Context Saturation in Zero-Shot Time-Series Foundation Models

Miguel Nogales [1] **Luca Butera** [1] **Alberto Ferrante** [1] **Cesare Alippi** [1 2]

Abstract

Despite *time series foundation model*s (TSFMs) supporting variable input lengths, they are typically evaluated using the longest possible input length, subject to data availability and model input capacity. This practice risks conflating different factors impacting performance, and leaves end-users lacking principled guidelines for input length selection. To study the relationship between input length W and model performance, we introduce the *context saturation length* (CSL): the minimum input length required to achieve a target fraction of a model's peak forecasting performance. For time series with a dominant seasonal period P, we show that context saturation can be reliably achieved with at least $W \approx 2P$. This is supported by results on both synthetic and real-world benchmarks, where *performance-to-input-length* curves align when considering the *period-normalized* input length W/P. Additionally, we demonstrate similar behavior on time series generated by auto-regressive processes, when normalizing the input length by the process memory length. Our findings demonstrate that, for periodic time series where the dominant period can be reliably estimated, the input length can be selected a-priori, avoiding hyper-parameter search, with negligible sacrifice of performance. Moreover, our work suggests that model-agnostic methodologies, based on the inherent characteristics of the target time series, can provide practical guidelines for input length selection, and lead to the design of principled benchmarks to evaluate TSFMs.

[1]Universittà della svizzera italiana, Lugano, Switzerland [2]Politecnico di Milano, Milano, Italy. Correspondence to: Miguel Nogales <miguel.nogales@usi.ch>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

*Time series foundation model*s (TSFMs) (Ansari et al., 2024; Woo et al., 2024) have recently emerged as a powerful paradigm for *zero-shot* time series forecasting, enabling a single *pre-trained* model to generalize across diverse temporal domains. A central design dimension in time series forecasting models is the input context length W, which determines how much historical information is provided to generate future predictions. Recent work on time series models has largely focused on extending the context length, motivated by the success of long-context architectures in other domains (Peng et al., 2024), and the assumption that more historical data should yield better forecasts (Zhang et al., 2025). As such, TSFMs are often designed to operate with larger context length than previous forecasting approaches (Ansari et al., 2025; Shi et al., 2025) and, while they can accept variable input lengths, current evaluation protocols emphasize performance under large contexts, typically using the maximum allowed by model design and data availability (Aksu et al., 2024).

At deployment time, context length is thus usually selected via empirical evaluation or fixed a-priori, with limited understanding of how the characteristics of the target time series contribute to determining the most effective input length. These observations lead to fundamental questions: when does additional context cease to provide meaningful gains? Can this threshold be estimated a-priori based on the characteristics of the target series? Addressing these questions can lead to more efficient context utilization and guide the design of principled evaluation frameworks for TSFMs.

In this work, we argue that more context is not universally useful, but instead its importance depends on the data-generating process of the target time series. We introduce the notion of *context saturation length* (CSL), defined as the minimal input length required to achieve a target fraction of a TSFM peak performance, and use it to characterize how forecasting accuracy evolves as a function of context length. Our central hypothesis is that CSL follows a process-dependent scaling law, and can be related to some intrinsic property of the target time series. This intuition allows us to move beyond retrospective performance comparisons and toward a data-centric understanding of context length requirements.

For time series with strong seasonal structure and dominant period P, we empirically show that CSL predictably depends on P, and corresponds to using an input length of at least $W \approx 2P$. For non-seasonal time series, we demonstrate a similar effect on *autoregressive* (AR) processes with memory p, where CSL relates to the memory size. We validate these findings through a controlled synthetic benchmark, and corroborate them on real-world datasets by relating CSL behavior to the estimated periodicity. Our results highlight how context requirements can be estimated in a TSFM-agnostic manner, and additional context is not universally beneficial. This work provides a first step toward data-centric context length selection and the design of controlled benchmarks for the evaluation of temporal inductive biases in TSFMs.

2. Problem Formulation

Let $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$ be a dataset of univariate time series, where $x^{(i)} = \{x_t^{(i)}\}_{t=1}^{T_i}$. For a forecasting horizon H, a TSFM f produces H-step-ahead predictions from the last W observations:

$$\hat{x}^{(i)}_{T_i+1:T_i+H}(W) = f\left(x^{(i)}_{T_i-W+1:T_i}; \theta\right), \tag{1}$$

where θ denotes the pre-trained model parameters. For each model, we evaluate a finite set of admissible context lengths $\mathcal{W}_f = \{W_1, \ldots, W_K\}$, with $W_K = W_{\max}(f)$.

Forecast quality is measured using standard point-forecast metrics $m \in \mathcal{M}$, including *mean absolute error* (MAE), *mean squared error* (MSE), *mean absolute scaled error* (MASE), and *symmetric mean absolute percentage error* (SMAPE). We denote by $\mathcal{E}_m(W; x^{(i)}, f)$ the error obtained by model f, on series $x^{(i)}$, for metric m, with context length W. Lower values indicate better performance.

We expect that increasing the context length reaches a point of diminishing returns, after which it can even hurt performance. To quantify this threshold, we define the *context saturation length* (CSL) in the following way. For each series, let $\mathcal{E}^\star_m(x^{(i)}, f) = \min_{W \in \mathcal{W}_f} \mathcal{E}_m(W; x^{(i)}, f)$ be the minimum error achieved over $\mathcal{W}_f$. We consider the ratio between this minimum error and the error achieved at a given context length W, resulting in $\mathcal{E}^\star_m(x^{(i)}, f)/\mathcal{E}_m(W; x^{(i)}, f) \in (0, 1]$. For a target threshold $\alpha \in (0, 1]$ (e.g., $\alpha = 0.95$), CSL $W_{\alpha,m}(x^{(i)}, f)$ is defined as the smallest context length that achieves a performance ratio of at least α:

$$W_{\alpha,m}(x^{(i)}, f) = \min\left\{ W \in \mathcal{W}_f \,\middle|\, \frac{\mathcal{E}^\star_m(x^{(i)}, f)}{\mathcal{E}_m(W; x^{(i)}, f)} \geq \alpha \right\}. \tag{2}$$

This definition is metric-agnostic, inherently normalized across diverse series, accommodates non-monotonic error curves, and straightforwardly captures the point beyond which longer contexts provide diminishing returns.

Eq. 2 can be extended to the dataset level by considering the aggregated (e.g., mean or median) error across time series in $\mathcal{D}$. We denote this with $W_{\alpha,m}(\mathcal{D}, f)$. This yields a summary statistic for each model and metric, while the series-wise CSL distribution captures heterogeneity across time series in $\mathcal{D}$.

Given the above definitions, our primary objective is to empirically assess whether CSL can be determined a-priori (i.e., without evaluating any specific TSFM) based on some intrinsic property of the process generating the target time series. For instance, in periodic data, we hypothesize CSL can be related to the dominant seasonality. Conversely, for AR processes, we can expect it to correlate with the system's memory size in the sense of Markov order.

3. Experiments

We conduct our experiments on three recent TSFMs: *TimesFM2.5* (Das et al., 2023), *TimeMoE* (Shi et al., 2025), and *Chronos2* (Ansari et al., 2025). For the latter, we consider two variants: *Chronos2-Small* and *Chronos2-Synth*. This selection spans key design dimensions within the TSFM landscape: autoregressive modeling (TimesFM2.5 and TimeMoE), non-autoregressive modeling (Chronos2), sparse Mixture-of-Experts architectures (TimeMoE), and training exclusively on synthetic data (Chronos2-Synth).

3.1. Synthetic Benchmarks

To test the hypotheses of Sec. 2 we first experiment in controlled environments, where the data-generating process is known by construction. We consider four complementary families of data-generating processes, where the scale of temporal dependencies can be controlled parametrically.

- **Noisy Sine.** The i-th time series follows

$$x_t^{(i)} = \sin\left(\frac{2\pi t}{P^{(i)}}\right) + \varepsilon_t.$$

  Each time series has its own seasonality with period $P^{(i)}$.

- **Harmonic Sweep.** The i-th time series follows

$$x_t^{(i)} = \sin\left(\frac{2\pi t}{P^{(i)}}\right) + \lambda_1 \sin\left(u \frac{2\pi t}{P^{(i)}}\right) + \varepsilon_t.$$

  a more complex scenario, with an extra harmonic.

- **AR.** The i-th time series is generated by a stationary AR processes of order $p^{(i)}$:

$$x_t^{(i)} = \sum_{k=1}^{p^{(i)}} \varphi_k^{(i)} x_{t-k}^{(i)} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

The resulting series are not periodic, and the memory depth $p^{(i)}$ controls the scale of temporal dependency.

- **Linear Trend.** The i-th time series follows

$$x_t^{(i)} = a^{(i)}\, t + \varepsilon_t.$$

The resulting time series are linear trends, and we consider them as having a temporal scale coefficient of 1 by construction. More details given in App. A.

Protocol. To assess whether CSL has a predictable relationship with the parameters governing the scale of temporal dependencies across the considered data-generating process families, we carry out the following experiment for each family: (i) we generate a dataset with time series spanning multiple values for the time scale parameter (i.e., $P^{(i)}$ or $p^{(i)}$); (ii) we plot each TSFM dataset-averaged error against the time scale-normalized input length (i.e., $W/P^{(i)}$ or $W/p^{(i)}$); (iii) we compute the dataset-level CSL, as defined in Sec. 2. For *Noisy Sine* and *Harmonic Sweep*, we consider $P^{(i)} \in \{64, 96, 128, 192, 256, 384\}$. For *AR*, we consider $p^{(i)} \in \{2, 3, ..., 12\}$. For *Linear Trend* $p^{(i)} = 1$ by construction. We evaluate the error over a dataset-specific grid of values for W (details in App. A), considering an horizon of $H = 32$ time steps for all the experiments.

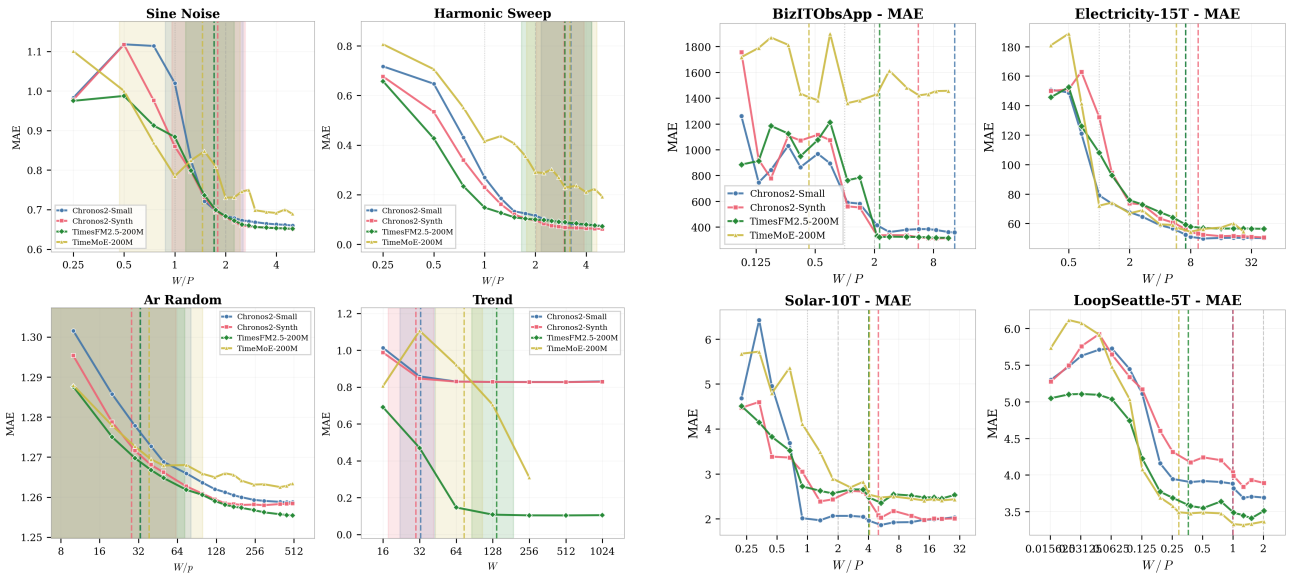
Results. The results of the experiments for the MAE metric are reported in Fig. 1a (see App. A for results on additional metrics, which are in general very close to the ones seen with MAE). For the periodic families (i.e., *Noisy Sine*, *Harmonic Sweep*), three regimes emerge as a function of $W/P^{(i)}$: (i) for $W/P^{(i)} < 1$, error remains high; (ii) for $1 \leq W/P^{(i)} \leq 2$, performance improves quasi-linearly; (iii) for $W/P^{(i)} > 2$, error plateaus. For TimeMoE, we additionally observe an oscillating behavior in the error curves, but the context saturation is present regardless. For the *AR* processes we observe a smoother performance improvement pattern, and CSL at $W/p^{(i)} \geq 32$. For the *Linear Trend* ($p^{(i)} = 1$), we observe a similar behavior to AR processes, despite in this case there is a bigger sensibility to model choice. With Chronos the CSL is at $W/p^{(i)} \geq 32$, similar to the AR case, while with the saturation is $W/p^{(i)} \geq 128$. Also, TimeMoE presents a big spike in error from context length 512 onward, so those values have been removed for the clarity of the figure. These results underscore how, across different TSFMs, we can observe consistent relationships between CSL and the parameter governing the scale of temporal dependencies. The plots also show the CSL ($\alpha = 0.95$) for each model as a vertical line of matching color. In the case of the *AR*, as the improvement is slight, the CSL is set to ($\alpha = 0.99$). In the *Noisy Sine* case we can observe a saturation close to the $W \approx 2P$ mark, while the *Harmonic Sweep* saturates around $W \approx 3P$, possibly due to the additional complexity introduced by the extra frequency. In the other cases (i.e., *AR* and *Linear Trend*), the variability is higher, possibly hinting at the fact that the temporal scale is not the only factor at play, but also the value of the parameters governing the generating process.

3.2. Real-World Case Studies

To test whether the CSL behavior observed on synthetic data persists on real-world seasonal time series, we complement the controlled benchmarks with four real-world case studies from the GiftEval (Aksu et al., 2024) TSFM benchmark: *Electricity* (energy consumption), *BizITObsApp* (IT applications observability metrics), *LoopSeattle* (traffic speed), and *Solar* (incident solar radiation) (details in App. B). Unlike the synthetic signals, these datasets are not generated from a known parametric process: they contain amplitude variation, local irregularities, and other non-ideal effects.

Protocol. Following Sec. 3, we associate each time series with a dominant period $P^{(i)}$ and evaluate forecasting performance against the time scale-normalized context length $W/P^{(i)}$. As $P^{(i)}$ is unknown for real-world data, we estimate it using the *autocorrelation function* (ACF), following the standard Box-Jenkins approach (Box et al., 2015) (ACF plots and estimated values for $P^{(i)}$ are reported in App. B). For each real-world dataset, we again evaluate TSFMs error against normalized input length $W/P^{(i)}$, using a dataset-dependent grid of values for W (details in App. B). This setup allows direct comparison with the synthetic benchmarks: if the rule identified for periodic time series in the previous section is relevant beyond synthetic signals, then the CSL should again fall around $W \approx 2P$.

Results. The results of the experiments for the MAE metric are reported in Fig. 1b (see App. B for results on additional metrics). Consistently with the synthetic benchmarks, we can observe CSL is linearly related to the dominant period $P^{(i)}$, such that $W_{\alpha,m}(x^{(i)}, f) = cP^{(i)}$, however, while on synthetic data we reliably observed a *saturation-to-period* coefficient $c \approx 2$, here different effects can be observed depending on the specific dataset. For BizITObsApp CSL is consistent with the observations on controlled seasonal benchmarks, likely as the time series have clear seasonality and low noise. We found a special case in TimeMoE, where the performance is relatively bad at every context length, and therefore no performance saturation is observed. In the case of Electricity we can also observe a behavior consistent with our previous analysis, yet we have a coefficient $c \approx 7$, which suggests that the estimated dominant period (daily) neglects the importance of the weekly periodicity (which is also highlighted by the ACF plot in App. B). For Solar CSL is less consistent, with a coefficient $1 \leq c \leq 4$ depending on the specific TSFM. While this dataset has a strong daily seasonality, we attribute these inconsistencies to the noise introduced by the occurrence of rainy or cloudy days, which are indeed strong random



(a) Synthetic results

(b) Real-world results

Figure 1. Forecasting MAE on different datasets as a function of the normalized context ratio W/P or W/p. (a) Synthetic datasets. (b) Real-world datasets.

effects from the considered modeling perspective, where precipitation/cloudiness information is not included. Lastly, for LoopSeattle, we can again observe a behavior consistent with our analysis, with a coefficient $c \approx 1$. In this case, this is less than the value $c \approx 2$ found for controlled benchmarks. Similar to Electricity, we attribute this result to the presence of multiple relevant seasonalities. Here the effect of the dominant period (weekly) is mitigated by the presence of another higher-frequency daily seasonality (see the ACF plot in App. B).

4. Discussion

Our results provide relevant insights to understand how the input length relates to the performance of TSFM. We find that CSL emerges as a quantity that depends mainly on the properties of the target time series, rather than on the capabilities of the chosen pre-trained model. This suggests that model-agnostic input length selection strategies can be based on the characterization of the target signal, rather than relying on direct model evaluation. For instance, we empirically demonstrated how, for seasonal time series with period P, an input length of $W \approx 2P$ is a good estimate for the point of diminishing returns. Our results on real-world data, however, suggest that more elaborate methodologies should be developed, in order to achieve higher reliability in more complex scenarios. For instance, considering how multiple seasonalities affect each other, or how periodic and autoregressive dependencies interact.

Our work also suggests a path towards more principled benchmarking practices for TSFM. Evaluating such models only at the largest possible context length, can conflate different factors affecting performance, and overestimate model capabilities when confronted with limited amounts of input information. To counteract this issue, TSFM could report error curves at different input lengths, as well as CSL, under different settings. Moreover, according to our analysis, controlled benchmarks can be designed as a tool to assess model capabilities in a more comprehensive manner. From this viewpoint, synthetic benchmarks can become hypothesis-testing instruments to assess the alignment of TSFMs behavior with a-priori known properties of the data-generating process. For instance, by considering time series for which the Bayes-optimal error can be computed or bounded by a function of the available sequence length, one could assess if a TSFM error curve aligns with this expectation.

5. Conclusion

In this work we introduce the concept of *context saturation length* (CSL) and use it to empirically demonstrate how the point of diminishing returns when increasing the input length of TSFM can be expressed as a function of latent data-generating process parameters that govern the scale of temporal dependencies. This underscores how input length selection can be based on properties of the target time series, rather than via direct evaluation of a chosen TSFM. For time series with a dominant seasonality P, we empirically show context saturation occurs when $W \approx 2P$. Similar relation-

ships are observed for AR processes, when considering their memory size, and for real-world data, when estimating the dominant seasonality via ACF.

Limitations and Future Work While our work offers a framework to understand context utility in TSFM, it is currently limited to the demonstration of empirical regularities. Furthermore, practical application in real-world settings relies heavily on the presence of seasonalities and the accuracy of their estimation. Future research must address these limitations via theoretical analysis of controlled settings, and by researching methodologies for data-centric CSL estimation in complex environments where periodic and autoregressive processes are coupled. Future benchmarking protocols should include standardized controlled environments and consider reporting CSL under such settings.

References

Aksu, T., Woo, G., Liu, J., Liu, X., Liu, C., Savarese, S., Xiong, C., and Sahoo, D. GIFT-Eval: A Benchmark For General Time Series Forecasting Model Evaluation. *arXiv*, October 2024. doi: 10.48550/arXiv.2410.10393.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, Y. Chronos: Learning the Language of Time Series. *arXiv*, March 2024. doi: 10.48550/arXiv.2403.07815.

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. *arXiv*, October 2025. doi: 10.48550/arXiv.2510.15821.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. *arXiv*, October 2023. doi: 10.48550/arXiv.2310.10688.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. YaRN: Efficient Context Window Extension of Large Language Models. *International Conference on Learning Representations*, 2024:31932–31951, May 2024.

Shi, X., Wang, S., Nie, Y., Li, D., Ye, Z., Wen, Q., and Jin, M. Time-moe: Billion-scale time series foundation models with mixture of experts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=e1wDDFmlVu.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. Unified Training of Universal Time Series Forecasting Transformers. *arXiv*, February 2024. doi: 10.48550/arXiv.2402.02592.

Zhang, J., Zhou, Z., Du, W., and Wang, Y. Enhancing the maximum effective window for long-term time series forecasting. *Advances in Neural Information Processing Systems*, 2025.

A. Full synthetic benchmarks.

Synthetic data hyperparameters For the different synthetic datasets, we used the following hyperparameters:

- **Noisy sine:** $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.5$,
- **Harmonic sweep:** $\lambda_1 \sim \mathcal{U}(0.2, 1)$ and $P_2 = \frac{P_1}{u}$, where $u \sim \mathcal{U}(1.5, 5)$,
- **Autoregressive process:** $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. The parameter vectors φ of the process are chosen randomly, ensuring processes are non divergent. For each $p \in \{2, .., 15\}$ we generate 10 different random processes.
- **Linear trend:** $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. Two different slopes a are tested, a gentle one consisting of a linear slope of $a = 0.005$ and a steeper one with $a = 0.05$.

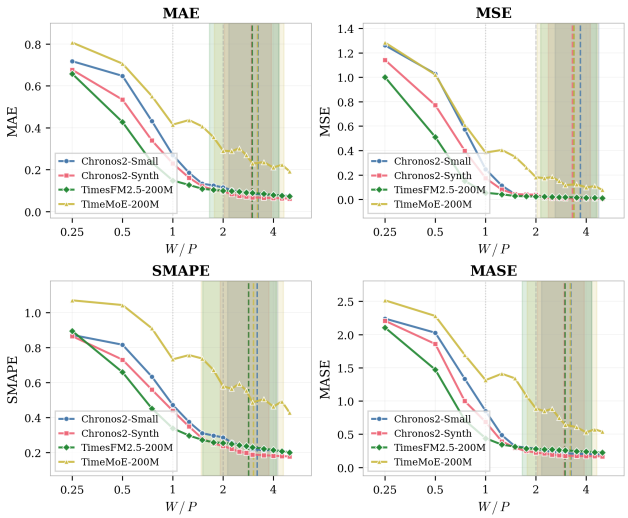
Error normalization for AR processes For the case of the autoregressive dataset, as different parameters converge to different error floors, we normalize by the optimal theoretical error, computed in closed form.

Normalized evaluation grids The set of normalized input length values W/P used for the experiments spans the values

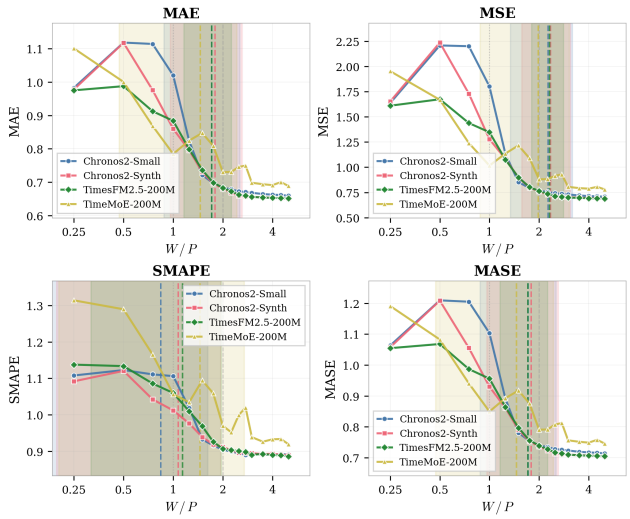
$$\{0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.5, 4.0, 4.5, 5.0\},$$

for the periodic data (i.e., *Noisy Sine*, *Harmonic Sweep* and real-world datasets). Conversely, for non-periodic data (i.e., *AR* and *Linear Trend*), we used the following set of normalized input length values W/p,

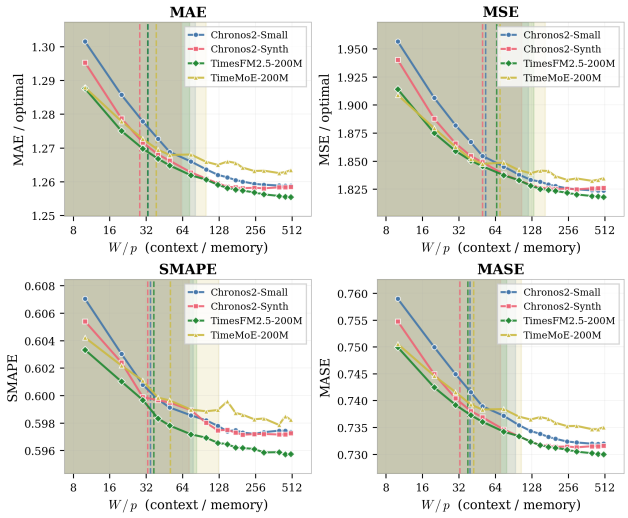
$$\{10, 20, 30, 40, 50, 75, 100, 125, 150, 175, 200, 250, 300, 400, 450, 500\}.$$



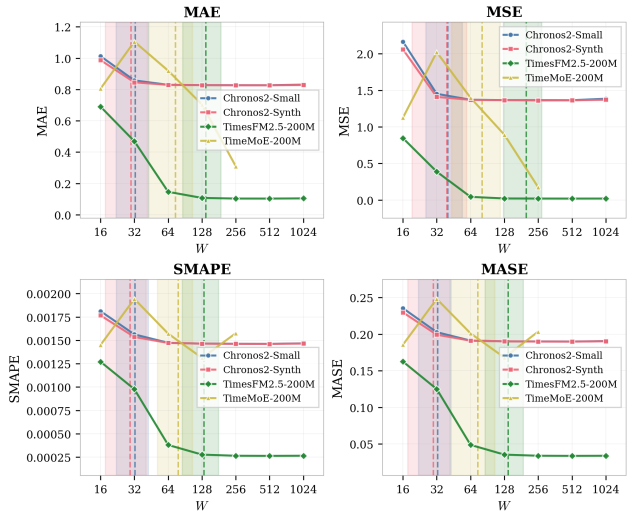
(a) Absolute Error: Harmonic Sweep



(b) Absolute Error: Noisy Sine

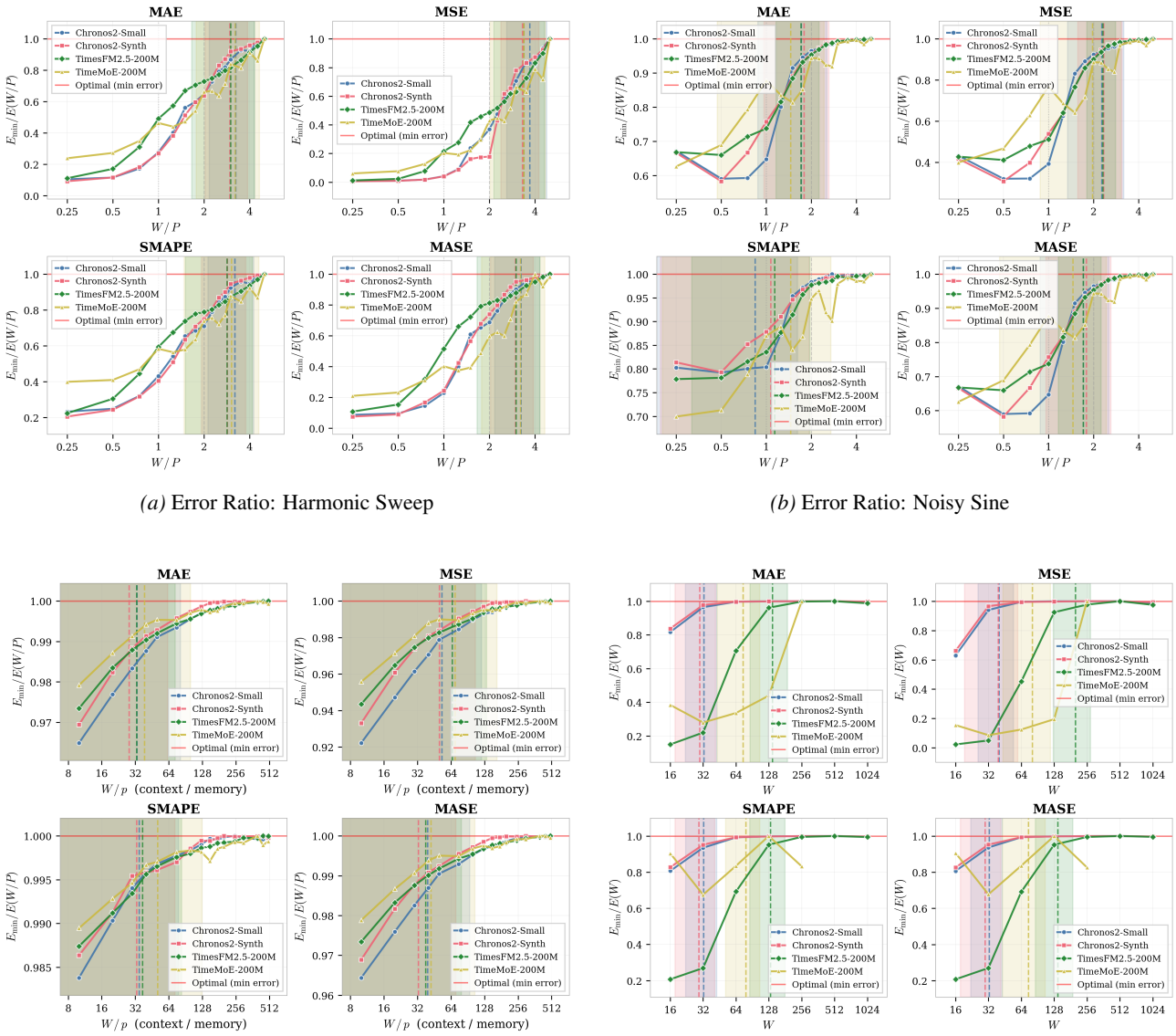


(c) Absolute Error: AR process



(d) Absolute Error: Linear Trend

Figure 2. Forecasting error on synthetic benchmarks. Metrics (MAE, MSE, SMAPE, and MASE) are plotted as a function of the normalized context ratio W/P across six base periods $P \in \{64, 96, 128, 192, 256, 384\}$ for periodic tests, $W/p \in \{8, 16, ..., 512\}$ for the AR case and from $W \in \{16, 32, ..., 1024\}$ in the trend case. Plots indicate absolute forecasting error for the (a) **Harmonic Sweep**, (b) **Noisy Sine** (with additive Gaussian noise $\varepsilon_t \sim \mathcal{N}(0, 0.5^2)$) (c) **AR process** and (d) **Trend** synthetic benchmarks.



(a) Error Ratio: Harmonic Sweep

(b) Error Ratio: Noisy Sine

(c) Error Ratio: AR process

(d) Error Ratio: Linear Trend

Figure 3. Error ratios on synthetic benchmarks. Metrics (MAE, MSE, SMAPE, and MASE) are plotted as a function of the normalized context ratio W/P across six base periods $P \in \{64, 96, 128, 192, 256, 384\}$ for periodic tests, $W/p \in \{8, 16, ..., 512\}$ for the AR case and from $W \in \{16, 32, ..., 1024\}$ in the trend case. Plots indicate absolute forecasting error for the (a) **Harmonic Sweep**, (b) **Noisy Sine** (with additive Gaussian noise $\varepsilon_t \sim \mathcal{N}(0, 0.5^2)$) (c) **AR process** and (d) **Trend** synthetic benchmarks.

B. Real-world benchmarks

Datasets evaluated:

Table 1. Summary of Selected GIFT-Eval Datasets

Dataset	**Domain**	**Sampling Rate (Δt)**	**Number of Time Series**	**Estimated Samples per Period**
Electricity	Energy	15 minutes	370	672
BizITObsApp	IT/Cloud Ops	10 seconds	1	360
Loop Seattle	Transportation	5 minutes	323	2016
Solar	Energy/Nature	10 minutes	137	144

The values for the grid of normalized lengths for assessment of performance in the real-world datasets were

$$L = \{32, 48, 64, 96, 128, 192, 256, 384, 512, 768, 1024, 1536, 2048, 2560, 3072, 4096\}$$

resulting on a different set of W/L for each dataset given the estimated strongest period.

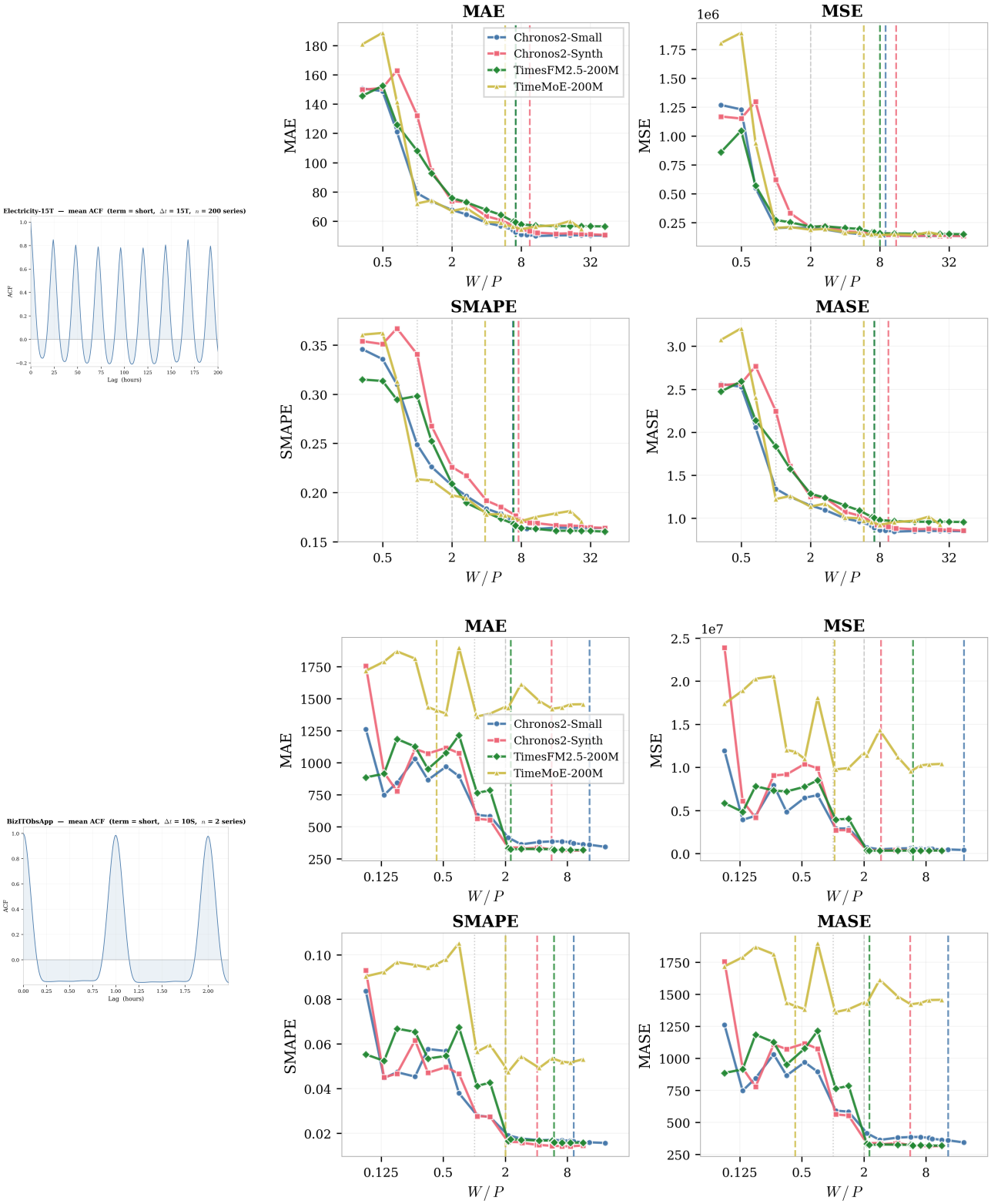


Figure 4. Real-world case studies on *Electricity* (top) and *BizITObsApp* (bottom). For each dataset, the left panel shows the empirical autocorrelation function (ACF), highlighting regularly spaced peaks associated with the dominant recurrence, while the right panel reports forecasting error as a function of the normalized context ratio W/P. Metrics shown are MAE, MSE, SMAPE, and MASE.

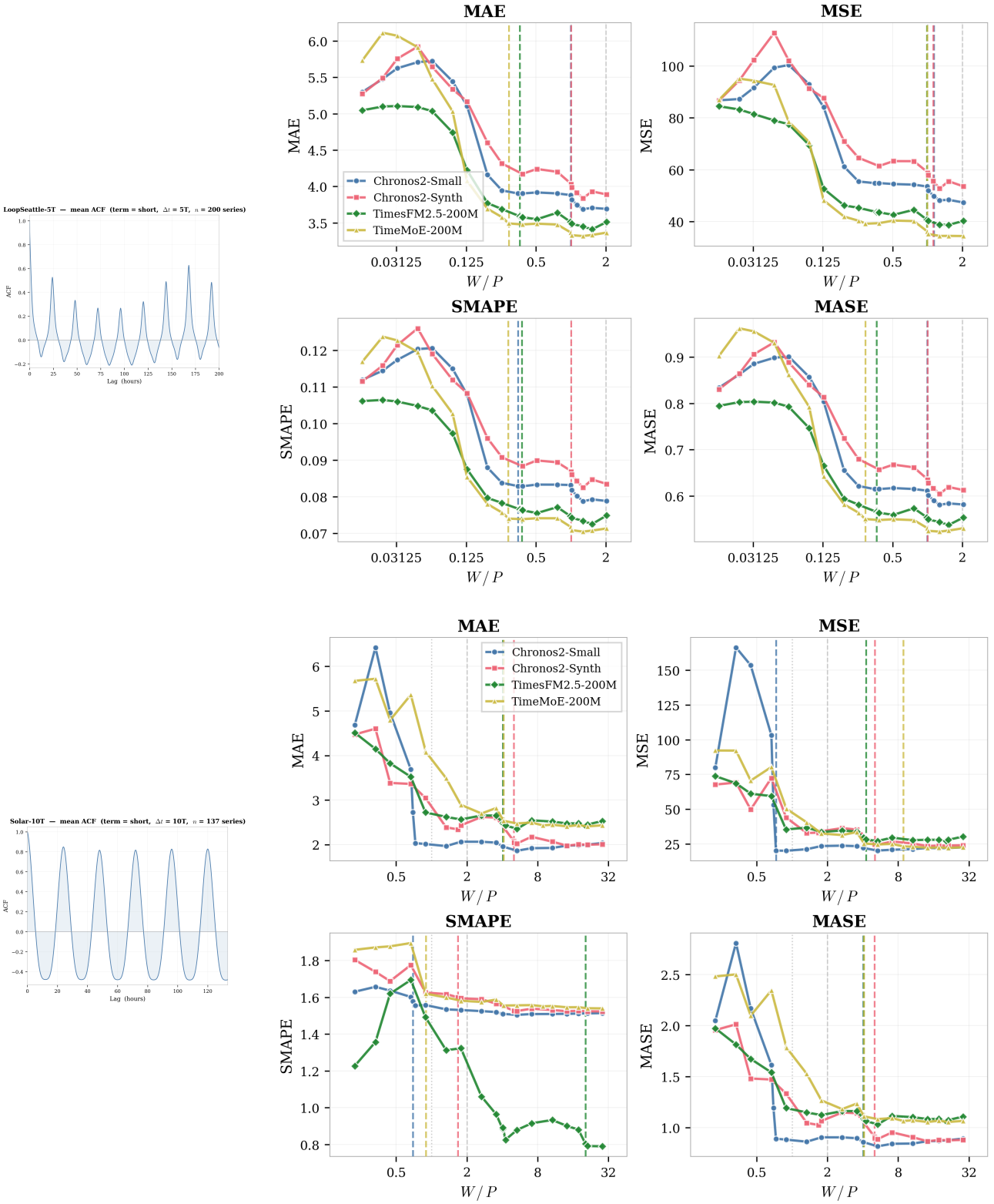
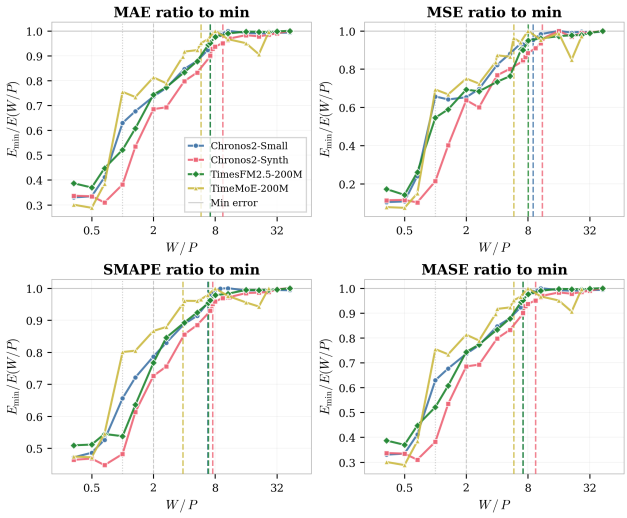
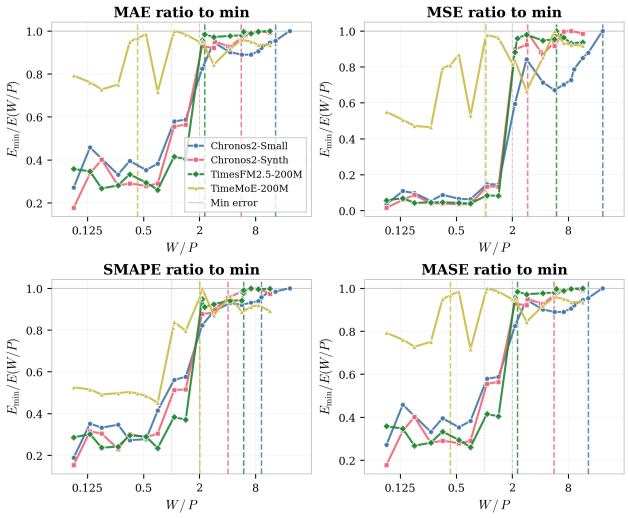


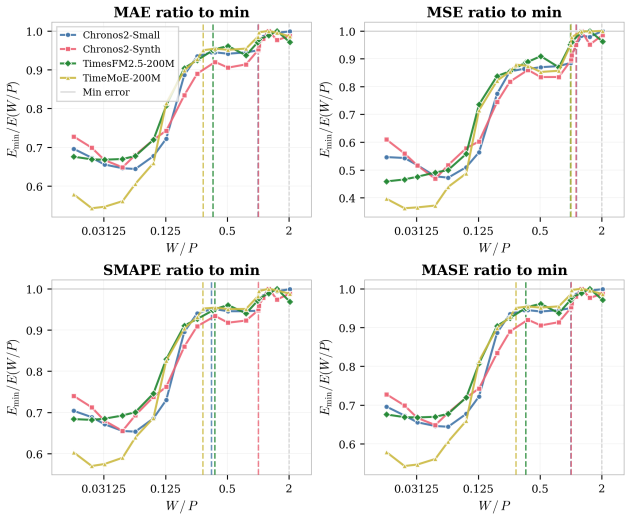
Figure 5. Real-world case studies on *Loop Seattle* (top) and *Solar* (bottom). For each dataset, the left panel shows the empirical autocorrelation function (ACF), highlighting regularly spaced peaks associated with the dominant recurrence, while the right panel reports forecasting error as a function of the normalized context ratio W/P. Metrics shown are MAE, MSE, SMAPE, and MASE.



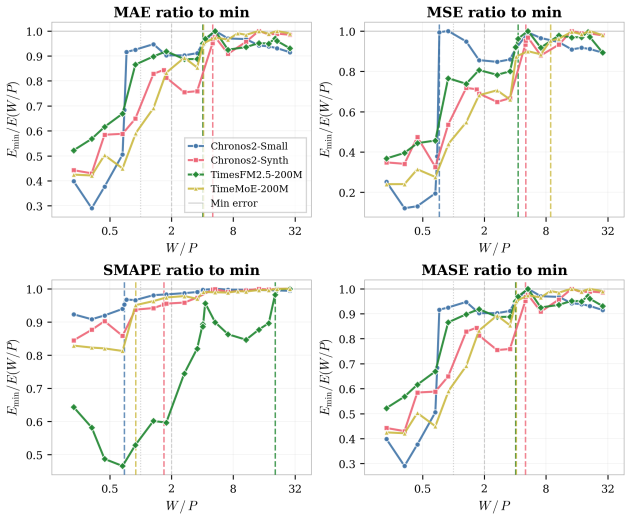
(a) Error ratio: Electricity



(b) Error ratio: BizITObsApps



(c) Error ratio: Loop Seattle



(d) Error ratio: Solar

Figure 6. Error ratios on real-world datasets. Metrics (MAE, MSE, SMAPE, and MASE) are plotted as a function of the normalized context ratio W/P. Plots indicate the error ratio relative to the minimum for the (a) **Electricity**, (b) **BizITObsApps**, (c) **Loop Seattle** and (d) **Solar** datasets.