. doi: 10.1145/3746027.3754729. URL https://dl.acm.org/doi/10.1145/3746027.3754729.

The Beth Israel Deaconess Medical Center, T. A. L. The mit-bih normal sinus rhythm database, 1990. URL https://physionet.org/content/nsrdb/.

A. RR Forecast Context Length and Prediction Horizon Experiment

Table 2. Subject-averaged forecasting performance for RR interval prediction across the 18-subject cohort, context lengths, and prediction horizons. Direct denotes direct RR interval prediction; extracted denotes RR intervals derived from model-predicted waveforms via peak extraction. Point estimates are the subject-averaged mean; each is followed by its 95% bootstrap confidence interval (lower, upper) over the 18 subjects.

Context Length	Prediction Horizon	Output Type	Model	RMSE (s) ↓	RMSE 95% CI	MAE (s) ↓	MAE 95% CI	KS ↓	KS 95% CI
2,048	256	Direct	Chronos-2	0.24	(0.17, 0.33)	0.14	(0.10, 0.18)	0.43	(0.38, 0.47)
2,048	256	Direct	TimesFM	0.20	(0.12, 0.29)	0.11	(0.07, 0.14)	0.49	(0.42, 0.57)
2,048	256	Direct	Moirai 2.0	0.49	(0.44, 0.55)	0.37	(0.33, 0.42)	0.48	(0.42, 0.53)
2,048	256	Direct	LSTM	0.18	(0.10, 0.28)	0.10	(0.06, 0.14)	—	—
2,048	256	Extracted	Chronos-2	0.14	(0.11, 0.18)	0.08	(0.06, 0.11)	0.50	(0.48, 0.52)
2,048	256	Extracted	TimesFM	0.16	(0.14, 0.19)	0.10	(0.08, 0.12)	0.51	(0.49, 0.53)
2,048	256	Extracted	Moirai 2.0	0.16	(0.14, 0.19)	0.11	(0.09, 0.14)	0.44	(0.43, 0.45)
2,048	256	Extracted	LSTM	0.24	(0.19, 0.30)	0.19	(0.14, 0.24)	—	—
2,048	512	Direct	Chronos-2	0.41	(0.32, 0.56)	0.26	(0.22, 0.30)	0.46	(0.43, 0.49)
2,048	512	Direct	TimesFM	0.23	(0.12, 0.39)	0.10	(0.07, 0.13)	0.44	(0.37, 0.52)
2,048	512	Direct	Moirai 2.0	0.63	(0.54, 0.77)	0.51	(0.46, 0.56)	0.61	(0.57, 0.64)
2,048	512	Direct	LSTM	0.19	(0.09, 0.35)	0.08	(0.06, 0.10)	—	—
2,048	512	Extracted	Chronos-2	0.20	(0.16, 0.25)	0.10	(0.08, 0.13)	0.46	(0.44, 0.48)
2,048	512	Extracted	TimesFM	0.20	(0.16, 0.24)	0.10	(0.08, 0.12)	0.55	(0.53, 0.58)
2,048	512	Extracted	Moirai 2.0	0.24	(0.16, 0.34)	0.12	(0.09, 0.16)	0.49	(0.47, 0.51)
2,048	512	Extracted	LSTM	0.31	(0.21, 0.44)	0.21	(0.14, 0.30)	—	—
2,048	1,024	Direct	Chronos-2	0.66	(0.50, 0.88)	0.42	(0.37, 0.46)	0.51	(0.48, 0.54)
2,048	1,024	Direct	TimesFM	0.34	(0.15, 0.60)	0.11	(0.09, 0.14)	0.40	(0.35, 0.45)
2,048	1,024	Direct	Moirai 2.0	0.68	(0.52, 0.90)	0.44	(0.38, 0.50)	0.67	(0.63, 0.70)
2,048	1,024	Direct	LSTM	0.32	(0.13, 0.59)	0.10	(0.07, 0.12)	—	—
2,048	1,024	Extracted	Chronos-2	0.35	(0.27, 0.45)	0.14	(0.11, 0.16)	0.44	(0.42, 0.46)
2,048	1,024	Extracted	TimesFM	0.25	(0.20, 0.33)	0.10	(0.08, 0.12)	0.56	(0.53, 0.58)
2,048	1,024	Extracted	Moirai 2.0	0.41	(0.25, 0.62)	0.16	(0.10, 0.23)	0.52	(0.50, 0.54)
2,048	1,024	Extracted	LSTM	0.28	(0.18, 0.41)	0.15	(0.10, 0.19)	—	—
4,096	256	Direct	Chronos-2	0.20	(0.14, 0.27)	0.10	(0.08, 0.12)	0.38	(0.35, 0.40)
4,096	256	Direct	TimesFM	0.15	(0.08, 0.22)	0.07	(0.06, 0.09)	0.46	(0.39, 0.53)
4,096	256	Direct	Moirai 2.0	0.52	(0.46, 0.58)	0.41	(0.36, 0.45)	0.52	(0.48, 0.57)
4,096	256	Direct	LSTM	0.14	(0.08, 0.22)	0.06	(0.05, 0.07)	—	—
4,096	256	Extracted	Chronos-2	0.14	(0.11, 0.17)	0.08	(0.06, 0.11)	0.49	(0.47, 0.51)
4,096	256	Extracted	TimesFM	0.16	(0.14, 0.18)	0.10	(0.08, 0.12)	0.51	(0.49, 0.53)
4,096	256	Extracted	Moirai 2.0	0.15	(0.13, 0.18)	0.11	(0.09, 0.14)	0.44	(0.43, 0.45)
4,096	256	Extracted	LSTM	0.21	(0.16, 0.26)	0.16	(0.11, 0.21)	—	—
4,096	512	Direct	Chronos-2	0.36	(0.27, 0.51)	0.21	(0.18, 0.24)	0.42	(0.39, 0.45)
4,096	512	Direct	TimesFM	0.19	(0.09, 0.35)	0.08	(0.06, 0.11)	0.42	(0.35, 0.49)
4,096	512	Direct	Moirai 2.0	0.64	(0.55, 0.78)	0.52	(0.47, 0.58)	0.63	(0.60, 0.67)
4,096	512	Direct	LSTM	0.19	(0.09, 0.35)	0.07	(0.06, 0.10)	—	—
4,096	512	Extracted	Chronos-2	0.23	(0.17, 0.29)	0.10	(0.08, 0.13)	0.45	(0.43, 0.47)
4,096	512	Extracted	TimesFM	0.20	(0.16, 0.25)	0.10	(0.09, 0.12)	0.55	(0.53, 0.57)
4,096	512	Extracted	Moirai 2.0	0.23	(0.16, 0.33)	0.12	(0.09, 0.15)	0.49	(0.46, 0.51)
4,096	512	Extracted	LSTM	0.24	(0.17, 0.30)	0.17	(0.12, 0.22)	—	—
4,096	1,024	Direct	Chronos-2	0.66	(0.50, 0.89)	0.41	(0.37, 0.46)	0.50	(0.47, 0.54)
4,096	1,024	Direct	TimesFM	0.33	(0.14, 0.60)	0.10	(0.08, 0.13)	0.39	(0.35, 0.43)
4,096	1,024	Direct	Moirai 2.0	0.69	(0.52, 0.92)	0.44	(0.39, 0.51)	0.67	(0.64, 0.70)
4,096	1,024	Direct	LSTM	0.33	(0.14, 0.61)	0.10	(0.08, 0.12)	—	—
4,096	1,024	Extracted	Chronos-2	0.39	(0.31, 0.48)	0.16	(0.13, 0.19)	0.44	(0.42, 0.46)
4,096	1,024	Extracted	TimesFM	0.24	(0.18, 0.30)	0.10	(0.08, 0.12)	0.55	(0.52, 0.57)
4,096	1,024	Extracted	Moirai 2.0	0.34	(0.22, 0.49)	0.13	(0.10, 0.18)	0.51	(0.49, 0.53)
4,096	1,024	Extracted	LSTM	0.28	(0.21, 0.34)	0.16	(0.12, 0.20)	—	—

Continued on next page

Table 2. (continued) Subject-averaged forecasting performance for RR interval prediction across the 18-subject cohort, context lengths, and prediction horizons. Direct denotes direct RR interval prediction; extracted denotes RR intervals derived from model-predicted waveforms via peak extraction. Point estimates are subject-averaged means; each is followed by its 95% bootstrap confidence interval over the 18 subjects.

Context Length	Prediction Horizon	Output Type	Model	RMSE (s) ↓	RMSE 95% CI	MAE (s) ↓	MAE 95% CI	KS ↓	KS 95% CI
8,192	256	Direct	Chronos-2	0.17	(0.11, 0.25)	0.08	(0.06, 0.10)	0.33	(0.30, 0.37)
8,192	256	Direct	TimesFM	0.15	(0.09, 0.23)	0.07	(0.05, 0.08)	0.40	(0.34, 0.47)
8,192	256	Direct	Moirai 2.0	0.55	(0.49, 0.62)	0.44	(0.40, 0.49)	0.56	(0.51, 0.61)
8,192	256	Direct	LSTM	0.15	(0.09, 0.23)	0.06	(0.05, 0.08)	—	—
8,192	256	Extracted	Chronos-2	0.19	(0.13, 0.28)	0.09	(0.07, 0.12)	0.47	(0.45, 0.49)
8,192	256	Extracted	TimesFM	0.21	(0.15, 0.28)	0.10	(0.08, 0.13)	0.49	(0.48, 0.51)
8,192	256	Extracted	Moirai 2.0	0.22	(0.15, 0.30)	0.13	(0.10, 0.15)	0.44	(0.42, 0.46)
8,192	256	Extracted	LSTM	0.28	(0.20, 0.35)	0.18	(0.13, 0.23)	—	—
8,192	512	Direct	Chronos-2	0.24	(0.14, 0.40)	0.12	(0.09, 0.15)	0.38	(0.34, 0.41)
8,192	512	Direct	TimesFM	0.18	(0.09, 0.35)	0.08	(0.06, 0.11)	0.39	(0.33, 0.46)
8,192	512	Direct	Moirai 2.0	0.66	(0.57, 0.80)	0.55	(0.51, 0.61)	0.69	(0.66, 0.72)
8,192	512	Direct	LSTM	0.20	(0.10, 0.36)	0.09	(0.06, 0.12)	—	—
8,192	512	Extracted	Chronos-2	0.23	(0.19, 0.28)	0.11	(0.09, 0.14)	0.45	(0.43, 0.47)
8,192	512	Extracted	TimesFM	0.20	(0.16, 0.24)	0.10	(0.08, 0.12)	0.54	(0.52, 0.56)
8,192	512	Extracted	Moirai 2.0	0.18	(0.14, 0.22)	0.10	(0.08, 0.12)	0.47	(0.45, 0.50)
8,192	512	Extracted	LSTM	0.30	(0.19, 0.44)	0.20	(0.12, 0.30)	—	—
8,192	1,024	Direct	Chronos-2	0.59	(0.40, 0.85)	0.31	(0.27, 0.35)	0.47	(0.43, 0.52)
8,192	1,024	Direct	TimesFM	0.34	(0.13, 0.63)	0.10	(0.08, 0.12)	0.37	(0.34, 0.41)
8,192	1,024	Direct	Moirai 2.0	0.76	(0.58, 1.00)	0.52	(0.46, 0.59)	0.74	(0.70, 0.77)
8,192	1,024	Direct	LSTM	0.35	(0.14, 0.64)	0.10	(0.08, 0.13)	—	—
8,192	1,024	Extracted	Chronos-2	0.41	(0.32, 0.50)	0.16	(0.13, 0.20)	0.44	(0.42, 0.47)
8,192	1,024	Extracted	TimesFM	0.25	(0.20, 0.31)	0.11	(0.09, 0.12)	0.54	(0.52, 0.56)
8,192	1,024	Extracted	Moirai 2.0	0.29	(0.18, 0.43)	0.12	(0.08, 0.17)	0.49	(0.47, 0.52)
8,192	1,024	Extracted	LSTM	0.35	(0.18, 0.65)	0.18	(0.11, 0.28)	—	—