WHAT CAN TIME-SERIES FOUNDATION MODELS ACTUALLY DO? TOWARDS CAPABILITY-LEVEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) achieve strong zero-shot forecasting performance, but why do they work, and what abilities support their predictions? Existing evaluations compare forecasting accuracy across datasets or under selected forecasting conditions, leaving unclear which abilities models possess and when those abilities fail. We investigate these questions through six forecasting abilities: Structure Induction, History Retrieval, State Tracking, Uncertainty Modeling, Channel Modeling, and Exogenous Modeling. For each ability, we construct controlled experiments that vary the information available to a model or the difficulty of using that information while holding other conditions fixed. We evaluate 21 frozen TSFMs through their supported interfaces, examining whether they can infer temporal structure, retrieve historical information, track changing states, represent uncertainty, combine information across channels, and respond to external conditions. The results reveal ability differences that aggregate forecasting rankings do not capture. Separating periodic structure from residual variation improves forecasting, whereas supplying additional informative channels does not consistently reduce error. Stochastic generation also does not ensure that predictive distributions express the uncertainty present in the target. These findings connect forecasting performance to how models represent, combine, and use predictive information. They motivate evaluating TSFMs through explicit ability profiles and examining whether model designs translate available information into accurate predictions and calibrated uncertainty. All code is available at https://anonymous.4open.science/r/sortinghat-32B3/.

1 INTRODUCTION

Time-series foundation models (TSFMs) achieve strong zero-shot forecasting performance across diverse datasets without task-specific parameter updates (Das et al., 2024; Ansari et al., 2024; Woo et al., 2024). Despite this progress, it remains unclear what forecasting capabilities these models have acquired. Can they infer temporal relationships, retrieve relevant historical observations, or adapt when the underlying dynamics change? Strong forecasting performance motivates these questions, but does not by itself answer them.

Existing benchmarks provide essential comparisons across datasets and forecasting conditions. GIFT-Eval evaluates models under a common protocol (Aksu et al., 2024), while TIME further examines performance across tasks and temporal characteristics (Qiao et al., 2026). However, these comparisons do not directly establish which capabilities support a model's predictions. For example, low error on a recurring sequence does not reveal whether a model can retrieve distant observations or infer the relationship governing their recurrence. Likewise, aggregate accuracy does not establish whether a model can use additional variables or represent conditional uncertainty. Models with similar forecasting errors may therefore possess different capabilities, and an aggregate ranking cannot fully describe these differences.

We investigate TSFMs through six forecasting capabilities: Structure Induction, History Retrieval, State Tracking, Uncertainty Modeling, Channel Modeling, and Exogenous Conditioning. For each capability, we construct controlled experiments that vary the information available to the model

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

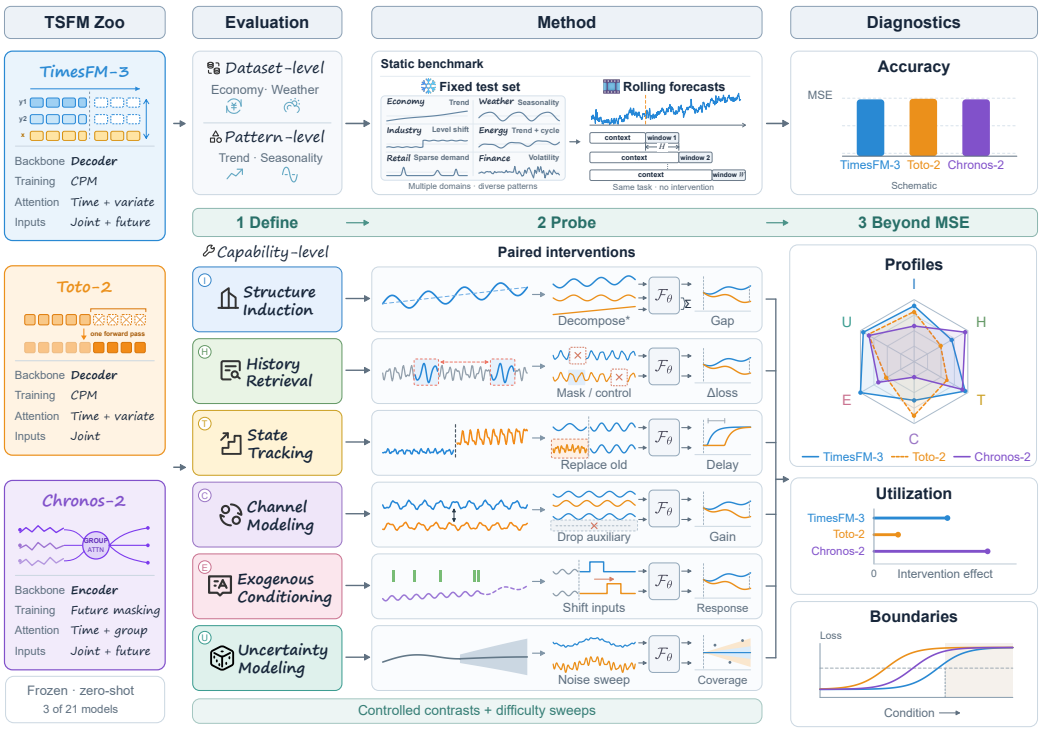


Figure 1: Overview of our capability-level evaluation. Models with different architectures and training objectives are evaluated through controlled probes of six forecasting capabilities. The resulting diagnostics characterize capability profiles, information use, and failure conditions beyond aggregate forecasting error. The radar chart uses ranks from our capability evaluation; the remaining plots are schematic.

or the difficulty of using that information. We evaluate 21 frozen TSFM versions through their supported interfaces. Comparisons of history length, dependency distance, temporal structure, and auxiliary information reveal when each capability produces a forecasting benefit. Known conditional distributions additionally allow uncertainty estimates to be compared with explicit references. Figure 1 provides an overview of the capability-level evaluation framework.

Our analysis yields three insights. First, the use of temporal structure remains uneven: additional historical repetitions benefit some models, output-block boundaries can interrupt otherwise smooth continuations, and explicit periodic decomposition can improve average forecasting performance. These findings motivate closer examination of how models learn temporal relationships and maintain them across prediction steps. Second, access to additional informative channels does not consistently reduce error, highlighting the need to improve how models combine information across channels. Third, stochastic generation does not ensure calibrated uncertainty: Sundial is less well calibrated on our random-walk diagnostic than two models with weaker aggregate forecasting ranks. This finding motivates evaluating predictive distributions against the uncertainty required by the task. For the community, these insights identify concrete directions for model development and support reporting explicit capability profiles alongside aggregate accuracy.

Our contributions are threefold:

- We organize forecasting evaluation around six functional capabilities and their controlled probes.
- We evaluate 21 frozen TSFM versions, reporting capability profiles and task-specific measurements while retaining unsupported interfaces as missing.
- We identify differences in structure continuation, historical evidence use, uncertainty calibration, and cross-channel information use that aggregate forecasting comparisons do not fully capture.

2 Related Work

Time-series foundation models. TimesFM, Chronos, and Moirai learn transferable forecasters from heterogeneous time series using different representations and predictive objectives (Das et al., 2024; Ansari et al., 2024; Woo et al., 2024). Their zero-shot forecasting performance motivates a further question: which forecasting capabilities have these models acquired, and how reliably do those capabilities transfer across tasks?

Evaluation of TSFMs. GIFT-Eval compares forecasting performance across datasets, domains, frequencies, and horizons (Aksu et al., 2024). TIME advances to pattern-level evaluation by relating performance to structural time-series features (Qiao et al., 2026). Our work takes a further step in abstraction, advancing evaluation from *dataset level* through *pattern level* to *capability level*. The unit of analysis becomes the forecasting capability a model exhibits. A capability can be tested through different patterns, while one pattern can demand different capabilities. Patterns thus serve as test settings for capabilities, and stress conditions specify the demands under which those capabilities must remain reliable.

Capability and compositional evaluation. Behavioral testing complements aggregate accuracy by evaluating specific capabilities under controlled input changes (Ribeiro et al., 2020). Potosnak et al. (2025) distinguish compositional reasoning from in-distribution generalization and investigate architectural choices through controlled studies. SimpleTimeBench diagnoses fundamental forecasting behavior using trends, periodic signals, and leading covariates (Ghoroghchian et al.). Simple target–covariate evaluations provide a related diagnostic perspective on whether models use supplied covariates (Berthelier et al., 2026). Our work organizes forecasting capabilities into a unified six-dimensional framework, connecting model profiles, paired information interventions, and stress sweeps.

Historical reuse and design biases. Context parroting shows that copying from observed histories can be highly competitive on dynamical-system forecasting tasks (Zhang & Gilpin, 2026). This motivates distinguishing historical reuse from generalization of the underlying temporal structure. Controlled studies also show that patch size, embeddings, and training objectives induce different forecasting biases (Yu et al., 2026). Our probes examine related behavioral differences across frozen models, while treating associations with model design as hypotheses rather than identified architectural causes.

Probabilistic evaluation. Adler et al. (2026) find that TSFMs are often better calibrated than competitive baselines across their evaluated settings. Our uncertainty probes address a complementary question: whether predictive marginals follow known conditional uncertainty as the terminal state and forecast horizon change. Favorable aggregate calibration does not establish reliable behavior in every controlled regime. Together, these comparisons motivate capability-level evaluation alongside aggregate forecasting benchmarks.

3 Capabilities, Probes, and Experimental Design

We decompose forecasting into six functional capabilities and evaluate each through controlled forecasting probes. A capability describes an operation that a model must perform. Each probe tests this operation through a forecasting task with specified inputs, predictive relationships, and controlled variations. The same forecasting task can be used to examine different capabilities through different experimental controls. For example, a Seasonal AR task can test whether a model infers a recurring temporal dependency, while interventions on its observed history can test whether the model uses relevant historical information. Appendix Table 28 summarizes the six capabilities, their diagnostic questions, and representative probes.

3.1 Structure Induction

Capability. This dimension evaluates whether a model can infer temporal rules from observations, such as the period of a recurring pattern or the dependence on previous values. These rules are then used to predict how the series continues.

The difficulty of this operation depends on the underlying temporal structure, how prediction errors propagate over time, and whether multiple components must be identified and combined. We examine this capability through controlled forecasting probes of non-chaotic and chaotic dynamics. The non-chaotic probes include elementary functions and component mixtures. For the dynamical systems, we vary initial conditions and parameters to test whether predictions adapt to the relationships expressed in the observed history. For elementary functions and mixtures, we examine rule continuation, consistency across output blocks, and the effect of separating temporal components.

Detailed probes, controls, and settings are provided in Appendix C and Appendix K.

3.2 History Retrieval

Capability. This dimension requires a model to find and use past observations that are relevant to the forecast. The needed information may lie in an earlier cycle or a distant segment of the history.

The difficulty of this operation depends on how far relevant observations are from the forecast origin and how much effective evidence the history provides for identifying their predictive relationship. We therefore organize our controlled forecasting probes around two factors: information distance and information density.

Information Distance. Information distance describes how far back a model must look to find observations relevant to the forecast. We use `Lag-copy` to control this distance through the lag parameter L: larger values require retrieving information from more distant parts of the history.

Information Density. Information density describes how much effective evidence the observed history provides for identifying a predictive relationship. It depends on how often a pattern repeats and how clearly these repetitions reveal the relationship. Detailed probes, controls, and settings are provided in Appendix D and Appendix K.

3.3 State Tracking

Capability. This dimension evaluates whether a model can recognize changes in the dynamics of a series from new observations and update its forecasts accordingly.

State tracking requires adapting when the active dynamics change and using relevant historical information when a previous state returns. We examine these behaviors through controlled forecasting probes of state transitions, transition speed, and state recurrence.

Detailed probes, controls, and settings are provided in Appendix E and Appendix K.

3.4 Uncertainty Modeling

Capability. This dimension evaluates whether a model can predict the distribution of future outcomes given the observed history, and represent how uncertainty changes with the forecast horizon and current state.

We construct controlled forecasting probes using stochastic processes with known conditional uncertainty structures. The probes cover horizon-dependent variance, conditional volatility, discrete event intensity, congestion, and uncertainty about latent states. A Gaussian random-walk probe additionally tests uncertainty that accumulates with the forecast horizon without mean reversion. Let $\mathcal{H}_T = (x_1, \ldots, x_T)$ denote the observed history. Our evaluation concerns the marginal predictive distribution at each future step, $p(x_{T+h} \mid \mathcal{H}_T)$.

Detailed probes, controls, and settings are provided in Appendix F and Appendix K.

3.5 Channel Modeling

Capability. This dimension evaluates whether a model can use predictive information from jointly observed channels to improve forecasts of a target series.

We construct controlled forecasting probes with specified dependencies among channels. These dependencies allow us to examine two factors: the number of observed channels and the structure of their predictive relationships. The model receives the observed histories without the dependency graph, parent identities, or latent-factor assignments.

Detailed probes, controls, and settings are provided in Appendix G and Appendix K.

3.6 Exogenous Conditioning

The distinction between historically observed variables and known future inputs is standard in multi-horizon forecasting (Lim et al., 2021). Simple target–covariate tests further motivate examining whether a model actually uses the supplied inputs (Berthelier et al., 2026).

Capability. This dimension evaluates whether a model can use future external inputs provided at prediction time to adjust its forecasts in accordance with the actual response of the target series.

We construct controlled forecasting probes in which the observed history remains unchanged while different future inputs produce different future trajectories. Each pair starts from the same state at the forecast origin. We examine periodic and non-periodic inputs using pharmacokinetic (PK) and spatial dynamics as representative probes. Additional probes cover epidemic, traffic-flow, and car-following dynamics.

Detailed probes, controls, and settings are provided in Appendix H and Appendix K.

3.7 Experimental Setup

We evaluate 21 TSFM versions with frozen pretrained weights through their supported native interfaces. Model versions, checkpoint sources, and interface coverage are documented in Appendix B. Unsupported settings remain unavailable rather than receiving a numerical score. Depending on the probe, we compare TSFMs with persistence, history-fitted baselines, and explicitly identified privileged references; the corresponding settings are provided in Appendices C–H.

We report point-forecast errors, probabilistic losses, and probe-specific diagnostic measures. Capability profiles summarize relative performance within each eligible model pool on a 0–100 rank scale. Metric definitions, reference scales, profile construction, and aggregation and resampling procedures are specified in Appendix A. Real-data splits and matched evaluation windows are described in Appendix I.

4 Results

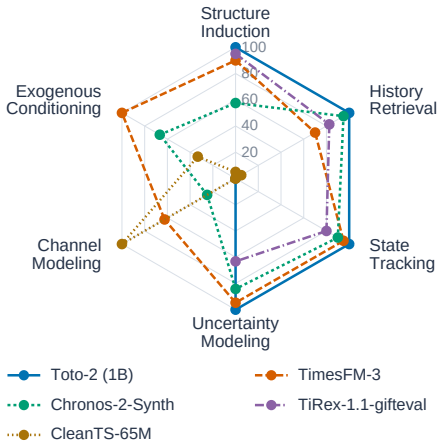
We first examine the six-capability profiles and their relationship to aggregate forecasting performance. We then examine temporal structure, repeated historical information, predictive uncertainty, and cross-channel information. Detailed State Tracking and Exogenous Conditioning results are reported in Appendices E and H.

4.1 What Capabilities Underlie Forecasting Performance?

Figure 2 shows five selected profiles from the original 21-model evaluation. The complete profiles and scores are reported in Appendix Figure 6 and Table 6. Models with similar aggregate forecasting performance can exhibit different capability profiles.

TiRex and Chronos-2-Synth have similar published GIFT-Eval MASE, 0.716 and 0.720, yet their Structure Induction scores are 95.0 and 57.5, while their History Retrieval scores are 82.5 and 95.0. Similar aggregate point estimates therefore do not imply matching diagnostic profiles; these values are not a statistical equivalence test.

We further compare capability scores with the GIFT-Eval point-forecasting leaderboard (Aksu et al., 2024), using the snapshot collected on September 18, 2026. These results indicate that GIFT-Eval rankings do not fully capture capability-level differences. In particular, the correlations for History Retrieval ($\rho = 0.545$) and State Tracking ($\rho = 0.599$) show only partial correspondence with aggregate forecasting performance, while Channel Modeling exhibits a negative association ($\rho = -0.564$).



Patching and attention. Patching is an established representation strategy for time-series Transformers (Nie et al., 2023). Controlled studies further show that patch size and training objectives can induce distinct forecasting biases (Yu et al., 2026). Several models with high scores in our capability evaluation combine patch representations with Transformer backbones. Patching groups local observations into representations, while attention allows the model to relate information across these groups. This combination provides a shared basis for representing local structure and dependencies across the context. However, the differences in Figure 2 show that this common architecture does not produce a uniform capability profile. The way these representations are used to construct predictions also matters.

4.2 How Do TSFMs Infer and Continue Temporal Structure?

Composite structures and periodic decomposition. Our composition experiments further show that adding two temporal components can make forecasting more difficult: directly forecasting $x = f_1 + f_2$ can produce larger errors than summing forecasts obtained separately from the true component histories. The separate-component branch receives the true decomposition and serves as a privileged diagnostic reference. We therefore examine whether explicitly separating a periodic component can also improve TSFM forecasts.

Given the period length, we estimate a periodic template using historical phase averages, subtract it from the input, forecast the residual with the frozen TSFM, and add the template back to the forecast. Across the tested mixture conditions, this procedure reduces average MSE for all five evaluated models, with reductions of 4.63–12.37% at $C = 128$ and 2.06–7.06% at $C = 512$ (Appendix Figure 8). Following the residual-cycle perspective of CycleNet (Lin et al., 2024), explicitly representing the recurring component allows the forecaster to focus on the remaining dynamics. The observed gains suggest that this separation can also reduce the burden of jointly modeling periodic and residual components in frozen TSFMs.

All five aggregate means improve, but some intervals include zero. More specifically, AR mixtures improve in both contexts, while all five white-noise mixtures worsen at $C = 128$. At $C = 512$, this harm is smaller and Chronos has a small positive mean. Repeating finite-sample noise in the phase template is consistent with this residual-dependent pattern, but the current contrasts do not uniquely identify the mechanism.

Discontinuities at patch boundaries. In a subset of our Structure Induction experiments, we evaluate forecasts of arithmetic, geometric, exponential, and logarithmic sequences. We find that several models exhibit abrupt changes in forecast residuals at multiples of their output-patch lengths, even though the underlying sequences are smooth (Figure 3). TiRex and CleanTS exhibit jumps at 32-step boundaries, while the Kairos variants show pronounced changes at their 64-step output boundaries. A plausible explanation is that training models to predict multi-step patches without an explicit continuity constraint across adjacent patches allows position-dependent errors to appear as jumps where patches meet. Pointwise decoding does not introduce this particular multi-step stitching boundary.

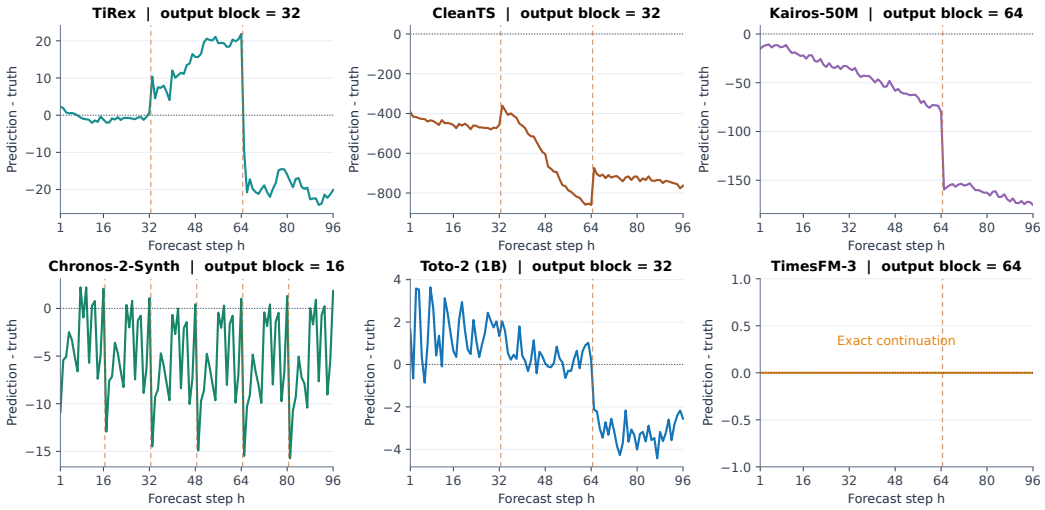


Figure 3: Forecast residuals for the arithmetic sequence $x(t) = 2t$, with $C = 1028$ and $H = 96$. Dashed vertical lines mark boundaries between native output blocks, whose lengths appear in the panel titles. Panels use independent vertical scales. TimesFM-3 exactly continues this particular saved sequence.

4.3 How Does Repeated Historical Information Affect Forecasting?

In our History Retrieval experiments, we examine whether longer histories containing more repetitions of the same dependency improve forecasting. We compare context lengths $C = 512$ and $C = 2048$ at fixed dependency distances $d \in \{9, 25, 33, 129\}$, increasing the number of visible cycles fourfold. All comparisons use noise-free lag-copy sequences and mean MSE over the first eight forecast steps, averaged across 20 seeds.

As shown in Appendix Figure 10, 9 of the 19 models with complete paired coverage (47.4%) achieve lower mean MSE at $C = 2048$ in at least three of the four dependency distances. These improvements are consistent with the possibility that repeated observations help models identify the predictive relationship and constrain plausible continuations. However, changing the context length in this experiment also changes the forecast origin and template phase, so the improvements cannot be attributed to repetition alone. The results support a model-dependent benefit from longer, more repetitive histories, without establishing a unified power-law relationship between repetition count and error.

4.4 Does Stochastic Generation Improve Uncertainty Modeling?

Existing evaluations report favorable calibration for TSFMs across their tested settings (Adler et al., 2026). Here, we examine a complementary controlled setting in which the conditional distribution is known and uncertainty must increase with the forecast horizon.

Stochastic generation and distributional accuracy. We compare Sundial (Liu et al., 2025) with PatchFM and Goia, which have weaker aggregate forecasting ranks. Within the matched 21-model set, they rank 16th, 17th, and 19th, respectively, by GIFT-Eval mean task MASE rank in the September 18, 2026 snapshot.

In the random-walk experiments with $C = 512$ and $H = 96$, Sundial's nominal 80% prediction intervals achieve 40.1% coverage, with a width ratio of 0.486 relative to the oracle interval. Under the same conditions, PatchFM and Goia achieve coverage rates of 58.5% and 75.0%, with width ratios of 0.683 and 0.951, respectively (Figure 4). Coverage is computed under the known conditional distribution and averaged over forecast steps and cases. The mean case-wise excess-loss ratios on the common nine-quantile grid are 27.25%, 8.98%, and 2.81%, respectively. Thus, weaker

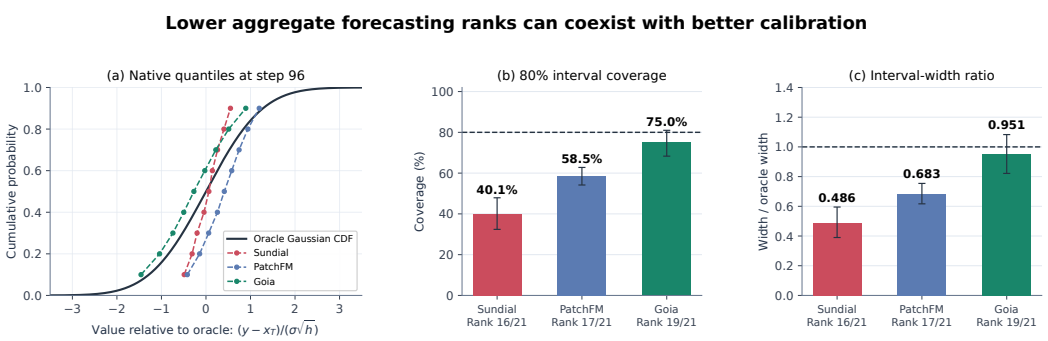


Figure 4: Uncertainty calibration for Sundial, PatchFM, and Goia at $C = 512$ and $H = 96$. (a) Native deciles at step 96 for a fixed history (seed 390000, $\sigma = 0.4$), compared with the oracle Gaussian CDF; segments connect available deciles. (b) Coverage of nominal 80% prediction intervals. (c) Interval widths relative to the oracle. Summary metrics average 96 forecast steps, ten history seeds, and three noise levels. Error bars show descriptive 95% paired-history bootstrap intervals. Model ranks refer to GIFT-Eval mean task MASE rank within the matched 21-model set.

aggregate forecasting ranks can coexist with better uncertainty calibration, and stochastic generation alone does not ensure calibrated forecasts.

From sampling randomness to predictive uncertainty. The dispersion of generated forecasts is determined by the learned conditional transformation of the input noise; introducing randomness does not ensure that this transformation reproduces the spread of the target distribution. Sundial's narrow intervals and low coverage indicate that its generated forecasts underrepresent conditional uncertainty in these tasks, whereas direct quantile prediction can represent the required spread by estimating the corresponding distributional quantiles.

4.5 How Do TSFMs Exploit Cross-Channel Information?

Do additional informative channels improve forecasting? We progressively reveal additional channels while keeping the target history, forecast origin, and future target fixed. The experiments consider noisy observations of a shared latent state, weak delayed relationships, and complementary information distributed across multiple parent channels. We evaluate each setting with context lengths $C = 128$ and $C = 512$.

As shown in Figure 5, increasing the number of informative channels does not consistently reduce prediction error. Some models obtain gains, whereas others plateau or deteriorate as additional channels are revealed. The pattern depends on the model, relationship type, and context length. For example, increasing the total channel count from 384 to 512 in the shared-factor setting raises TimesFM-3's mean NMSE by 11.73% at $C = 128$ and 6.87% at $C = 512$. In the distributed-parent setting at $C = 512$, the same increase raises Toto-1's mean NMSE by 22.04%, even though the reference conditional variance decreases by 41.37%.

The latter comparison separates the predictive information available in the input from the information used by the model. Additional channels reduce the remaining uncertainty about the target, but the model must combine their contributions to obtain this benefit. The observed error increases show that providing more predictive information does not ensure that a TSFM will incorporate it into a more accurate forecast.

When does cross-channel interaction occur? To understand the difference in channel modeling performance between Toto-2 and TimesFM-3, we examine how the two models incorporate information from other channels. The evaluated Toto-2 1B configuration applies variate attention only in the last of its 36 Transformer layers (Khwaja et al., 2026).[1] Within a forward pass, each channel is

[1]Evaluated configuration: https://huggingface.co/Datadog/Toto-2.0-1B/blob/main/config.json.

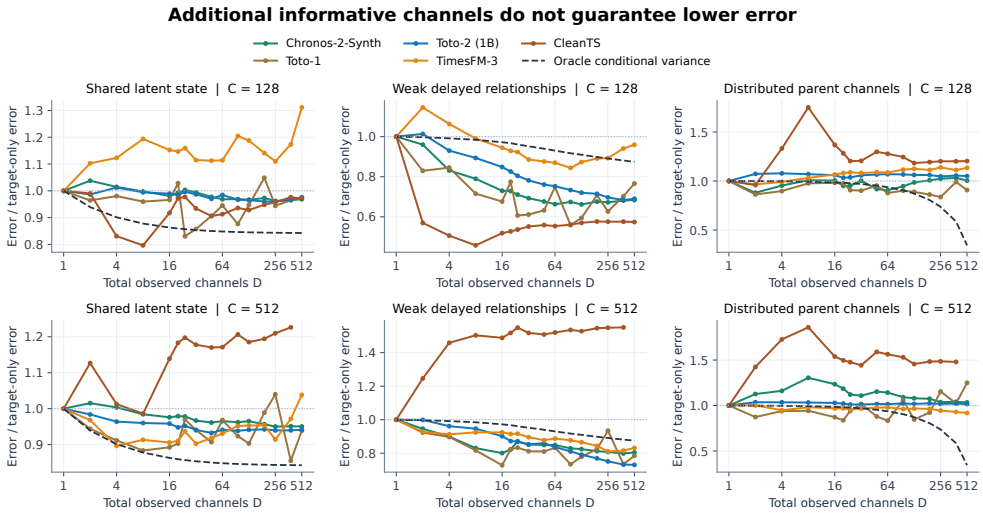


Figure 5: Forecasting with additional informative channels. Rows correspond to $C = 128$ and $C = 512$. Columns show shared-state observations, weak delayed relationships, and distributed parent channels. The horizontal axis gives the total number of observed channels, including the target. Model curves show mean first-step NMSE over 16 seeds, normalized by each model's own target-only value. Oracle conditional variances use their own target-only references. These ratios do not directly compare absolute errors across models. CleanTS outputs at $C = 512$, $D = 512$ are unavailable.

therefore processed temporally before cross-channel fusion, with no subsequent temporal attention layer to further process the fused information. TimesFM-3 instead alternates temporal and variate attention across multiple layers (Das et al., 2024),[2] allowing information from other channels to inform subsequent temporal processing.

The observed performance difference is consistent with the benefit of integrating cross-channel information throughout temporal processing. Strong forecasting performance on individual series does not by itself establish that a model can incorporate dependencies between series; the placement of channel interaction determines how those dependencies enter its temporal representations.

Together, these observations associate capability differences with specific stages of forecasting: the representation of history, the construction of output blocks, the integration of channel information, and the formation of predictive distributions. The six-capability evaluation makes these distinctions visible beyond aggregate forecasting performance.

5 Conclusion

In this paper, we showed that strong forecasting performance does not fully characterize the capabilities acquired by time-series foundation models. We evaluated 21 TSFMs using controlled probes across six capabilities and found that aggregate rankings capture shared strengths while obscuring differences in information use and failure conditions. Our experiments show that accurate component forecasts do not guarantee equally accurate composition, success in stable regimes does not ensure timely adaptation, and additional informative inputs do not consistently improve predictions. These findings highlight the need to assess the operations underlying forecasting success alongside overall accuracy. We encourage future work to use capability diagnosis to guide model development and selection, with greater attention to how models combine temporal structure, update relevant evidence, and maintain reliability as forecasting demands change.

[2]TimesFM-3 architecture description: https://research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/.

AI USE STATEMENT

In this work, we used generative AI tools to help refine the conceptual framework of capability-level evaluation, to provide feedback on portions of the research methodology and experimental design, and to assist in analyzing and interpreting some experimental results. In addition, we used generative AI tools to retrieve and summarize literature, to polish the manuscript and improve its readability, to design and revise scientific figures, and to edit portions of the software code.

We did not use generative AI tools to propose or prove mathematical claims, to write mathematical proofs, to conduct qualitative or thematic analysis, or to autonomously generate the experimental results reported in this paper. We manually reviewed all AI-assisted research ideas, experimental design suggestions, code, text, and figures; we obtained the relevant experimental results by actually running the programs and verified them against the raw experimental outputs and the relevant literature. We take full responsibility for the final content, claims, and research artifacts presented in this paper.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 describes the capability definitions, diagnostic probes, and experimental design. Appendix A specifies the prediction protocol, evaluation metrics, reference scales, aggregation procedures, and statistical reporting. Appendix B documents the evaluated model versions, checkpoint sources, supported interfaces, and detailed results. Appendices C–H provide capability-specific experimental details, including data-generating processes, parameter settings, and paired controls. Appendix I describes the real-data splits, forecast-window construction, and information interventions. Code for data generation, model evaluation, and result analysis will be made available at https://anonymous.4open.science/r/sortinghat-32B3/.

REFERENCES

Coen Adler, Yuxin Chang, Samar Abdi, Felix Draxler, and Padhraic Smyth. Beyond accuracy: Are time series foundation models well-calibrated? In *International Conference on Learning Representations*, volume 2026, pp. 95913–95944, 2026.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR. Expert Certification.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *Advances in Neural Information Processing Systems*, 38:57529–57580, 2026.

Gaspard Berthelier, Mariia Baranova, Andrei-Tiberiu Pantea, Etienne Le Naour, Adrien Petralia, Tahar Nabil, and Themis Palpanas. Investigating simple target-covariate relationships for chronos-2 and tabPFN-TS. In *1st ICLR Workshop on Time Series in the Age of Large Models*, 2026. URL https://openreview.net/forum?id=H8p5Yc0c3W.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, et al. This time is different: An observability perspective on time series foundation models. *Advances in neural information processing systems*, 38:50907–50951, 2026.

Carlos F Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation research part B: methodological*, 28(4): 269–287, 1994.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Kun Feng, Shaocheng Lan, Yuchen Fang, Wenchao He, Sihan Lu, Shuqi Gu, Lintao Ma, Xingyu Lu, and Kan Ren. Kairos: Toward adaptive and parameter-efficient time series foundation models. *arXiv preprint arXiv:2509.25826*, 2025.

Xinghong Fu, Yanhong Li, Georgios Papaioannou, and Yoon Kim. Reverso: Efficient time series foundation models for zero-shot forecasting. *arXiv preprint arXiv:2602.17634*, 2026.

Nafiseh Ghoroghchian, Haipeng Zhang, Shuyi Han, Alex Labach, and George Stein. Foundations without fundamentals: Zero-shot blind spots in time series fms. In *2nd ICML Workshop on Foundation Models for Structured Data*.

Milo Gibaldi and Donald Perrier. *Pharmacokinetics*. crc Press, 1982.

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

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

Lars Graf, Thomas Ortner, Stanisław Woźniak, and Angeliki Pantazi. Flowstate: Sampling-rate-equivariant time-series forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=9yNGBueKgZ.

James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pp. 357–384, 1989.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, et al. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764, 2021.

Shengsheng Lin, Weiwei Lin, Xinyi HU, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=clBiQUgj4w.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liu25be.html.

Edward N Lorenz. Deterministic nonperiodic flow. In *Universality in Chaos, 2nd edition*, pp. 367–378. Routledge, 2017.

Helmut Lütkepohl et al. *New introduction to multiple time series analysis*, volume 1. springer Berlin, 2005.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Willa Potosnak, Cristian Challu, Mononito Goswami, Kin G Olivares, Michał Wiliński, Nina Żukowska, and Artur Dubrawski. Investigating compositional reasoning in time series foundation models. *arXiv preprint arXiv:2502.06037*, 2025.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's TIME: Towards the next generation of time series forecasting benchmarks. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=79TgfXHbsK.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

Michael Scheuerer and Thomas M Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.

John F Shortle, James M Thompson, Donald Gross, and Carl M Harris. *Fundamentals of queueing theory*. John Wiley & Sons, 2018.

Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000.

Ming Chen Wang and George Eugene Uhlenbeck. On the theory of the brownian motion ii. *Reviews of modern physics*, 17(2-3):323, 1945.

Xue Wang, Tian Zhou, Jinyang Gao, Bolin Ding, and Jingren Zhou. Output scaling: Yinglong-delayed chain of thought in a large pretrained time series forecasting model. *arXiv preprint arXiv:2506.11029*, 2025.

Yunshi Wen, Wesley M Gifford, Chandra Reddy, Lam M Nguyen, Jayant Kalagnanam, and Anak Agung Julius. Revisiting the generic transformer: Deconstructing a strong baseline for time series foundation models. *arXiv preprint arXiv:2602.06909*, 2026.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Annan Yu, Danielle Maddix, Boran Han, Xiyuan Zhang, Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos, Andrew Gordon Wilson, Michael W Mahoney, and Bernie Wang. Understanding the implicit biases of design choices for time series foundation models. In *International Conference on Learning Representations*, volume 2026, pp. 71998–72049, 2026.

Yuanzhao Zhang and William Gilpin. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning. In *International Conference on Learning Representations*, volume 2026, pp. 58314–58340, 2026.

A Notation and Evaluation Protocol

This appendix specifies the notation, prediction protocol, metrics, and aggregation used in our experiments. The capability order throughout the appendix is Structure Induction, History Retrieval, State Tracking, Uncertainty Modeling, Channel Modeling, and Exogenous Conditioning. We distinguish the original capability profiles from subsequent diagnostic experiments: additional probes provide further evidence but do not silently change the conditions or scores underlying the original profiles.

A.1 Common Notation

Table 1 fixes the shared notation. The same quantity retains the same symbol across probes; distinct quantities use different letters. Subscripts identify instances of the same quantity.

Table 1: Shared notation. Lengths and distances count observations unless a physical sampling interval is explicitly specified.

Symbol	Meaning
$t,\ T$	Observation index and forecast origin; T is the last observed index.
C	Number of observations supplied as history.
$H,\ h$	Requested forecast length and future-step index, $1 \le h \le H$.
$V,\ j$	Number of observed channels and channel index.
$x_{t,j},\ \hat{x}_{T+h,j}$	Observed value and point forecast of the same series.
$\mathcal{O}_T$	Information supplied to the model at the forecast origin.
$P,\ d$	Oscillation period and actual source–target dependency distance.
ϕ	Autoregressive coefficient in a scalar seasonal recurrence.
$\sigma,\ \epsilon_t$	Innovation standard deviation and innovation random variable.
ν	Declared reference scale used to normalize prediction error.
η	Mean absolute historical increment used as the MASE denominator.
e_h	Forecast residual, $\hat{x}_{T+h} - x_{T+h}$.
$q,\ \hat{x}^{(q)}_{T+h}$	Quantile probability level and predicted value at that level.
ℓ_q	Pinball loss at probability level q.
N	Number of independent seeds or histories in a reported comparison.
$M,\ r$	Number of eligible models and a model's rank within that set.

For a scalar target, the history is $x_{T-C+1:T}$ and the forecast target is $x_{T+1:T+H}$. We omit the channel index in scalar metric definitions. A multichannel input contains V aligned histories; jointly providing these histories is different from batching V independent univariate calls. Only the target channels specified by a probe are scored.

Lag-copy is written as $x_t = x_{t-d} + \epsilon_t$; its legacy generator argument `lag` equals $d - 1$. Thus, settings 8, 24, 32, and 128 mean $d = 9, 25, 33, 129$. By contrast, $x_t = \phi x_{t-P} + \epsilon_t$ has distance P. The noiseless lag-copy quantity C/d counts visible cycle lengths, not independent observations.

A.2 Prediction and Information Protocol

Frozen models and native interfaces. TSFM weights remain frozen. Inputs comprise observed histories, supported masks and time information, the forecast length, and auxiliary inputs allowed by the probe. Future conditions require a supported interface. Generator parameters, latent states, and future targets remain with the evaluator. Known-mechanism and true-component controls are explicitly labeled privileged.

Checkpoint-specific preprocessing and native output conventions are retained and documented with the model settings. In particular, a point forecast may be a native mean, median, or another declared statistic; it is not silently replaced by a different statistic for an analysis. We do not impose a single external normalization on every experiment. When a probe applies a common historical normalization to paired branches, the same historical statistics are used and predictions are returned

to the target units before scoring. Model-internal preprocessing is separate from the reference scales used by the evaluator.

Paired comparisons. History and channel interventions preserve the target and its future within each pair. Fixed-origin context comparisons expose suffixes of the same underlying history. Component experiments satisfy the same additive identity in both the observed history and the future target. Exogenous condition branches share the history and initial simulator state but have branch-specific futures generated under their respective future inputs. Replacing a supplied condition with an incorrect one is evaluated against the intended branch's target.

The fixed-origin property is experiment-specific. In the arithmetic, geometric, exponential, and logarithmic continuation experiments, changing C also changes the forecast origin and value range. Their context contrast therefore does not isolate the effect of additional history. The added sinusoidal protocol instead uses a shared future across contexts. Requested horizon and scored horizon are also reported separately: scoring the first eight steps of a 96-step forecast does not imply that the model was called with an eight-step horizon.

Coverage and baselines. Unavailable interfaces or native length restrictions are recorded as unsupported, rather than as high prediction errors. Execution failures, missing saved outputs, and undefined metric scales are recorded separately. A planned condition is not counted as an observed result. Classical baselines may fit the supplied history; this does not fine-tune the TSFM. Persistence, local extrapolation, historical lag search, fitted linear models, and empirical-distribution references are used where specified. Privileged references are distinguished from baselines that use only information available to the forecasting model.

A.3 Metrics and Reference Scales

Point prediction. For one target channel, define $e_h = \hat{x}_{T+h} - x_{T+h}$. Full-horizon errors are

$$\mathrm{MSE} = \frac{1}{H}\sum_{h=1}^{H} e_h^2, \qquad \mathrm{MAE} = \frac{1}{H}\sum_{h=1}^{H} |e_h|. \tag{1}$$

A prefix or segment analysis averages over its stated forecast steps. Channel-specific errors precede channel aggregation. Normalized errors are

$$\mathrm{NMSE} = \frac{\mathrm{MSE}}{\nu^2}, \qquad \mathrm{NRMSE} = \frac{\sqrt{\mathrm{MSE}}}{\nu}. \tag{2}$$

The definition of ν is fixed by each probe, not inferred from its name alone. It may be a declared reference-history standard deviation or a known process scale. Direct-composition and component-sum comparisons use the common denominator $\nu^2 = \nu_1^2 + \nu_2^2$, where ν_1 and ν_2 are the two component reference standard deviations. Separately normalized function-family summaries are not subtracted to estimate a composition effect. Where available, both raw and normalized errors are reported so that a small normalized error is not mistaken for a small absolute error.

For nonseasonal MASE (Hyndman & Koehler, 2006), using the visible history as the scaling reference,

$$\eta = \frac{1}{C-1}\sum_{t=T-C+2}^{T} |x_t - x_{t-1}|, \qquad \mathrm{MASE} = \frac{\mathrm{MAE}}{\eta}. \tag{3}$$

Some fixed-origin diagnostics instead compute η from a common reference history shared across contexts. They explicitly distinguish this fixed denominator from the visible-history version. The latter is used for the simple sequence-continuation experiments and the original main-panel MASE table. The scoring reference does not add observations to the model input. Degenerate or numerically invalid denominators are reported as undefined under the corresponding scorer's validity check; they are not replaced by an arbitrary small denominator.

Marginal probabilistic prediction. For a predicted quantile $\hat{x}^{(q)}_{T+h}$, define the standard quantile loss (Koenker & Bassett Jr, 1978):

$$\ell_q\left(x_{T+h}, \hat{x}^{(q)}_{T+h}\right) = \left(q - \mathbf{1}\{x_{T+h} < \hat{x}^{(q)}_{T+h}\}\right)\left(x_{T+h} - \hat{x}^{(q)}_{T+h}\right). \tag{4}$$

The original probability profile uses its declared scale-normalized pinball losses. The common nine-quantile diagnostic uses

$$\mathrm{CRPS}_{\mathrm{Q9}} = \frac{2}{9H} \sum_{h=1}^{H} \sum_{q \in \{0.1, 0.2, \ldots, 0.9\}} \ell_q \left(x_{T+h}, \hat{x}_{T+h}^{(q)} \right). \tag{5}$$

This finite-grid score approximates the quantile integral defining CRPS (Gneiting & Raftery, 2007); it is not exact continuous-distribution CRPS. Native quantile labels are retained, including documented floating-point rounding, without inventing missing levels or extrapolating tails. For sample-output models, the common grid is obtained from their native samples. Empirical-sample CRPS, when available, is reported separately. Missing probability outputs are not reconstructed from a point forecast. Explicit point-mass references have exact CRPS equal to MAE and are labeled accordingly; they do not establish native uncertainty modeling.

For two available levels $q_{\mathrm{lo}} < q_{\mathrm{hi}}$, the interval's nominal probability is $q_{\mathrm{hi}} - q_{\mathrm{lo}}$. Coverage and mean width are

$$\mathrm{Coverage} = \frac{1}{H} \sum_{h=1}^{H} \mathbf{1}\{\hat{x}_{T+h}^{(q_{\mathrm{lo}})} \le x_{T+h} \le \hat{x}_{T+h}^{(q_{\mathrm{hi}})}\}, \tag{6}$$

$$\mathrm{Width} = \frac{1}{H} \sum_{h=1}^{H} \left(\hat{x}_{T+h}^{(q_{\mathrm{hi}})} - \hat{x}_{T+h}^{(q_{\mathrm{lo}})} \right). \tag{7}$$

We report the actual nominal probability implied by the native endpoints. Coverage is interpreted together with width and probabilistic loss, following the distinction between calibration and sharpness (Gneiting et al., 2007). Quantile crossings are retained and audited rather than repaired by sorting the output values.

Conditional-distribution scores average over reference futures for the same history: the original stochastic suite uses 100 such futures per history, not 100 independent histories. Random-walk losses and coverage use analytic Gaussian expectations; realized-future errors are separate. The known-mechanism Gaussian reference is privileged. Shared comparisons concern future marginal distributions, not complete joint paths.

A.4 Aggregation and Statistical Reporting

The independent resampling unit is a generating seed or independently sampled history, according to the suite. Parameter settings, intervention branches, forecast origins, and forecast steps that share this unit are not treated as additional independent replicates. For deterministic probes, a seed may vary only selected components such as observation noise; it does not necessarily produce a new underlying dynamical system. Table 2 summarizes the principal repetition budgets without pooling experiments that use different protocols.

Table 2: Repetition units in the analyzed experiment suites. Expanded parameter grids and repeated model calls do not increase these budgets.

Suite	Repetition budget
Original core induction, seasonal-history, and Switching AR	5 seeds
Lag-copy	20 seeds
Original smooth AB / ABA	10 / 5 seeds
Original uncertainty, channel, and conditioning suites	3 histories/seeds
Accepted independent synthetic supplements	16 seeds/histories
Informative-channel count experiment	16 seeds
Random walk; simplified periodic decomposition	10 histories/seeds
Simple sequence and added sine parameter grids	Fixed seed 1723

Within a reported comparison, we first aggregate the applicable settings and origins within each independent unit, then average across units with the stated weights. The original main-panel MASE/CRPS table averages origins within a scenario and seed, scenarios within a seed, and then

seeds. A displayed cell requires the specified coverage; missing conditions are not silently removed to produce a complete-looking mean. Losses from different systems are not pooled in raw units into a global forecasting score.

Seed-based diagnostic intervals are paired: the same resampled units are used for both sides of a contrast. Most accepted supplementary analyses use 4,000 bootstrap resamples; the simplified periodic-decomposition analysis uses 2,000. Other intervals retain their explicitly stated estimator. These intervals are descriptive unless a particular comparison specifies an adjusted testing procedure. Parameter-grid counts from a fixed-seed function experiment do not support a seed-based confidence interval or a significance claim.

Real-data comparisons aggregate by the declared time blocks, keeping repeated targets and interventions from a block together. Nonoverlapping forecast windows can still have overlapping histories or temporal dependence. Tracking analyses report the fraction recovered alongside delay among recovered trajectories; trajectories that do not recover within the observation window are right-censored. A short delay among a small subset of recovered trajectories does not establish faster adaptation overall.

A.5 Construction of the Capability Profiles

The six-axis profiles summarize selected panels of the original experiments. Table 3 gives their order, coverage, and primary measurements. Coverage counts denote eligible models in these panels, not successful forecasting behavior on every condition or eligibility for all later diagnostic experiments.

Table 3: Panels underlying the original capability profiles. H is the requested forecast length; the measurement column states any shorter scoring window. Later probes do not redefine these profiles.

Capability	Models	C/H	Primary measurement
Structure Induction	21	512/16	NRMSE across Seasonal AR, Lorenz, and SEIR settings.
History Retrieval	21	512/96	First-step MSE relative to the zero-prediction reference on noiseless lag-copy, $d = 9, 25, 33, 129$.
State Tracking	21	256/96	Clean NMSE over the first 16 steps in the new-state-dominant AB panels.
Uncertainty Modeling	20	512/96	Scale-normalized marginal pinball loss, aggregated by stochastic family.
Channel Modeling	5	512/96	First-step cross-copy error and first-eight-step gains in three 20-channel systems.
Exogenous Conditioning	4	512/96	Response errors in the six non-SEIR controlled systems.

Within each selected subprobe, eligible models are ranked with average ranks for ties and with the direction chosen so that rank one is best. For $M > 1$ eligible models, rank r is mapped to

$$\mathrm{Score}(r; M) = 100\frac{M - r}{M - 1}. \tag{8}$$

Subprobe scores are averaged with equal weights within a capability. Models are then reranked by this average and mapped once more using Equation 8. The channel profile gives equal weight to the copying term and the three system-specific gain terms. The probability profile uses one family-aggregated term per stochastic family. Unsupported capabilities remain unscored.

These scores describe relative model positions, not absolute capability or error reductions. MASE/CRPS tables and additional diagnostics retain their own conditions and coverage. External comparisons report the matched model set and its size for each capability.

B Evaluated Models and Overall Results

We report the evaluated model versions, complete capability profiles, and selected absolute forecasting results. The profiles summarize the original controlled panels defined in Appendix A; subsequent diagnostic experiments are distinguished from this fixed comparison. Results follow the order Structure Induction, History Retrieval, State Tracking, Uncertainty Modeling, Channel Modeling, and Exogenous Conditioning. In numerical result tables, bold and underlined entries mark the best and second-best distinct values within each column, including ties. Emphasis is determined before rounding. These marks describe point estimates and do not denote statistical significance.

B.1 Evaluated Models and Interfaces

Model coverage. Table 4 lists the 21 evaluated versions. The set includes Transformer encoders and decoders, encoder–decoder models, an xLSTM model, state-space models with functional-basis decoding, and hybrid architectures. Backbone, pretraining task, and predictive output are recorded separately. CPM denotes contiguous patch masking; NTP denotes next-token or next-patch prediction. FlowState's functional-basis decoder is distinct from Sundial's stochastic flow-based generation. CleanTS's evaluated implementation is causal, but its checkpoint-specific pretraining objective remains unverified.

Table 4: Evaluated versions and interfaces. Joint and Future indicate eligibility for the joint-channel and future-input probes, respectively; a dash denotes no eligible interface in these experiments. Interface support does not imply successful use of the information.

Model	Backbone	Training task	Output	Joint	Future
Chronos-2-Synth	Encoder	Future masking	Quantiles	Yes	Yes
PatchTST-FM-r1	Encoder	CPM	Quantiles	–	–
Granite-PatchTST-r1	Encoder	CPM	Quantiles	–	–
YingLong-110M	Encoder	Masked recovery	Quantiles	–	–
YingLong-300M	Encoder	Masked recovery	Quantiles	–	–
TimesFM-3	Decoder	CPM	Quantiles	Yes	Yes
Toto-2 (1B)	Decoder	CPM	Quantiles	Yes	–
TimesFM-2.5	Decoder	NTP	Quantiles	–	–
PatchFM	Decoder	NTP	Quantiles	–	–
Toto-1	Decoder	NTP	Samples	Yes	Yes
Sundial-128M	Decoder	Flow generation	Samples	–	–
TiRex-1.1-gifteval	xLSTM	CPM	Quantiles	–	–
Kairos-10M	Encoder–decoder	Direct forecast	Quantiles	–	–
Kairos-23M	Encoder–decoder	Direct forecast	Quantiles	–	–
Kairos-50M	Encoder–decoder	Direct forecast	Quantiles	–	–
FlowState-9.1M	SSM + basis decoder	CPM	Quantiles	–	–
FlowState-r1.1	SSM + basis decoder	CPM	Quantiles	–	–
Granite-FlowState-r1.1	SSM + basis decoder	CPM	Quantiles	–	–
Reverso-Small	Conv. + DeltaNet	Direct forecast	Point	–	–
Goia-nano-v0	Conv. + attention	Forecast + distill.	Quantiles	–	–
CleanTS-65M	Causal Transformer	Unverified	Quantiles	Yes	Yes

Model sources. The evaluated checkpoints include Chronos-2-Synth (Ansari et al., 2025); PatchTST-FM-r1 and Granite-PatchTST-r1 (Wen et al., 2026); YingLong-110M and YingLong-300M (Wang et al., 2025); TimesFM-2.5 and TimesFM-3 (Das et al., 2024); Toto-1 (Cohen et al., 2026) and Toto-2 (1B) (Khwaja et al., 2026); Sundial-128M (Liu et al., 2025); TiRex-1.1-gifteval (Auer et al., 2026); Kairos-10M, Kairos-23M, and Kairos-50M (Feng et al., 2025); FlowState-9.1M, FlowState-r1.1, and Granite-FlowState-r1.1 (Graf et al., 2026); and Reverso-Small (Fu et al., 2026). We additionally evaluate PatchFM,[3] Goia-nano-v0,[4] and CleanTS-65M.[5] The repository and revision records below identify the exact evaluated releases.

[3] https://huggingface.co/vilhess/PatchFM
[4] https://huggingface.co/gredio/goia-forecast-nano-v0
[5] https://huggingface.co/EINK/CleanTS-65M

Checkpoint identity. Table 5 records the repositories and revision prefixes where a snapshot revision was retained. Local releases are identified by their evaluated version names. The declared checkpoint paths agree between the original experiment groups; this agreement is not a new verification of weight bytes. Short display names throughout this appendix refer to these specific versions: Toto-2 denotes the 1B release, and TiRex denotes TiRex-1.1-gifteval.

Table 5: Checkpoint identifiers for the original comparison. Revision prefixes identify recorded snapshots; local release labels are not weight hashes. Repository links identify the associated public model.

Model	Repository or local release	Revision
Chronos-2-Synth	autogluon/chronos-2-synth	`3607918a9f`
PatchTST-FM-r1	ibm-research/patchtst-fm-r1	`67dc5f9a1d`
Granite-PatchTST-r1	ibm-granite/granite-timeseries-patchtst-fm-r1	`151f9c6d57`
YingLong-110M	qcw2333/YingLong_110m	`e3738d65f4`
YingLong-300M	qcw2333/YingLong_300m	`c57b37f6d4`
TimesFM-3	google/timesfm-3.0-pytorch	`43046b85ec`
Toto-2 (1B)	Datadog/Toto-2.0-1B	`1604e1a524`
TimesFM-2.5	google/timesfm-2.5-200m-pytorch	`1d952420fb`
PatchFM	vilhess/PatchFM	`4a8c35a5a9`
Toto-1	Datadog/Toto-Open-Base-1.0	`0411ceb27b`
Sundial-128M	thuml/sundial-base-128m	`3212e42564`
TiRex-1.1-gifteval	NX-AI/TiRex-1.1-gifteval	`01b1ebc83a`
Kairos-10M	mldi-lab/Kairos_10m	`786ecafbb3`
Kairos-23M	mldi-lab/Kairos_23m	`5123edcc9d`
Kairos-50M	mldi-lab/Kairos_50m	`b788434dd9`
FlowState-9.1M	ibm-research/flowstate	`ec815e4160`
FlowState-r1.1	ibm-research/flowstate	`940580f969`
Granite-FlowState-r1.1	ibm-granite/granite-timeseries-flowstate-r1	`b80e0f1b1c`
Reverso-Small	Reverso/reverso_small	`70f539a4e9`
Goia-nano-v0	gredio/goia-forecast-nano-v0	`a3d17d8261`
CleanTS-65M	EINK/CleanTS-65M	`1df59376c9`

Prediction interfaces. Weights remain frozen, and each adapter preserves its declared preprocessing, length restrictions, and output convention. Most quantile models use the $q = 0.5$ prediction as their point forecast; the two PatchTST variants use the adapter's mean estimate from quantile integration. Toto-1 and Sundial use the sample median, with 100 native samples in the original probabilistic protocol. Reverso uses its deterministic regression output. A native API field named "mean" does not override the underlying statistic: TiRex returns a median through this field in the evaluated interface.

All 21 models enter the first three capability profiles. Twenty provide eligible marginal probabilistic outputs, five enter Channel Modeling, and four enter Exogenous Conditioning. Reverso is excluded from the probabilistic profile. Toto-2 enters the joint-channel panel but has no eligible future-input result in the original comparison. Univariate batching is not treated as joint-channel modeling. Unsupported conditions remain missing; they are not assigned zero capability scores.

Reference predictors. Depending on the probe, references include last-value, local-line, seasonal copying, Fourier, AR, Ridge, ARX, and empirical-history-quantile predictors. History-fitted references may estimate their coefficients from the supplied history; this is distinct from updating a TSFM's pretrained weights. Controls using a true component or a known mechanism are identified as privileged. References establish the difficulty of a task and are not included in the 21-model rank pools.

B.2 Complete Capability Profiles

Table 6 gives all six capability scores alongside published GIFT-Eval aggregate MASE. Within each capability, a higher score means a better relative position under the fixed aggregation in Appendix A.5. A score of 100 identifies the highest-ranked eligible model; it is not a success rate. The different eligible sets and missing interfaces preclude a common six-capability total.

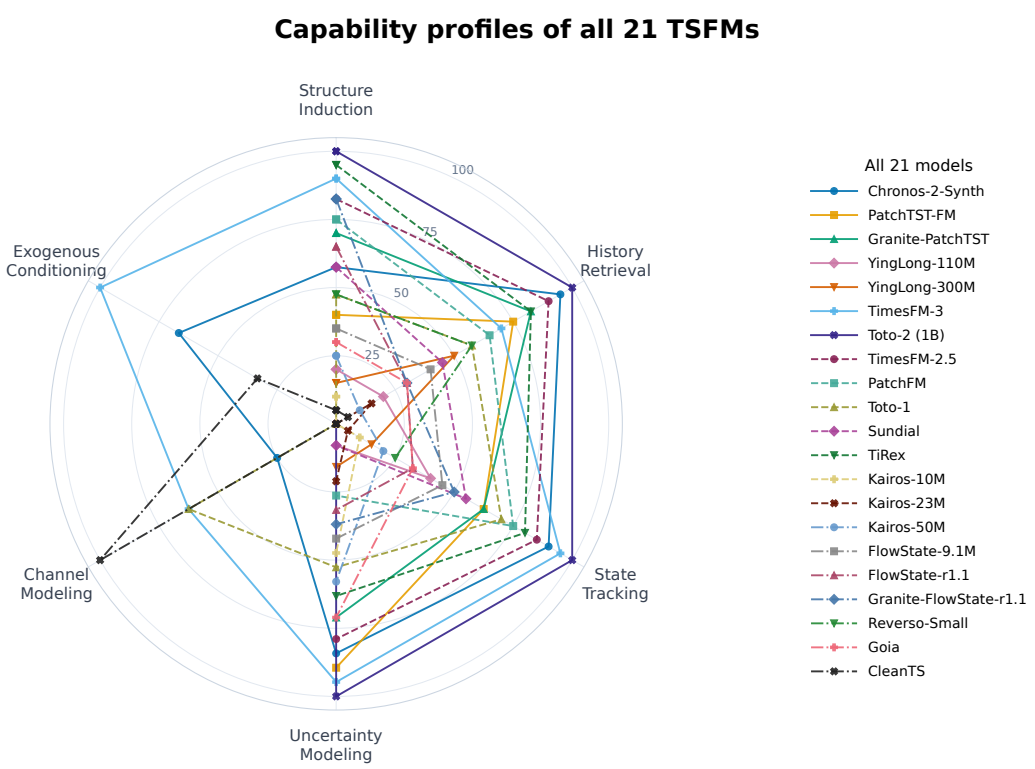


Figure 6: Six-capability profiles of all 21 evaluated TSFMs. All models share one radar and a common 0–100 scale. Higher scores indicate better within-capability relative performance under the original frozen evaluation. Unsupported interfaces remain missing. The axes have different eligible model pools and do not define an overall capability score.

Shared strengths and different profiles. Toto-2 ranks first in Structure Induction, History Retrieval, State Tracking, and Uncertainty Modeling. TimesFM-3 ranks first in Exogenous Conditioning and second in State Tracking and Uncertainty Modeling. The broad capabilities are positively associated: Spearman correlations are 0.749 between Structure Induction and History Retrieval, 0.855 between Structure Induction and State Tracking, and 0.869 between History Retrieval and State Tracking. Thus, the profiles capture shared forecasting strength together with differences in relative performance, rather than six statistically independent abilities.

The differences remain informative. Chronos-2-Synth ranks second in History Retrieval and third in State Tracking, compared with a tied rank of 9.5 in Structure Induction. TiRex and Chronos-2-Synth have similar published GIFT-Eval MASE, 0.716 and 0.720, yet their Structure Induction scores are 95.0 and 57.5, while their History Retrieval scores are 82.5 and 95.0. Similar aggregate point estimates therefore do not imply matching diagnostic profiles; these values are not a statistical equivalence test.

B.3 Absolute Results on the Controlled Panels

Rank scores discard the size of error differences. To retain that information, Tables 7–9 report selected absolute losses and reference-relative measurements. Every column retains its own metric, scoring window, and reference scale. These columns are not averaged into a cross-task error score and do not redefine the profiles.

Structure Induction, History Retrieval, and State Tracking. Table 7 reports MASE for all three original Structure Induction systems and diagnostic errors for long-distance copying and frequency

Table 6: Complete original capability profiles (higher is better) and GIFT-Eval aggregate MASE (lower is better; September 18, 2026 snapshot). Ind., Ret., State, Unc., Chan., and Exog. follow the six-capability order. Eligible model counts are 21, 21, 21, 20, 5, and 4. Dashes are ineligible interfaces, not zero scores. Row order follows the model registry and is not an overall ranking.

Model	Ind.	Ret.	State	Unc.	Chan.	Exog.	GIFT MASE
Chronos-2-Synth	57.5	95.0	90.0	84.2	25.0	66.7	0.720
PatchTST-FM-r1	40.0	75.0	62.5	89.5	–	–	0.707
Granite-PatchTST-r1	70.0	82.5	62.5	71.1	–	–	0.717
YingLong-110M	20.0	20.0	40.0	7.9	–	–	0.809
YingLong-300M	15.0	50.0	15.0	15.8	–	–	0.798
TimesFM-3	90.0	70.0	95.0	94.7	62.5	**100.0**	**0.667**
Toto-2 (1B)	**100.0**	**100.0**	**100.0**	**100.0**	0.0	–	0.699
TimesFM-2.5	82.5	90.0	85.0	78.9	–	–	0.705
PatchFM	75.0	65.0	75.0	26.3	–	–	0.778
Toto-1	47.5	57.5	70.0	52.6	62.5	0.0	0.750
Sundial-128M	57.5	45.0	55.0	7.9	–	–	0.750
TiRex-1.1-gifteval	95.0	82.5	80.0	63.2	–	–	0.716
Kairos-10M	10.0	0.0	10.0	47.4	–	–	0.753
Kairos-23M	0.0	15.0	5.0	21.1	–	–	0.748
Kairos-50M	25.0	10.0	20.0	57.9	–	–	0.742
FlowState-9.1M	35.0	40.0	45.0	42.1	–	–	0.726
FlowState-r1.1	65.0	30.0	32.5	31.6	–	–	0.701
Granite-FlowState-r1.1	82.5	30.0	50.0	36.8	–	–	0.701
Reverso-Small	47.5	57.5	25.0	–	–	–	0.726
Goia-nano-v0	30.0	30.0	32.5	71.1	–	–	0.815
CleanTS-65M	5.0	5.0	0.0	0.0	**100.0**	33.3	0.798

transitions. The MASE values were recovered from saved forecasts using the evaluation protocol in Appendix A. Original profiles instead use their declared NRMSE or NMSE measurements. For noiseless lag-copy, the saved MASE scale is undefined in the selected panel; the table retains the original first-step MSE relative to the zero-prediction reference. Undefined scale is different from model failure.

Uncertainty Modeling. Table 8 reports the common nine-quantile approximation $\mathrm{CRPS}_{\mathrm{Q9}}$ for the five stochastic families. It is computed from saved outputs on the common native deciles, with 100 reference futures per observed history. These losses are in the original target units; comparisons are within columns. The original uncertainty profile instead aggregates ranks of scale-normalized marginal pinball loss. The raw loss table therefore need not reproduce its family-aggregated order. Marginal loss does not by itself measure interval calibration or dependence across a sampled future path.

Channel Modeling and Exogenous Conditioning. Table 9 separates copying error, the benefit of joint prediction relative to independent prediction, and response error under changed future inputs. CleanTS has the highest original Channel Modeling score, while Toto-2 has the lowest, despite the reverse ordering on the four broad profiles. CleanTS's leading score chiefly reflects short-range gains relative to its independent-channel reference. Replacing those gain terms with joint-prediction absolute errors moves CleanTS to fourth and TimesFM-3 to first. Interface support, incremental benefit, and final prediction accuracy are therefore distinct observations.

For Exogenous Conditioning, TimesFM-3 has a mean response-error ratio of 0.275, compared with 0.753 for Chronos-2-Synth and approximately 0.996–0.998 for CleanTS and Toto-1. The latter pair receives different rank scores despite nearly equal response errors. These values summarize the six non-SEIR systems; they do not establish uniformly accurate timing or successful controlled-SEIR prediction. The history-fitted ARX2 reference obtains 0.070 on this response measure and remains outside the TSFM ranking.

Table 7: Selected point-forecast results (lower is better). Seasonal AR, SEIR, and Lorenz–63 use MASE, $C = 512$, $H = 16$, and five seeds. Copying uses $C = 512$, $H = 96$, $h = 1$, $d = 129$, no noise, and 20 seeds; its entry is MSE divided by the zero-prediction MSE. Frequency tracking reports clean NMSE on the original dominant-new-state AB panel, $C = 256$, $H = 96$, first 16 steps, and ten seeds. Settings are averaged within each seed before taking the mean.

Model	Seasonal AR	SEIR	Lorenz–63	Copying	Frequency
Chronos-2-Synth	0.926	1.053	1.965	**0.056**	0.173
PatchTST-FM-r1	1.090	1.055	1.082	0.978	0.177
Granite-PatchTST-r1	1.083	0.738	0.958	0.980	0.201
YingLong-110M	3.880	42.137	2.315	1.142	0.154
YingLong-300M	3.036	17.059	3.379	1.003	0.364
TimesFM-3	1.069	0.314	0.772	1.191	0.155
Toto-2 (1B)	**0.791**	**0.220**	**0.432**	0.082	0.089
TimesFM-2.5	1.008	0.495	1.047	0.749	0.238
PatchFM	0.927	0.579	1.298	1.035	**0.005**
Toto-1	1.277	0.315	1.694	1.138	0.409
Sundial-128M	1.048	0.931	1.053	1.199	0.208
TiRex-1.1-gifteval	0.837	0.266	1.422	0.359	0.310
Kairos-10M	3.137	14.739	4.642	1.138	0.649
Kairos-23M	3.347	58.195	6.034	1.317	0.736
Kairos-50M	2.683	3.423	3.132	1.058	0.678
FlowState-9.1M	1.175	0.774	3.716	0.958	0.896
FlowState-r1.1	0.837	0.570	3.176	1.132	0.649
Granite-FlowState-r1.1	0.820	0.403	2.529	1.138	0.645
Reverso-Small	1.038	0.295	3.135	1.015	0.618
Goia-nano-v0	2.523	0.824	3.371	1.043	0.885
CleanTS-65M	4.536	30.440	4.851	0.951	0.587

Table 8: Marginal probability losses, $\mathrm{CRPS}_{\mathrm{Q9}}$ (lower is better), at $C = 512$ and $H = 96$. Means are over parameter settings within each of three history seeds and then over seeds. Quantile and sample models use the same nine-decile scoring grid. A dash for Reverso denotes the absence of an eligible predictive distribution; a point-mass substitute is not used.

Model	OU	GARCH	Counts	Queue	HMM
Chronos-2-Synth	0.423	0.593	1.320	**4.449**	0.895
PatchTST-FM-r1	0.428	0.594	1.281	4.819	0.901
Granite-PatchTST-r1	0.425	0.593	1.294	4.984	0.903
YingLong-110M	0.500	0.624	1.377	5.422	1.031
YingLong-300M	0.482	0.609	1.302	6.928	0.930
TimesFM-3	**0.422**	0.593	1.311	4.689	0.892
Toto-2 (1B)	0.437	**0.591**	**1.262**	4.638	**0.891**
TimesFM-2.5	0.425	0.593	1.317	4.565	0.906
PatchFM	0.439	0.596	1.341	4.679	0.932
Toto-1	0.463	0.596	1.341	4.596	0.906
Sundial-128M	0.453	0.629	1.406	5.321	1.010
TiRex-1.1-gifteval	0.443	0.596	1.284	4.773	0.904
Kairos-10M	0.444	0.594	1.343	4.989	0.910
Kairos-23M	0.437	0.603	1.535	6.133	0.932
Kairos-50M	0.457	0.593	1.286	4.910	0.909
FlowState-9.1M	0.439	0.601	1.299	4.684	0.915
FlowState-r1.1	0.446	0.602	1.294	4.724	0.921
Granite-FlowState-r1.1	0.442	0.599	1.309	4.721	0.906
Reverso-Small	–	–	–	–	–
Goia-nano-v0	0.452	0.594	1.275	4.453	0.906
CleanTS-65M	0.523	0.615	1.423	7.186	1.045

Table 9: Original restricted-interface panels at $C = 512$, $H = 96$, and three seeds. Copying is first-step cross-copy NRMSE (lower is better). VAR, Factor, and Graph are first-eight-step NRMSE gains from joint prediction over independent prediction in the 20-channel systems (higher is better); negative values indicate harm. Response is the paired response-error ratio averaged equally over six non-SEIR systems (lower is better). Toto-2 has no eligible future-input result.

Model	Copying	VAR gain	Factor gain	Graph gain	Response
Chronos-2-Synth	0.769	-0.009	<u>0.055</u>	0.012	<u>0.753</u>
TimesFM-3	**0.731**	<u>-0.008</u>	0.052	-0.021	**0.275**
Toto-2 (1B)	0.786	-0.056	0.003	<u>0.025</u>	–
Toto-1	<u>0.760</u>	**0.062**	0.015	0.002	0.998
CleanTS-65M	0.768	-0.146	**0.519**	**0.185**	0.996

B.4 Comparison with Aggregate Forecasting Benchmarks

We compare the profiles with GIFT-Eval (Aksu et al., 2024) and TIME (Qiao et al., 2026), using the archived public tables from September 18, 2026. Published variant names and release information match 21 GIFT-Eval entries and seven TIME entries. The evaluated TiRex-1.1-gifteval checkpoint is excluded from the TIME match because the listed version differs. We do not claim that external weight hashes, runtimes, or preprocessing are identical to ours, and we do not rerun the external benchmarks.

Table 10 reports Spearman correlations, denoted by ρ, between higher capability scores and better external performance. GIFT-Eval uses its mean task MASE rank for this comparison, whereas the MASE column in Table 6 is the published aggregate loss; these are different quantities. TIME uses normalized MASE. Rounded leaderboard ties receive average ranks. The table also gives a GIFT-Eval comparison in which each model family has equal weight.

Table 10: Agreement with external forecasting performance. Positive ρ indicates better-to-better agreement. GIFT columns show model-weighted and family-weighted comparisons; TIME uses the matched model set. A coefficient is not reported for the two-model future-input overlap. These descriptive correlations do not establish independence or causal effects of model design.

Capability	GIFT M	ρ	Families	Family ρ	TIME M	ρ
Structure Induction	21	0.721	13	0.619	7	0.714
History Retrieval	21	0.545	13	0.608	7	0.750
State Tracking	21	0.599	13	0.591	7	0.893
Uncertainty Modeling	20	0.696	12	0.650	7	0.893
Channel Modeling	5	-0.564	4	-0.200	3	0.000
Exogenous Conditioning	4	0.800	4	0.800	2	–

The broad panels show positive agreement with external performance, but agreement is incomplete. In the GIFT-Eval comparison, the correlations range from 0.545 for History Retrieval to 0.721 for Structure Induction. The negative Channel Modeling association, $\rho = -0.564$, concerns only five models and a score that emphasizes incremental gains. Under equal family weighting its magnitude decreases to 0.200. This small panel does not support a general trade-off between aggregate forecasting quality and cross-channel modeling. The positive Exogenous Conditioning coefficient likewise rests on only four models. The seven-model TIME overlap is a separate comparison and cannot be used as if it covered all 21 models.

B.5 Sensitivity and Scope of the Overall Results

Table 11 compares alternative summaries of the saved predictions with the original profiles. It reports changes in model order, not new model runs. Structure Induction and History Retrieval remain highly correlated with the original ordering under the listed changes. State Tracking is more sensitive to the type of transition: restricting it to frequency changes gives $\rho = 0.692$ and a maximum shift of ten rank positions. Adding calibration to the uncertainty criterion preserves substantial agreement ($\rho = 0.937$), but exchanges the first two positions of TimesFM-3 and Toto-2.

Table 11: Sensitivity to the profile definition. Correlations compare each alternative with the original rank on its eligible set. Maximum change is measured in rank positions; half positions arise from ties. No Exogenous Conditioning alternative is included in this audit.

Alternative	Models	ρ	Max. rank change
Induction: all available C/H	21	0.961	4.0
Induction: without SEIR	21	0.960	3.5
Retrieval: all noise levels	21	0.976	4.5
Tracking: frequency changes only	21	0.692	10.0
Tracking: periodic changes	21	0.885	7.0
Tracking: including recovery	21	0.977	2.5
Uncertainty: loss and calibration	20	0.937	4.5
Channels: full forecast horizon	5	0.205	3.0
Channels: absolute joint accuracy	5	0.205	3.0
Channels: near and full horizon	5	0.359	2.0

Relation to later diagnostics. The fixed profiles are complemented by function continuation, composition, periodic decomposition, history interventions, uncertainty-horizon analysis, and informative-channel-count experiments. Their model sets and forecast conditions are reported separately. In particular, the accepted arithmetic, geometric, shifted-exponential, and logarithmic continuation results cover 546 requests for the original 21 models: 390 valid forecasts and 156 unsupported requests. These deterministic cases are Structure Induction diagnostics, not additional independent seeds for the original profiles.

Later induction evaluations also include YingLong-50M and Toto-2 at 313M, 1B, and 2.5B. These versions do not enlarge the original 21-model tables. A rerun of the 1B version is an additional evaluation of an existing version, not a new model. Coverage must be matched within each scale comparison before attributing differences to model size.

Metric availability. The saved original results support more point-error comparisons than probabilistic comparisons. In particular, probabilistic outputs were not retained for many of the early Structure Induction, History Retrieval, and State Tracking cases. A model's native probabilistic interface is not evidence that those distributions were saved in every run. We therefore do not fill missing probabilistic scores from point forecasts or treat an undefined MASE denominator as a zero error. The tables here summarize available, verified measurements; capability-specific analyses retain their own conditions, reference predictors, and uncertainty estimates.

C Structure Induction

This section examines whether a model can continue a predictive structure from its observed history. We organize the probes into non-chaotic and chaotic dynamics. Elementary function continuation and component mixtures belong to the non-chaotic group. Periods, growth rates, latent states, and true components are evaluator information unless a control explicitly supplies them. The evaluated TSFM remains frozen. We write unobserved simulator coordinates as $\mathbf{y}$ and observed targets as $\mathbf{x}$.

The original three-system results are reported in Appendix B. The additional panels address two ambiguities: changing C should not silently move the forecast origin, and low absolute SEIR error can arise when the target is nearly flat. Last-value and local-linear references therefore remain necessary; a high rank among TSFMs does not imply that a model beats these simple solutions.

C.1 Non-chaotic Dynamics

C.1.1 Seasonal AR and SEIR

The seasonal recurrence is

$$x_t = \phi x_{t-P} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{9}$$

The original structure panel uses $P = 24$, $\sigma = 0.08$, and $\phi = 0.9, 0.85, 0.98$ for the ordinary, interpolation, and extrapolation settings, relative to protocol support $\{0.7, 0.8, 0.9\}$. These labels describe our parameter construction, not verified inclusion in or exclusion from pretraining. Initial-state changes and observation noise are separate conditions.

The SEIR state (Hethcote, 2000) comprises susceptible, exposed, infected, and recovered fractions, denoted by the superscripts below:

$$\begin{aligned} \dot{y}^{\mathrm{sus}} &= -\beta y^{\mathrm{sus}} y^{\mathrm{inf}}, \qquad \dot{y}^{\mathrm{exp}} = \beta y^{\mathrm{sus}} y^{\mathrm{inf}} - \gamma y^{\mathrm{exp}}, \\ \dot{y}^{\mathrm{inf}} &= \gamma y^{\mathrm{exp}} - \upsilon y^{\mathrm{inf}}, \quad \dot{y}^{\mathrm{rec}} = \upsilon y^{\mathrm{inf}}. \end{aligned} \tag{10}$$

Here β is transmission, γ progression, and υ recovery. The original settings use $\gamma = 0.2$, $\upsilon = 0.1$, $\beta \in \{0.5, 0.45, 0.8\}$, sampling interval 0.25, and the same integration scheme. Additional active-dynamics tests use $\gamma = 1/5.2$, $\upsilon = 0.1$, $\beta \in \{0.35, 0.5, 0.7\}$, interval 0.1, forecast boundary 320, and $C \in \{64, 128, 256\}$. Early, active, and depleted initial states are kept separate.

C.1.2 Elementary Functions and Component Mixtures

Let $\tilde{t} = t - T - 1$ be time relative to the first future observation, so that histories have $\tilde{t} < 0$. Amplitudes use a, phases use ψ, and all periods retain P. The shared second component is $f_2(\tilde{t}) = a_2 \sin(2\pi\tilde{t}/P_1 + \psi_2)$. The four deterministic first-component families are

$$\begin{aligned} f_1(\tilde{t}) &= b\tilde{t}, && b = a_1/256, \\ f_1(\tilde{t}) &= a_1 \sin(2\pi\tilde{t}/P_2 + \psi_1), \\ f_1(\tilde{t}) &= a_1 e^{-\max(\tilde{t}+128,0)/256} \sin(2\pi\tilde{t}/P_2 + \psi_1), \\ f_1(\tilde{t}) &= a_1 \sin\Big(2\pi[\tilde{t}/P_2 + 0.512 \log(1 + e^{\tilde{t}/512})] + \psi_1\Big). \end{aligned} \tag{11}$$

Amplitudes are sampled from $[0.5, 1.5]$ and phases from $[-\pi, \pi]$. The grid panel uses $(P_1, P_2) = (24, 96)$; the off-grid panel samples periods from $[20, 44]$ and $[70, 110]$. A fifth family replaces f_1/a_1 with a stationary AR(1) process, $\phi = 0.8$, $\sigma = 0.15$, and is analyzed separately because its future remains random. Component observation noise is either zero or 0.05 times the corresponding amplitude, and the observed sum contains the same component noises. The design uses $C \in \{128, 512, 2048\}$, $H \in \{96, 256\}$, a fixed forecast boundary at 2304, and 16 independent seeds.

The saved component calls also provide pure linear, fast-period, and slow-period forecasts. Double-period and trend-plus-period forecasts use the sum branch. Function-specific errors normalize by each function's own reference-history variance; composition comparisons instead use one common denominator. These normalizations must not be interchanged when subtracting errors.

Direct versus component-wise prediction. Both branches predict $x = f_1 + f_2$, but the component branch receives the two true histories and makes two forecasting calls. On the common squared scale $\nu_1^2 + \nu_2^2$, define

$$\mathrm{Gap}_{\mathrm{comp}} = \frac{\mathrm{MSE}(\hat{x}^{\mathrm{direct}}, x) - \mathrm{MSE}(\hat{f}_1 + \hat{f}_2, x)}{\nu_1^2 + \nu_2^2}. \tag{12}$$

The numerator evaluates the sum of component predictions, not the sum of component MSEs; component errors can cancel. A positive gap quantifies benefit from privileged decomposition and does not establish that a deployable history-only decomposition would achieve the same gain.

Table 12: Complete deterministic composition comparison at $C = 512$, $H = 96$, and no observation noise. Four families and two period panels yield 128 paired cases per model, grouped within 16 seeds. Both errors use the common component-variance denominator. The gap is computed before rounding.

Model	Direct NMSE	Component NMSE	Gap
Chronos-2-Synth	0.0103	0.0016	0.0087
PatchTST-FM	0.0045	0.0016	0.0028
Granite-PatchTST	0.0039	0.0015	0.0024
YingLong-110M	0.5307	0.6165	-0.0858
YingLong-300M	0.7194	0.7526	-0.0333
TimesFM-3	0.0020	0.0009	0.0012
Toto-2	0.0039	0.0011	0.0028
TimesFM-2.5	0.0034	0.0011	0.0023
PatchFM	0.0134	0.0065	0.0069
Toto-1	0.0423	0.0051	0.0372
Sundial-128M	0.0406	0.0199	0.0207
TiRex	0.0055	0.0119	-0.0064
Kairos-10M	0.5790	0.7515	-0.1725
Kairos-23M	0.8827	0.7291	0.1536
Kairos-50M	0.4586	0.4400	0.0187
FlowState-9.1M	0.4247	0.3808	0.0439
FlowState-r1.1	0.5399	0.3571	0.1828
Granite-FlowState	0.3715	0.2215	0.1499
Reverso-Small	0.0800	0.0166	0.0635
Goia-nano-v0	0.2975	0.1663	0.1311
CleanTS-65M	1.4594	1.5214	-0.0620
Ridge AR32	0.0021	0.0017	0.0004

Chronos-2-Synth, TimesFM-3, and Toto-2 retain all 128 cases when both component errors are required to be at most 0.05 on the common scale, and their positive composition gaps remain. TiRex retains 124 cases under this criterion but has a negative gap. CleanTS also has a negative gap, with large errors in both branches. These exceptions distinguish accurate direct composition from a small or negative difference caused by poor component predictions. Off-grid periods do not impose a universal penalty: the direct-error change is negative for Chronos-2-Synth but positive for TimesFM-3 and TiRex. This is parameter sensitivity, not evidence that off-grid periods were absent from pretraining.

C.1.3 Arithmetic, Geometric, Exponential, and Logarithmic Continuation

The noiseless continuation functions are

$$x_t = bt, \qquad x_t = \chi^t, \qquad x_t = e^{t \log \chi} - 1, \qquad x_t = \log(1 + t/\xi). \tag{13}$$

The slope is $b \in \{2, 4, 8, 16\}$, the growth factor is $\chi \in \{1.001, 1.002, 1.004\}$, and the logarithmic time scale is $\xi \in \{1, 32, 256\}$. We reserve d for dependency distance and r for model rank; neither is reused for sequence parameters. Each C contains 13 parameter cases, with $C = 1028$ or 2056 and $H = 96$. Histories start at $t = 0$, so $T = C - 1$: increasing C also changes the forecast origin and value range. Exponential and geometric cases differ by a constant shift of one and are not independent structural families.

The original 21-model set produces 546 requests: 390 valid forecasts and 156 unsupported requests. Eighteen models have at least one supported context; PatchFM and both YingLong versions have no valid forecast at these two lengths. The two batches use identical inputs and targets. Unsupported requests are retained rather than cropped or externally padded. All cases use the fixed execution seed 1723; parameter values are not independent random replications.

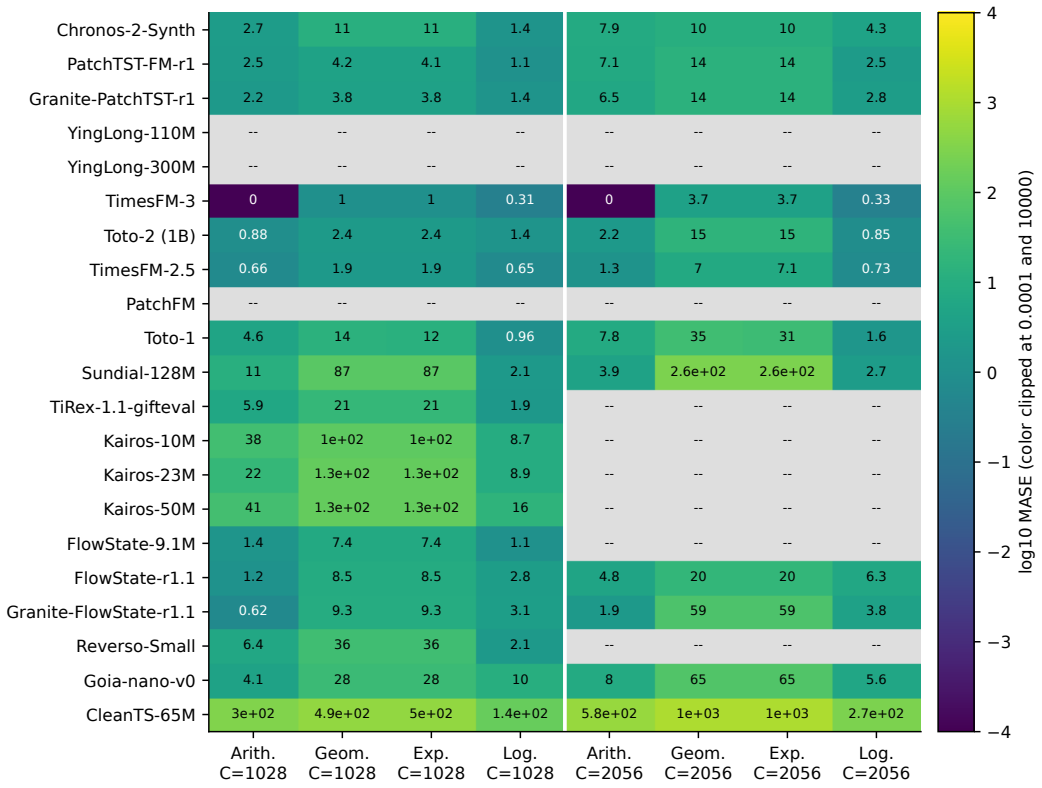


Figure 7: Mean MASE within each function and context, using all original 21 models. Grey cells denote unsupported requests. Numbers retain exact zeros; the logarithmic color map clips values below 10^{-4} and above 10^{4} for display. This is a deterministic parameter-grid comparison, not a seed-based significance test.

TimesFM-3 has the lowest mean MASE in all eight function-by-context groups, including exact saved predictions for all arithmetic cases. This does not establish exact continuation for arbitrary slopes or horizons. The earlier single arithmetic test, $b = 2$, $C = 1024$, $H = 96$, also has zero saved error for TimesFM-3, whereas Toto-2 has MAE 1.785 and RMSE 2.291. Simple references remain competitive in other families: at $C = 2056$, last-increment linear extrapolation has logarithmic MASE 0.182, below the best model's 0.332.

C.1.4 Forecast Residuals at Output-Block Boundaries

Residuals retain the definition $e_h = \hat{x}_{T+h} - x_{T+h}$ from Appendix A. For a recorded native output-block length L, the local jump is

$$\text{Jump}(h) = |e_{h+1} - e_h|, \qquad 1 \le h < H. \tag{14}$$

We compare its mean at boundaries $h = L, 2L, \ldots < H$ with its mean at other positions. A ratio above three is a descriptive flag, not a significance threshold or a minimum absolute error requirement. Zero divided by zero is undefined and retained as such.

Boundary reinforcement is common for some models but not universal. PatchTST-FM has no flagged case in either context, whereas Chronos-2-Synth has 11/13 in each. TimesFM-2.5, Sundial, and Goia do not cross their recorded full output blocks within this 96-step request; these cases

Table 13: Cases with boundary-to-interior mean residual jump above three, out of 13 parameter cases at each context. Only models with a recorded fixed block length shorter than $H = 96$ and valid predictions are included. Dashes denote unavailable comparisons. TimesFM-3 has four zero-jump arithmetic cases at each context; their ratios are undefined.

Model	Block length L	$C = 1028$	$C = 2056$
Chronos-2-Synth	16	11/13	11/13
PatchTST-FM-r1	64	0/13	0/13
Granite-PatchTST-r1	64	5/13	0/13
TimesFM-3	64	0/13	0/13
Toto-2 (1B)	32	3/13	6/13
Toto-1	64	8/13	10/13
TiRex-1.1-gifteval	32	13/13	–
Kairos-10M	64	13/13	–
Kairos-23M	64	13/13	–
Kairos-50M	64	13/13	–
Reverso-Small	48	13/13	–
CleanTS-65M	32	13/13	13/13

cannot test full-block stitching. FlowState's variable-horizon basis decoder is not assigned a stitching period from its internal decoder patch length. The observed discontinuities motivate continuity diagnostics but do not isolate the causal role of patch training.

C.1.5 History-Estimated Periodic Decomposition

Following the residual-cycle construction of CycleNet (Lin et al., 2024), we test subtraction and restoration of a periodic component around a frozen forecaster. Our template is estimated from the supplied history with a known period; it is not a jointly trained CycleNet component. In a shared historical normalization, let $\mathrm{PhaseMean}_P(x)$ denote the zero-centered phase-average template repeated with period P. The forecast is

$$\hat{x}^{\mathrm{decomp}}_{T+1:T+H} = \mathrm{Forecast}(x_{T-C+1:T} - \mathrm{PhaseMean}_P(x)_{T-C+1:T}) + \mathrm{PhaseMean}_P(x)_{T+1:T+H}. \tag{15}$$

Predictions are returned to the original units before scoring. The template is aligned by phase at the forecast origin and uses no future observations. A separate true-periodic-component branch is privileged.

The residual is a constant, white noise, or stationary AR(1) with $\phi = 0.8$; the periodic component is single-frequency or double-harmonic, with $P \in \{24, 48\}$ and $a \in \{0.25, 1, 4\}$. We use $C \in \{128, 512\}$, $H = 96$, and ten new seeds. Five models selected from the earlier function comparison each have 2160 records across the direct, history-template, and true-template branches; 3600 direct/history pairs are complete.

Table 14: Full-horizon MSE reduction in percent: $100(1 - \overline{\mathrm{MSE}}_{\mathrm{decomp}}/\overline{\mathrm{MSE}}_{\mathrm{direct}})$. Conditions are averaged within each seed before aggregation; brackets are descriptive paired-seed 95% bootstrap intervals using 2000 resamples. Positive values favor decomposition.

Model	$C = 128$	$C = 512$
TimesFM-3	10.31 [0.43, 20.01]	4.48 [-0.57, 9.17]
Toto-2 (1B)	8.60 [-2.09, 18.19]	2.06 [-0.53, 5.19]
TimesFM-2.5	4.63 [-2.02, 11.75]	5.35 [-0.28, 10.05]
Granite-PatchTST-r1	12.37 [1.73, 22.30]	4.88 [0.30, 9.13]
Chronos-2-Synth	7.07 [-1.51, 16.27]	7.06 [4.80, 9.68]

Figure 8 reports the residual-specific results.

All five aggregate means improve, but some intervals include zero. More specifically, AR mixtures improve in both contexts, while all five white-noise mixtures worsen at $C = 128$. At $C = 512$, this harm is smaller and Chronos has a small positive mean. Repeating finite-sample noise in the

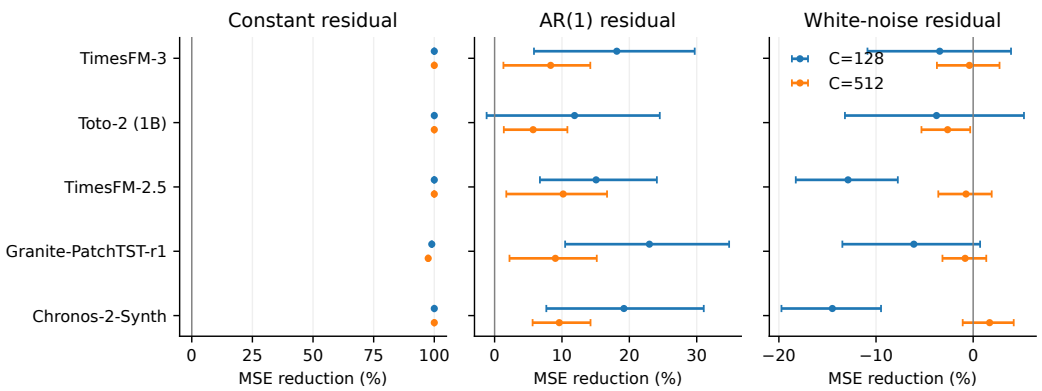


Figure 8: History-template decomposition by residual type, $H = 96$. Points and intervals use the same ten-seed aggregation as Table 14. Horizontal scales differ across residual types. Constant mixtures can have very small direct errors, so a large relative reduction need not imply a large absolute benefit.

phase template is consistent with this residual-dependent pattern, but the current contrasts do not uniquely identify the mechanism. The accepted experiment contains the selected five models, not the separately planned bottom-five group.

C.1.6 Model Size and Fixed-Origin Sine Tests

The accepted scale supplement evaluates Toto-2 at 313M, 1B, and 2.5B on the 26 sequence cases above and 12 sine cases per model. Sine histories use $x_t = \sin(2\pi\tilde{t}/P)$, $P \in \{32, 64, 128, 63\}$, $C \in \{512, 1024, 2056\}$, and $H = 192$. All contexts share the same future and are suffixes of the same history. This differs from the sequence protocol where changing context moves the origin. YingLong-50M contributes eight valid sine forecasts; its other 30 requests are unsupported by native context rules.

Table 15: Toto-2 scale comparisons: mean MASE within each fixed function and context. Sequence families use 3 or 4 parameter cases and $H = 96$; sine uses four periods and $H = 192$. Values are descriptive parameter-grid means, not independent-seed estimates.

Function	C	313M	1B	2.5B
Arithmetic	1028	0.899	0.883	0.530
Arithmetic	2056	0.705	2.191	0.808
Geometric	1028	3.879	2.370	3.248
Geometric	2056	19.251	15.484	8.746
Shifted exponential	1028	3.879	2.371	3.248
Shifted exponential	2056	19.251	15.484	8.746
Logarithmic	1028	0.531	1.370	0.809
Logarithmic	2056	0.457	0.851	0.603
Sine	512	0.279	0.182	0.120
Sine	1024	0.222	0.122	0.073
Sine	2056	0.086	0.086	0.053

The 2.5B version has the lowest average sine error at all three contexts. Logarithmic continuation instead favors 313M at both contexts, and sequence results are not uniformly monotonic in size. Across 38 paired cases, the 313M/1B/2.5B versions have the lowest MASE in 15/6/17 cases, respectively; only 13 cases have non-increasing error across all three sizes. This supports function-dependent scale effects, not a universal scaling law.

The larger four-family function grid, the complete cycle-mixture attenuation experiment, and the original five-model sine batch have no accepted full analysis in the evidence used here. Their

planned cases do not contribute performance values or additional completed experiments to this section.

C.2 Chaotic Dynamics

C.2.1 Lorenz–63

The Lorenz–63 system (Lorenz, 2017) follows

$$\dot{y}_1 = 10(y_2 - y_1), \quad \dot{y}_2 = y_1(\zeta - y_3) - y_2, \quad \dot{y}_3 = y_1 y_2 - \tfrac{8}{3} y_3, \tag{16}$$

where $\zeta \in \{28, 26, 40\}$ is the Lorenz forcing parameter. It is distinct from the correlation coefficient ρ used in Appendix B. Coordinates are observed at intervals of 0.02 and forecast independently in the original structure experiment. Numerical integration uses fourth-order Runge–Kutta with five substeps per observation.

Additional Lorenz tests hold the forecast boundary at 2304 or 3072 while varying the supplied history.

D HISTORY RETRIEVAL

History Retrieval asks whether a model finds and uses predictive observations at the relevant historical locations. We separate distance, repeated evidence, and history-content interventions. A longer input can contain more relevant observations, more distractors, and a different target phase; these changes must be identified before interpreting a context-length effect.

D.1 LAG-COPY DISTANCE AND AVAILABLE EVIDENCE

Context-parroting results motivate examining the contribution of historical reuse to forecasting (Zhang & Gilpin, 2026). Our lag-copy construction separately controls dependency distance and the historical evidence available to the model.

The copying process is

$$x_t = x_{t-d} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{17}$$

The implementation's requested lags $8, 24, 32, 128, 512, 2048, 8192$ correspond to actual distances $d = 9, 25, 33, 129, 513, 2049, 8193$. The tested contexts are $C = 512, 2048, 8449$ and innovation standard deviations are $\sigma = 0, 0.05, 0.1$. A cell requires $C - d \geq 256$ observable source–successor pairs, in addition to the model's native context restrictions. Unavailable cells are retained. Noiseless copying repeats a random initial template, so repeated pairs are not statistically independent evidence.

The original profile uses $C = 512$, no noise, the first four distances, and first-step MSE relative to the zero-prediction reference. Distance stress at $C = 8449$ keeps the input length fixed while changing d. The later repetition analysis uses the first eight steps of saved $H = 96$ predictions, with 20 seeds; all eight direct sources lie inside the history because the shortest d is 9. It therefore answers a different question from a long-horizon rollout test.

D.2 REPETITION AND CONTEXT-LENGTH EFFECTS

The quantity C/d counts visible cycle lengths. It is not the seed count N and is not the number $C-d$ of historical source–successor pairs. For the first eight future steps, the exact number of same-phase occurrences in the history is $\lfloor (C+h-1)/d \rfloor$, for $h = 1, \ldots, 8$. Fractional C/d is used as a compact common horizontal coordinate.

At $d = 129$, increasing C from 512 to 8449 reduces Toto-2's mean MSE from 0.02889 to 0.001376 and TimesFM-3's from 0.84669 to 0.003106. Shorter distances can saturate or reverse: at $d = 9$, both models have higher mean error at $C = 8449$ than at $C = 2048$. Neither observation alone establishes a universal law relating repetition to error.

For an all-model comparison, we fix the common change $C = 512 \rightarrow 2048$ and the distances $9, 25, 33, 129$. We count how many distances have lower mean first-eight-step MSE in the longer context. Nine of the 19 models with complete paired coverage improve at at least three distances. This criterion describes mean directions; it is not a significance test or a claim of large benefit.

The original generator changes the forecast origin and template phase when C changes. Across distances, the initial template dimension also changes. There are at most three contexts per distance. Consequently, fitting a single power law to pooled points would confound distance, input length, and phase. The observed curves support model-dependent benefit from repeated history; they do not identify the causal effect of repetition alone.

D.3 SEASONAL-HISTORY INTERVENTIONS

The seasonal process in Equation 9 is also evaluated with dependency periods $P = 24, 48, 96$. Full history is compared with source removal, removal of other positions, a 16-step recent window, a source-focused crop, and added noise prefixes of length 32 or 128. Masking retains original timestamps when the interface supports it. Masks, shorter sequences, and prefix additions are different interventions and are not pooled as one "history removal" operation.

Each contrast keeps its forecast target and uses the declared common pre-intervention reference scale. A source-focused crop can make localization easier by providing evaluator-selected informa-

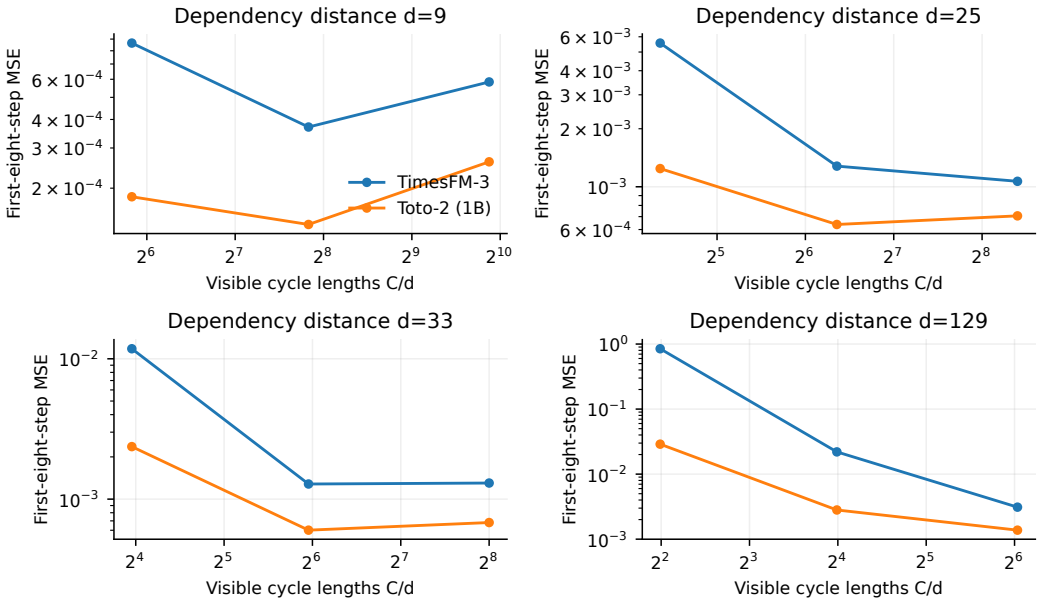


Figure 9: First-eight-step MSE versus visible cycle lengths C/d for TimesFM-3 and Toto-2, using the three available contexts and four fixed distances. Each point averages 20 seeds. Both axes are logarithmic. The two models illustrate strong long-distance improvements and short-distance saturation; the complete model comparison is retained in Table 16.

Table 16: Common context comparison across all 21 original models. Median change is the median percentage MSE change over the four distances; negative values favor $C = 2048$. The criterion requires lower mean MSE at at least three of four distances. Dashes indicate no complete common comparison.

Model	Distances improved	Median change (%)	Meets criterion
Chronos-2-Synth	3/4	-19.6	Yes
PatchTST-FM-r1	3/4	-63.8	Yes
Granite-PatchTST-r1	3/4	-28.8	Yes
YingLong-110M	1/4	2.2	No
YingLong-300M	2/4	5.1	No
TimesFM-3	4/4	-83.1	Yes
Toto-2 (1B)	4/4	-61.7	Yes
TimesFM-2.5	–	–	–
PatchFM	–	–	–
Toto-1	1/4	14.2	No
Sundial-128M	3/4	-44.8	Yes
TiRex-1.1-gifteval	1/4	32.2	No
Kairos-10M	2/4	-0.7	No
Kairos-23M	3/4	-7.8	Yes
Kairos-50M	3/4	-3.4	Yes
FlowState-9.1M	2/4	-1.5	No
FlowState-r1.1	2/4	-1.3	No
Granite-FlowState-r1.1	2/4	3.1	No
Reverso-Small	4/4	-36.4	Yes
Goia-nano-v0	2/4	2.8	No
CleanTS-65M	2/4	-1.3	No

tion. It is therefore a control, not an equal-information baseline. Similarly, sensitivity to a removed source indicates dependence on those observations but does not by itself identify an internal search or retrieval algorithm.

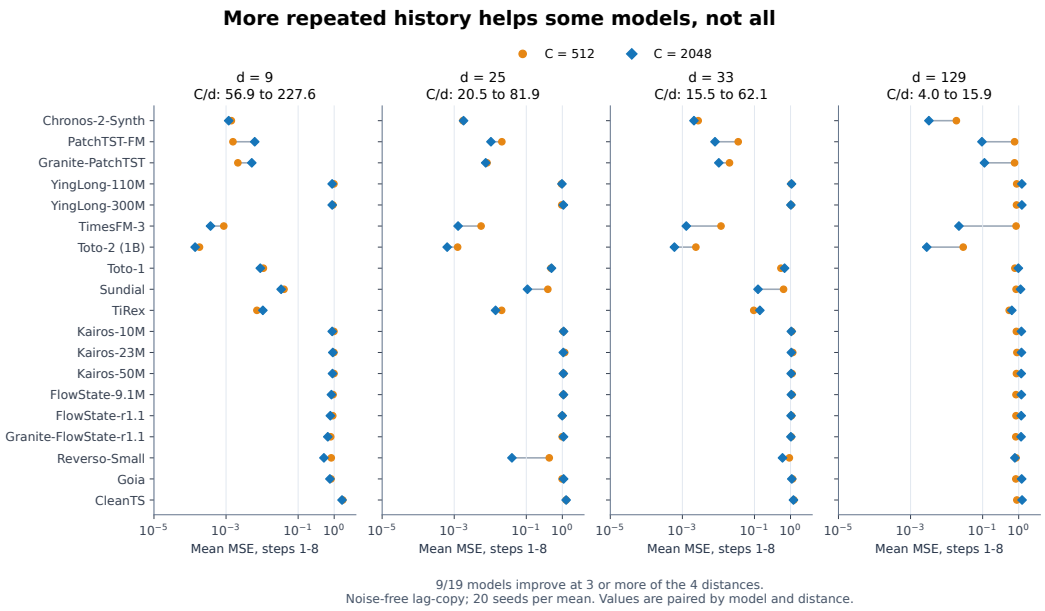


Figure 10: Prediction errors under a paired context-length comparison. Each panel fixes the dependency distance and compares $C = 512$ with $C = 2048$. Values are mean MSE over the first eight forecast steps, averaged across 20 seeds under noise-free conditions. Of the 19 models with complete paired coverage, nine (47.4%) achieve lower mean MSE at the longer context in at least three of the four distances. This criterion summarizes the direction of mean changes. Forecast origins and template phases also change with context length.

Returning-state old-history interventions are reported in Appendix E, where their value can be compared with adaptation to the active state. Real same-length interventions appear in Appendix I.

E State Tracking

State Tracking evaluates whether forecasts adapt when the active dynamics change. It separates prediction of each stable state, preference for the new state, and accurate recovery after transition. The state is evaluator metadata; no state-label input is supplied to the models in these experiments.

E.1 Switching Autoregressive and Smooth-Transition Processes

Regime-switching models provide a classical framework for changes in time-series dynamics (Hamilton, 1989). Here, the state schedules and paired controls are specified by our experimental design to examine adaptation and state recurrence.

The switching autoregressive process is

$$x_t = \mu_{z_t} + \phi_{z_t}(x_{t-1} - \mu_{z_t}) + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, 0.15^2), \tag{18}$$

where z_t is the active discrete state. Its mean and autoregressive coefficient are $(\mu_A, \phi_A) = (-2, 0.65)$ and $(\mu_B, \phi_B) = (2, -0.55)$. Abrupt AB, returning ABA, gradual drift, and stochastic switching schedules vary how the process moves between states. Here μ always denotes a level mean and ϕ an autoregressive coefficient.

For the smooth probes, let t_{switch} be the first transition center, t_{return} the return center, and τ the transition width. Define

$$\begin{aligned} w_{AB}(t) &= \frac{1}{1 + \exp(-(t - t_{\text{switch}})/\tau)}, \\ w_{ABA}(t) &= w_{AB}(t) - \frac{1}{1 + \exp(-(t - t_{\text{return}})/\tau)}. \end{aligned} \tag{19}$$

The transition width is τ; V continues to denote channel count. For frequency and amplitude transitions, the clean signal is

$$f(t) = [1 - w(t)]f_A(t) + w(t)f_B(t). \tag{20}$$

Frequency states have periods P_A, P_B with the same amplitude and phase; amplitude states retain the period and phase while the amplitude changes from $0.25a$ to a in the original protocol.

For valley, growth-to-plateau, and plateau-to-growth processes, the transition acts on the slope, rather than directly mixing levels:

$$\frac{df(t)}{dt} = [1 - w(t)]b_A + w(t)b_B, \tag{21}$$

with slopes proportional to $(-1, 1)$, $(1, 0)$, and $(0, 1)$, respectively. The implementation integrates this expression analytically using the softplus integral of the sigmoid. This distinction preserves the smooth path actually used in the experiment. Slopes retain b, as in the arithmetic probe.

Original AB transitions use widths 8 or 24, $C = 256$, $H = 96$, ten seeds, and noise fractions 0 or 0.03. ABA uses width 16, a return after 256 steps, and five seeds. The principal score covers the first 16 forecast steps. Additional frequency tests use $(P_A, P_B) = (64, 8), (64, 16), (96, 24)$, widths 8 or 32, $C = 64, 256, 1024$, and 16 seeds. Additional returning $(64, 16)$ transitions use width 16 and returns after 256 or 512 steps; noise fractions are 0 or 0.05.

E.2 Qualification, Recovery, and Censoring

Pure-state controls first establish whether the model can predict both stationary regimes. A transition then becomes eligible once the new state has weight at least 0.8 and the reference futures are sufficiently separated. Directional preference asks which pure-state future is closer to the forecast. Accurate recovery additionally requires clean NMSE at most 0.05 at two consecutive eligible origins. Choosing the new direction alone is insufficient.

Recovery delay is measured from new-state dominance to the first origin of a successful consecutive pair and is reported in units of the new state's period. Curves that do not recover in the observation range are censored. Delays are averaged only over recovered, qualified curves, alongside the recovery count. If no curve qualifies, the recovery rate is undefined; if no curve recovers, the delay is undefined rather than zero.

E.3 Old-State Content After a Return

The returning-state comparison retains $C = 1024$, recent observations, the forecast origin, and the future after return to A. Only observations earlier than the first switch minus its transition width are eligible for replacement, and at least 32 old observations are required. Shuffling preserves old values; a wrong-frequency replacement uses period $0.7P_A$ and matches the old segment's mean and variance. An unrelated AR(1) replacement uses $\phi = 0.7$. These replacements change old content while retaining its length and timestamps.

Table 17: Recovery and old-state dependence for all 21 models. Recovery uses noiseless AB frequency transitions, $C = 256$, with 96 curves per model from 16 seeds and six settings. Counts are recovered/qualified; delay is in new-state cycles among recovered curves. Old-state columns are NMSE increases after shuffling or wrong-frequency replacement at $C = 1024$, $H = 96$, averaged over returning-state settings within 16 seeds. These are different diagnostic panels.

Model	Recovered/qualified	Delay	Shuffle	Wrong frequency
Chronos-2-Synth	83/96	2.58	0.628	1.308
PatchTST-FM	94/96	2.64	0.699	0.727
Granite-PatchTST	76/96	2.17	0.879	0.981
YingLong-110M	30/52	2.66	-0.088	-0.058
YingLong-300M	1/54	1.19	0.188	0.039
TimesFM-3	87/96	2.24	0.469	0.633
Toto-2	89/96	1.81	0.707	0.950
TimesFM-2.5	49/96	2.27	–	–
PatchFM	96/96	1.03	0.361	0.411
Toto-1	31/96	3.40	0.143	0.287
Sundial-128M	0/32	–	0.683	1.334
TiRex	70/96	2.48	0.833	1.097
Kairos-10M	0/0	–	0.338	0.465
Kairos-23M	0/24	–	0.169	0.445
Kairos-50M	0/64	–	0.704	1.195
FlowState-9.1M	15/42	1.43	0.069	-0.003
FlowState-r1.1	15/38	1.80	0.057	-0.001
Granite-FlowState	8/51	1.82	0.216	0.048
Reverso-Small	0/77	–	0.569	1.015
Goia-nano-v0	0/68	–	0.499	0.711
CleanTS-65M	0/0	–	0.009	-0.118

PatchFM recovers all 96 qualified curves, with mean delay 1.03 cycles, whereas Reverso has 77 qualified curves and no recovery. This separates stable-state predictability from adaptation in the evaluated window. Toto-2 recovers 89/96 and TimesFM-3 87/96; their delays must be interpreted together with the censored fraction. Ridge AR32, a history-fitted reference, recovers 96/96 with mean delay 1.64 cycles. The 96 curves are not 96 independent seeds.

After a return, replacing old A with a matched wrong-frequency segment increases error for Chronos, TiRex, Toto-2, and TimesFM-3. Their predictions therefore depend on old structural content. CleanTS instead improves under that replacement, and several FlowState effects are small. The unrelated-AR replacement is another alteration of old A, not a placebo intervention on irrelevant timestamps. These controls support model-specific historical content dependence without distinguishing explicit retrieval from statistical fitting or other internal implementations.

F UNCERTAINTY MODELING

We evaluate marginal predictive distributions conditional on the supplied history. Distributional loss, interval coverage, and interval width measure different aspects of uncertainty. Native samples are converted to marginal quantiles for the common-grid comparison; a sampling interface alone does not establish calibrated uncertainty. This section does not infer joint-path or probability-composition consistency from marginal scores.

Assessing multivariate dependence requires additional joint evaluation tools, such as variogram-based scoring rules (Scheuerer & Hamill, 2015). These joint scores are outside the scope of the present marginal comparisons.

The evaluated processes are OU, GARCH, self-exciting counts, queues, HMMs, and Gaussian random walks. OU and random walks both probe uncertainty across forecast horizons: OU has mean reversion and a limiting conditional variance, whereas random-walk variance grows linearly without such a limit. The random-walk experiment uses its own matched protocol and is reported in Appendix F.3; it does not replace the original five-system scores used to construct the frozen capability profiles.

F.1 STOCHASTIC GENERATORS AND CONDITIONAL REFERENCES

Ornstein–Uhlenbeck process. For background on the Brownian-motion theory underlying this process, see Wang & Uhlenbeck (1945). The exact discrete observation transition is

$$x_{t+1} = e^{-\kappa} x_t + \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim \mathcal{N}\left(0, \frac{\vartheta^2}{2\kappa}(1 - e^{-2\kappa})\right). \tag{22}$$

The observation interval is one and the long-run mean is zero. Here κ is the mean-reversion rate and ϑ the continuous-time diffusion amplitude. The discrete innovation standard deviation remains σ from Appendix A and equals $\vartheta\sqrt{(1 - e^{-2\kappa})/(2\kappa)}$. Thus diffusion amplitude and observation-level innovation noise do not share one symbol. Original settings use $\kappa = 0.2$ and $\vartheta = 0.1, 0.4, 0.8$.

GARCH. The conditional-variance generator follows the GARCH framework (Bollerslev, 1986). The conditional innovation variance evolves as

$$x_t = \epsilon_t, \qquad \sigma_t^2 = \omega + \alpha x_{t-1}^2 + \delta \sigma_{t-1}^2, \qquad \omega = 1 - \alpha - \delta. \tag{23}$$

Original settings use $\alpha = 0.12$ and $\delta = 0.8$, giving unit stationary variance. Conditional innovations are Gaussian or standardized Student-t with five degrees of freedom, each with variance σ_t^2. The GARCH persistence parameter is δ, not the SEIR transmission parameter β. Reference futures inherit the conditional variance reached at the end of the visible history.

Self-exciting counts. Self-excitation provides the conceptual background (Hawkes, 1971); the finite-memory discrete-time generator used here is specified below. Counts follow a finite-memory discrete Poisson process,

$$x_t \mid x_{<t} \sim \text{Poisson}(\lambda_t), \qquad \lambda_t = 0.5 + \sum_{d=1}^{16} \varpi_d x_{t-d}. \tag{24}$$

Weights are proportional to $e^{-0.25(d-1)}$ and sum to 0.6 or 0.9. The symbol λ denotes arrival intensity, as in the queue below, and ϖ_d is a lag-specific excitation coefficient. This generator is not an exact continuous-time Hawkes process. Its last 16 observed counts initialize future branches.

Queues. The generator follows the standard Markovian queueing framework (Shortle et al., 2018). An M/M/K queue has arrival rate λ and per-server service rate $\iota = 1$. At customer count x, upward events have rate λ and downward events have rate $\iota \min(x, K)$. We simulate continuous-time exponential events and observe the count at unit intervals. Original pairs (λ, K) are $(0.8, 1), (2.7, 3), (0.95, 1), (1.1, 1)$, including a transient unstable configuration. Future paths start from the observed terminal count.

Hidden Markov models. The latent-state model and posterior filtering follow the standard HMM framework (Rabiner, 1989). The discrete state z_t switches between two values with probability 0.08 per step. Observations satisfy $x_t = \mu_{z_t} + \epsilon_t$, with means $\mu_0 = -1$, $\mu_1 = 1$ and Gaussian innovation standard deviation 0.8 or 1.2. The reference samples the terminal state from its posterior conditional on the observed history, not from the simulator's private realized state.

Together these settings give 13 original configurations. Every observed history has 100 independent conditional reference futures. Original profiles use three history seeds and $C = 512$, $H = 96$; saved additional panels use $C = 128$ and $H = 256$ as well. A separate 16-history supplement varies OU (κ, ϑ) over $[0.1, 0.4] \times [0.1, 0.6]$, GARCH (α, δ) over $[0.08, 0.16] \times [0.65, 0.8]$, and queue arrival rate over $[0.7, 1.1]$. Histories, not reference futures or forecast steps, are the independent statistical units.

F.2 History Baselines and State-Dependent Coverage

The empirical-history-quantile baseline tests whether conditional predictions improve over a distribution estimated only from historical marginal values. The separate supplement generally shows clearer gains on queues than on GARCH. At $C = 512$, low and high terminal queue states are split at count four. Ties produce 5 low-state and 11 high-state histories, rather than equal halves.

Table 18: All 20 probabilistic models on the independent-history supplement. Loss differences are model minus history-quantile scaled pinball loss, averaged over the reported contexts within 16 histories; negative values favor the model. Queue coverage is for requested 80% intervals at $C = 512$, with 5 low-state and 11 high-state histories. Coverage values are percentages.

Model	Queue loss difference	GARCH difference	Low coverage	High coverage
Chronos-2-Synth	-0.0520	0.0004	69.8	77.6
PatchTST-FM	-0.0453	0.0019	79.7	75.7
Granite-PatchTST	-0.0507	0.0012	83.0	74.4
YingLong-110M	0.0135	0.0311	47.4	55.3
YingLong-300M	0.0026	0.0340	72.1	56.3
TimesFM-3	-0.0388	0.0003	81.7	70.8
Toto-2	-0.0515	0.0010	80.1	77.9
TimesFM-2.5	-0.0514	0.0004	82.5	79.7
PatchFM	-0.0272	0.0047	59.6	58.3
Toto-1	-0.0475	0.0024	72.7	72.5
Sundial-128M	-0.0072	0.0331	43.8	44.2
TiRex	-0.0532	0.0025	80.5	74.7
Kairos-10M	-0.0214	0.0064	76.7	82.5
Kairos-23M	-0.0228	0.0055	74.2	66.8
Kairos-50M	-0.0281	0.0041	72.5	68.5
FlowState-9.1M	-0.0439	0.0092	68.7	61.8
FlowState-r1.1	-0.0419	0.0092	59.7	56.8
Granite-FlowState	-0.0386	0.0080	60.8	57.0
Goia-nano-v0	-0.0534	0.0020	78.9	73.9
CleanTS-65M	0.0476	0.0216	42.6	28.3

For example, TimesFM-3's queue intervals cover 81.7% in the low-state group and 70.8% in the high-state group, while Toto-2 covers 80.1% and 77.9%. Mean interval widths also change with state: 6.57/14.90 for TimesFM-3 and 6.63/18.02 for Toto-2. These small strata reveal a conditional diagnostic; they do not establish population calibration. Negative average pinball excess can coexist with undercoverage or histories on which the model loses to the baseline.

F.3 Random-Walk Diffusion Across Forecast Horizons

A separate 20-model experiment uses the Gaussian random walk

$$x_{t+1} = x_t + \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim \mathcal{N}(0, \sigma^2), \qquad x_{T+h} \mid \mathcal{O}_T \sim \mathcal{N}(x_T, h\sigma^2). \tag{25}$$

For a normal quantile function with zero mean and unit variance,

$$x^{(q)}_{T+h,\text{oracle}} = x_T + \sigma\sqrt{h}\ \text{NormalQuantile}(q),$$
$$\mathbb{E}[(\hat{x}_{T+h} - x_{T+h})^2 \mid \mathcal{O}_T] = (\hat{x}_{T+h} - x_T)^2 + h\sigma^2. \tag{26}$$

The oracle interval width grows as $\sqrt{h}$. Analytic coverage and expected pinball losses separate forecast errors from the increasing irreducible uncertainty. The oracle knows σ; a feasible Gaussian reference estimates it from historical increments. A fixed-width Gaussian forecast is a negative control. None of these reference parameters is supplied to the model.

The grid uses $C = 128, 256, 512$, $H = 96, 256$, $\sigma = 0.1, 0.4, 0.8$, ten history seeds, and a common forecast boundary at 512. All 3600 model requests have valid saved outputs, comprising 180 cases per model and 633600 predicted time points across 20 models. Reverso-Small is excluded because the evaluated interface has no native probabilistic output. Shorter contexts are suffixes of longer contexts ending at the same origin; different requested horizons share future prefixes, and noise levels use paired standardized innovations. The ten independent history seeds are 390000–390009. Neither the true future, the noise parameter, nor oracle quantiles are supplied to a model. Parameter and horizon combinations share histories or innovations and are not independent replicates. Table 19 uses only the fixed $C = 512$, $H = 96$ panel. Its Q9 excess is the mean case-wise ratio of the model's expected nine-quantile loss to the oracle loss, minus one. Q9 uses the nine deciles and is a discrete approximation to CRPS, not exact CRPS. The history-estimated Gaussian reference knows the random-walk form but estimates its noise scale from the supplied increments. Statistical replication is by independent history, not by the paired context, horizon, or noise-level conditions. Only marginal forecast distributions are evaluated.

Table 19: Random-walk marginal uncertainty at $C = 512$, $H = 96$: mean over three noise levels and ten histories. Q9 excess and coverage are percentages; width is divided by the oracle width. The width exponent is the fitted log-width versus log-horizon slope, whose oracle value is 0.5. These are descriptive means, not significance rankings.

Model	Q9 excess	80% coverage	Width ratio	Exponent
Chronos-2-Synth	2.19	74.0	0.909	0.456
PatchTST-FM-r1	6.06	72.7	0.939	0.450
Granite-PatchTST-r1	6.11	73.3	0.957	0.433
YingLong-110M	21.31	57.2	0.776	0.363
YingLong-300M	26.27	61.6	0.924	0.215
TimesFM-3	7.17	78.1	1.058	0.486
Toto-2 (1B)	2.49	75.7	0.935	0.451
TimesFM-2.5	2.82	79.0	1.036	0.455
PatchFM	8.98	58.5	0.683	0.249
Toto-1	7.82	71.4	0.933	0.475
Sundial-128M	27.25	40.1	0.486	0.529
TiRex-1.1-gifteval	15.71	72.6	1.013	0.493
Kairos-10M	10.24	79.8	1.256	0.184
Kairos-23M	25.10	72.8	1.355	0.003
Kairos-50M	4.43	68.1	0.842	0.216
FlowState-9.1M	7.15	61.6	0.728	0.430
FlowState-r1.1	9.36	56.3	0.649	0.409
Granite-FlowState-r1.1	9.12	58.2	0.679	0.416
Goia-nano-v0	2.81	75.0	0.951	0.482
CleanTS-65M	49.32	23.0	0.367	-0.234
Known-noise oracle	0.00	80.0	1.000	0.500
History-estimated Gaussian	0.03	80.3	1.009	0.500
Fixed-width control	20.55	18.8	0.189	0.000

Full-grid and fixed-panel results. Across all 180 conditions per model, the mean case-wise Q9 excess is 4.05% for Chronos-2-Synth, 4.22% for Toto-2, 5.04% for Goia, and 6.28% for TimesFM-2.5. All 20 models have full-grid mean coverage below the nominal 80%. These full-grid aggregates differ from the fixed $C = 512, H = 96$ values in Table 19 and are not additional independent replications or significance rankings.

Sundial's stochastic forecasts have 40.1% coverage and width ratio 0.486 in this panel, compared with 75.7%/0.935 for Toto-2 and 79.0%/1.036 for TimesFM-2.5. Introducing sampling randomness therefore does not guarantee that the resulting marginal spread matches conditional uncertainty.

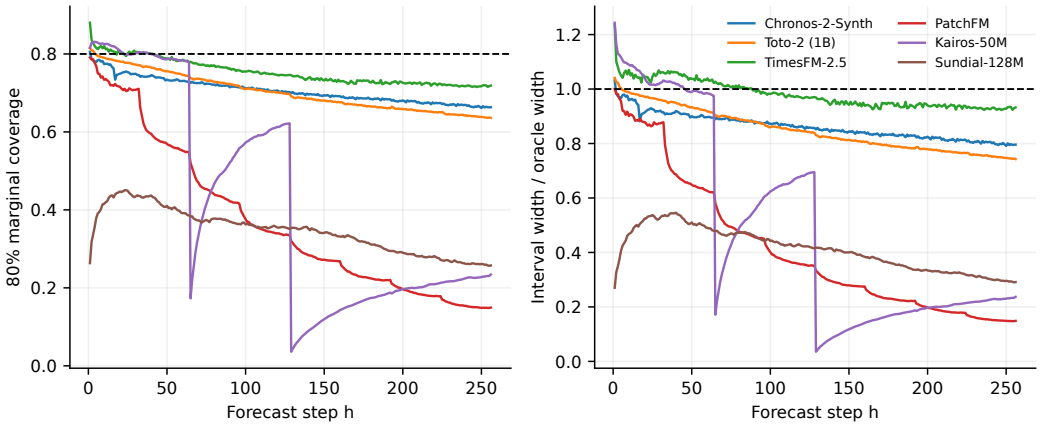


Figure 11: Analytic coverage and interval-width ratios at $C = 512$ for the same $H = 256$ requests, averaged over three noise levels and ten histories. Dashed lines denote oracle coverage 0.8 and width ratio 1. These selected models illustrate distinct horizon profiles; Table 19 gives the complete fixed-$H = 96$ panel.

For $H = 256$, PatchFM's coverage decreases from 76.8% in the first eight steps to 15.5% in the last 32, with width ratio decreasing from 0.954 to 0.154. Kairos-50M changes from 82.7% to 22.3%, while TimesFM-2.5 retains 71.8% coverage in the last 32 steps. These are matched portions of a single requested horizon. Comparing raw losses alone would mix changing task noise with model error.

Adjacent quantile crossings occur in 18 CleanTS, 30 YingLong-110M, and 15 YingLong-300M cases out of 180 per model. The available central interval endpoints are not inverted, so their coverage remains computable. We preserve native quantiles without sorting and do not manufacture unavailable levels.

F.4 Horizon Profiles in the Original Stochastic Suite

The original five-family $C = 512$, $H = 96$ forecasts can also be divided into steps 1–32, 33–64, and 65–96. TimesFM-2.5's mean 80% coverage is 80.1/80.6/81.1%, Toto-1's is 77.7/78.3/79.1%, and PatchFM's is 78.7/67.4/58.7%. The marked late-horizon decline is therefore not shared by all three NTP models. The $H = 96$ request also fits within TimesFM-2.5's 128-step output block, so NTP cannot be interpreted as one-step recursive decoding for every evaluated model.

CPM cumulative-window analyses retain the same distinction between raw loss and reference-adjusted loss. Their finite-reference loss estimate is not an oracle lower bound; a small negative difference does not imply performance better than the Bayes optimum. These horizon diagnostics complement the family-aggregated profile without replacing it or isolating the causal effect of a pretraining objective.

G Channel Modeling

G.1 Dependency Structures and Evaluation Units

Channel modeling asks whether a forecast benefits from information in other observed channels. The target channel, forecast origin, and target future are held fixed within each comparison. We distinguish additional informative observations from additional distractors: a larger input can contain more channels without containing more information about the target. Only models with a supported joint-channel interface enter this comparison; independent calls to a univariate model do not constitute joint prediction.

The original panel includes cross-copy, vector autoregression (VAR), latent-factor groups, and graph diffusion. Cross-copy places the relevant delayed observation in another channel. In the VAR construction (Lütkepohl et al., 2005), the target has coefficients 0.45 on its own previous value and 0.25 and 0.20 on two parent channels; the other channels have autoregressive coefficient 0.65. The latent-factor construction uses four autoregressive factors with characteristic-root modulus 0.94 and periods 24, 48, 64, and 96. Graph diffusion uses a path graph, with coefficient 0.55 on the identity and 0.30 on the row-normalized adjacency operator. These systems use innovation standard deviation $\sigma = 0.2$ and a 256-step burn-in. Nested observations of the original 64-channel systems use $V \in \{20, 24, 32, 64\}$. The extra VAR channels are distractors, whereas additional factor or graph observations can carry predictive information.

We report the gain as a ratio of aggregate losses,

$$\mathrm{Gain}_{\mathrm{joint}} = 100 \left(1 - \frac{\overline{\mathrm{NMSE}}_{\mathrm{joint}}}{\overline{\mathrm{NMSE}}_{\mathrm{target\ only}}} \right). \tag{27}$$

Positive values indicate lower error with joint input. The same target-specific reference scale ν is used for both losses. This quantity is not an average of seed-wise percentage gains, which can be dominated by very small denominators. Supplemental dependency experiments use 16 independent seeds and separately score the first 8 steps and the complete $H = 96$ forecast.

Table 20: Joint-input gains (%) at $C = 512, V = 20, H = 96$, with descriptive 95% paired-seed intervals. The first-eight-step and full-horizon columns use the same forecasts. All eligible models, baselines, and dependency families in this supplemental comparison are retained.

Model	System	First 8 steps	Full horizon
Chronos-2-Synth	Cross-copy	-0.19 [-3.75, 3.27]	-0.18 [-0.66, 0.24]
Chronos-2-Synth	VAR	-0.36 [-2.74, 1.32]	-0.25 [-0.56, 0.11]
Chronos-2-Synth	Factors	-4.22 [-21.20, 6.89]	1.25 [-3.71, 8.46]
CleanTS-65M	Cross-copy	2.77 [-33.40, 35.96]	14.17 [-3.49, 30.73]
CleanTS-65M	VAR	1.85 [-33.61, 24.62]	11.29 [4.12, 18.28]
CleanTS-65M	Factors	4.65 [-89.08, 47.40]	13.25 [-1.22, 23.84]
Toto-1	Cross-copy	12.36 [-0.95, 25.13]	6.27 [0.35, 13.20]
Toto-1	VAR	10.96 [-0.57, 19.43]	4.23 [1.92, 6.41]
Toto-1	Factors	20.15 [-1.76, 32.47]	0.19 [-4.91, 6.17]
Independent baseline	Cross-copy	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Independent baseline	VAR	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Independent baseline	Factors	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Linear baseline	Cross-copy	-6.72 [-22.23, 5.64]	-1.02 [-2.10, 0.13]
Linear baseline	VAR	3.05 [-8.45, 11.70]	0.00 [-1.70, 1.48]
Linear baseline	Factors	-3.20 [-22.58, 14.38]	-17.37 [-33.31, -5.01]
TimesFM-3	Cross-copy	3.82 [-2.59, 9.89]	1.14 [-0.02, 2.42]
TimesFM-3	VAR	0.19 [-2.42, 1.53]	-0.40 [-1.05, 0.20]
TimesFM-3	Factors	7.21 [-17.71, 29.29]	0.93 [-2.31, 3.81]
Toto-2 (1B)	Cross-copy	16.75 [5.13, 27.14]	1.90 [-0.55, 4.67]
Toto-2 (1B)	VAR	5.95 [-3.06, 12.67]	1.67 [-0.04, 2.99]
Toto-2 (1B)	Factors	1.49 [-8.34, 9.09]	-0.73 [-1.93, 0.47]

The dependence on forecast span is substantial. For Toto-2, cross-copy gains decrease from 16.75% over the first eight steps to 1.90% over the complete horizon. The corresponding intervals are

[5.13,27.14]% and [−0.55,4.67]%. Thus, useful short-range cross-channel information need not materially improve the full-horizon average. The complete table also retains negative gains and intervals that include zero; the existence of a channel interface alone does not establish successful information use.

G.2 Increasing the Number of Informative Channels

A separate experiment increases V through

$$1, 2, 4, 8, 16, 20, 24, 32, 48, 64, 96, 128, 192, 256, 384, 512.$$

It uses three linear-Gaussian constructions: shared observations of a latent process, weak delayed observations, and distributed parent signals. Latent persistence is sampled between 0.93 and 0.98, with a 1024-step burn-in. The target trajectory and origin are identical across nested channel sets. Histories have $C \in \{128, 512\}$, forecasts have $H = 96$, and 16 additional seeds are used. One-step error is the primary endpoint; the first 8, 16, and 96 steps are supplementary endpoints.

A filter with the known generating mechanism supplies the conditional mean and variance using observed histories only. Future observations are not provided. For a fixed prediction, conditional squared-error risk decomposes as

$$\begin{aligned}&\mathbb{E}\left[(\widehat{x}_{T+h,j} - x_{T+h,j})^2 \mid \mathcal{O}_T^{(V)}\right] \\ &\quad = \mathrm{Var}(x_{T+h,j} \mid \mathcal{O}_T^{(V)}) + \left[\widehat{x}_{T+h,j} - \mathbb{E}(x_{T+h,j} \mid \mathcal{O}_T^{(V)})\right]^2. \end{aligned} \tag{28}$$

Here $\mathcal{O}_T^{(V)}$ contains the available V channel histories. The reference scale is fixed across V. The first term measures remaining uncertainty under the supplied information, while the second measures distance from its conditional mean. We retain both realized error and conditional risk because their finite-sample changes can differ.

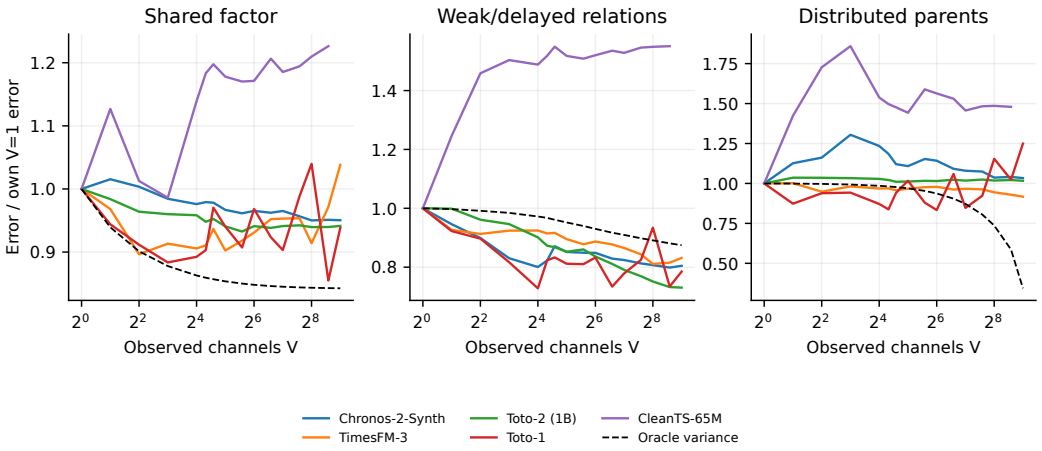


Figure 12: One-step NMSE versus informative channel count at $C = 512$, normalized by each model's own $V = 1$ error. The dashed curves show the corresponding conditional-variance ratios from the known-mechanism filter. Each panel uses 16 paired seeds. Missing configurations indicate unavailable runs, not interpolated performance.

G.3 Local Degradation and the Limits of a Channel Threshold

More informative channels do not guarantee lower model error. In the shared-observation system, TimesFM-3's realized NMSE increases by 6.87% from 384 to 512 channels at $C = 512$, with an adjusted interval of [2.66,24.47]%. Its conditional-risk increase is 7.31%, but the corresponding interval includes zero. In the distributed-parent system, Toto-1's realized NMSE rises by 22.04% over the same channel increment while conditional variance falls by 41.37%. Its conditional-risk

interval also includes zero. CleanTS provides a more consistent local example at $C = 128$: increasing shared channels from 8 to 16 raises realized NMSE by 15.17% and conditional risk by 13.84%, with both intervals above zero.

These are local comparisons, not a universal maximum usable channel count. The predefined persistent-boundary criterion requires two consecutive supported plateaus or deteriorations, at least 0.5% oracle improvement at both increments, and at least eight paired seeds. Intervals adjust for the 15 adjacent comparisons within each model, system, history length, and scoring span. No tested configuration establishes a persistent boundary under this criterion. CleanTS's unavailable large-channel runs, including $C = 512, V = 512$, are resource failures and do not supply performance evidence.

There are also counterexamples to a uniformly harmful channel-count effect. In the weak-delayed system at $C = 512$, Toto-2's realized NMSE decreases from 0.6197 to 0.4527 between the smallest and largest channel sets, although conditional risk increases from 0.6785 to 0.6983. This disagreement reinforces the need to distinguish an observed future realization from conditional expected performance. The experiments support model- and system-dependent use of additional information; they do not isolate an architectural cause.

Does aggregate forecasting performance reflect channel modeling capability? In Section 4.1, we found that channel capability scores are negatively correlated with GIFT-Eval point-forecasting performance ($\rho = -0.564$). To examine this difference, we compare model rankings under two measures of channel modeling: the improvement obtained from auxiliary channels and the absolute error of joint prediction. Toto-2 ranks last under our channel capability score, which combines cross-channel copying with incremental forecasting gains. Replacing the gain terms with absolute joint-prediction errors, while retaining the copying term, places TimesFM-3 first and leaves Toto-2 last. These results indicate that aggregate forecasting rankings do not fully describe how models use information from other channels.

H EXOGENOUS CONDITIONING

H.1 CONTROLLED FUTURE INPUTS

Exogenous conditioning concerns external inputs that are known over the forecast horizon. The model receives the observation history and, when its native interface supports them, the future inputs $u_{T+1:T+H}$. This differs from channel modeling, where auxiliary channels are observed only in the past. The evaluated system is restarted from the same state when comparing two future-input branches, so the observed history is identical and only the future intervention changes.

The original panel contains seven deterministic systems sampled at time interval 0.2, with no observation noise. We denote continuous states by y, reserving x for their sampled observations. The one- and two-compartment pharmacokinetic systems (Gibaldi & Perrier, 1982) are

$$\dot{y} = -0.25y + u, \tag{29}$$

$$\dot{y}_1 = -0.45y_1 + 0.12y_2 + u, \qquad \dot{y}_2 = 0.20y_1 - 0.12y_2. \tag{30}$$

The first compartment is observed. In the SEIR-based epidemic system (Hethcote, 2000), the contact input determines the transmission coefficient $\beta_t = u_t^{\text{contact}}$, while u_t^{vac} controls vaccination:

$$\begin{aligned}
\dot{y}^{\text{sus}} &= -\beta_t y^{\text{sus}} y^{\text{inf}} - u^{\text{vac}} y^{\text{sus}}, \\
\dot{y}^{\text{exp}} &= \beta_t y^{\text{sus}} y^{\text{inf}} - \gamma y^{\text{exp}}, \\
\dot{y}^{\text{inf}} &= \gamma y^{\text{exp}} - \upsilon y^{\text{inf}}, \\
\dot{y}^{\text{rec}} &= \upsilon y^{\text{inf}} + u^{\text{vac}} y^{\text{sus}}.
\end{aligned} \tag{31}$$

The progression and recovery rates retain the meanings used in Appendix C: $\gamma = 0.2$ and $\upsilon = 0.1$. The observed channel is the infected fraction.

The traffic system uses a cell-transmission formulation (Daganzo, 1994) with eight cells. Each cell has demand $\min(y_j, 1/3)$ and supply $\min(0.5(1 - y_j), 1/3)$. Interior flow is the minimum of upstream demand and downstream supply; the cell derivative is incoming minus outgoing flow. Boundary inputs determine inflow and outflow. The intelligent-driver system (Treiber et al., 2000) uses the leader's future speed as input. Its gap derivative is $u - y^{\text{speed}}$, and its desired gap is

$$\text{Gap}_{\text{desired}} = 2 + \max\left(0, 1.5y^{\text{speed}} + \frac{y^{\text{speed}}(y^{\text{speed}} - u)}{2\sqrt{1.2 \times 2}}\right). \tag{32}$$

The speed derivative is $1.2[1 - (y^{\text{speed}}/25)^4 - (\text{Gap}_{\text{desired}}/y^{\text{gap}})^2]$, with the numerical update enforcing nonnegative speed. The speed is observed.

The two spatial systems use 16 finite-volume cells and zero boundary values. Writing the spatial coordinate as s, the advection–diffusion–reaction and Darcy-type diffusion equations are

$$\partial_t y = -0.05\partial_s y + 0.005\partial_{ss} y - 0.05y + u, \tag{33}$$

$$\partial_t y = \partial_s(\text{Perm}(s)\partial_s y) + u. \tag{34}$$

The implemented spatial discretization uses upwind advection and diffusive face fluxes; the Darcy face permeability is a harmonic mean of positive cell permeabilities. All spatial cells are observed. These equations specify the generating mechanisms, not information made available to a forecast model.

H.2 PAIRED RESPONSES AND INEFFECTIVE-INPUT CONTROLS

The supplemental experiment uses $C = 512, H = 256$ and 16 seeds for PK-one, PK-two, Darcy, and SEIR. Eight-step input pulses have amplitudes 0.5 or 1.5 and begin at forecast steps

$$15, 16, 17, 31, 32, 33, 40, 63, 64, 65, 191, 192, 193.$$

This placement covers both short and long forecast distances, including neighborhoods of several candidate block boundaries. System parameters also vary across seeds: elimination rates range from 0.18 to 0.35, transfer rates from 0.12 to 0.28, recovery rates from 0.08 to 0.12, and permeability from 0.007 to 0.014.

For two branches with identical histories, let Δx be the true future response and $\Delta\widehat{x}$ the difference between their forecasts. For an error array $\mathbf{e}$, define the reference-normalized norm

$$\|\mathbf{e}\|_\nu^2 = \frac{1}{HV}\sum_{h=1}^{H}\sum_{j=1}^{V}\frac{e_{h,j}^2}{\nu_j^2}. \tag{35}$$

Each ν_j is determined from the common reference history. The paired response error is

$$\mathrm{RespErr} = \frac{\mathbb{E}\|\Delta\widehat{\mathbf{x}} - \Delta\mathbf{x}\|_\nu^2}{\mathbb{E}\|\Delta\mathbf{x}\|_\nu^2}. \tag{36}$$

Empirical expectations aggregate the accepted paired cases before taking the ratio. A prediction that ignores a nonzero intervention has response error one. A value near zero indicates accurate response differences, but need not imply accurate absolute trajectories.

A separate ineffective-input control changes an auxiliary input that does not drive the target system. Its score is the mean squared change in the forecast under that manipulation, on the same reference scale. This measures spurious sensitivity and has no division by true response energy. Hidden-input and deliberately incorrect-input branches are diagnostics; they are not included in the correct-input response ratio.

Table 21: Paired response error (Resp.) and ineffective-input sensitivity (Spur.) for all four eligible TSFMs and the linear baseline, $C = 512$, $H = 256$, 16 seeds. Lower values are preferable, but the two metrics have different denominators. Values rounded to zero need not be exact zeros.

	PK-one		PK-two		Darcy	
Model	Resp.	Spur.	Resp.	Spur.	Resp.	Spur.
Chronos-2-Synth	0.456	0.553	0.394	0.426	0.681	0.014
TimesFM-3	0.012	0.014	0.047	0.013	0.111	0.018
Toto-1	0.999	0.003	0.999	0.002	1.000	0.001
CleanTS-65M	0.983	0.065	0.983	0.051	0.996	0.000
Linear	0.000	0.000	0.048	0.001	0.000	0.000

H.3 Observed Response Patterns and Scope

TimesFM-3 follows the controlled input most closely among the evaluated TSFMs on the three systems in Table 21. Its response errors are 0.012, 0.047, and 0.111. Chronos-2-Synth shows partial responsiveness, whereas Toto-1 and CleanTS have response errors close to one. Their small ineffective-input scores therefore cannot, by themselves, establish selective conditioning: an input-insensitive predictor also produces a small spurious response. Correct response and resistance to ineffective inputs must be read together.

The linear baseline is highly competitive on these locally linear systems. Its performance establishes that the input-response relation can be recovered under the supplied information; it does not imply that the task requires nonlinear reasoning. Conversely, nonzero errors cannot identify a particular architectural failure without controlled changes to the model.

SEIR is not included in this normalized-response table because the measured true response energy is only 8.53×10^{-12} in the relevant supplemental configuration. Dividing by that quantity would make the ratio unstable and obscure the negligible intervention effect. This exclusion is a denominator-validity decision, not evidence of either success or failure on epidemic conditioning. The original exogenous profile and this targeted paired-response experiment therefore have different eligible system sets.

The pulse-onset grid permits inspection of where responsiveness changes across the forecast, but proximity to a candidate output-block boundary does not establish a block-induced failure mechanism. The principal supported conclusion is narrower: models with future-input interfaces differ substantially in whether their forecast differences track the imposed intervention.

I Real-Data Diagnostic Validation

I.1 Data Splits and Matched Forecast Windows

The synthetic experiments isolate a known dependency or intervention. The real-data validation asks whether related model sensitivities remain observable when the generating process is unknown. It does not treat a real-data error change as direct confirmation of a synthetic mechanism.

The original validation uses ETT1 at 15-minute resolution and Jena Weather at 10-minute resolution, with targets 0 and 1 and disjoint development and evaluation periods. A history-change score compares the means and standard deviations of the two most recent 32-step segments, scaled by the development-period standard deviation. The development 30th and 70th percentiles define stable and change strata. Forecast origins advance by $H = 96$, and up to 20 eligible windows per stratum are retained at equally spaced chronological positions. Future windows do not overlap, although their histories can overlap.

There are 134 of 160 planned target–window combinations in each original history/channel validation panel. ETT1 target-0 stable windows contribute 10 of 20, and Jena target-1 change windows contribute 4 of 20; the other groups contribute 20 each. Different targets can share an origin. Consequently, these combinations are not 134 independent events. Uncertainty summaries resample time blocks and remain descriptive when blocks share historical information.

I.2 Intervening on Old History at Fixed Input Length

For $C = 1024$, the old-history intervention shuffles the observations before the most recent 256 steps. It preserves the input length, recent observations, origin, and complete target future. The reported effect is the increase in NMSE from the original input to the shuffled input. Positive values mean that this alteration worsens the forecast; negative values mean that the altered history happens to improve it. Because the intervention changes temporal order rather than only removing information, neither sign alone identifies the model's internal retrieval mechanism.

Table 22: NMSE increase under same-length old-history shuffle, with descriptive 95% time-block intervals. Here $C = 1024$, $H = 96$; ETT1 target-0 change has 20 windows and Jena target-1 change has four. Recent observations and future targets are unchanged. Missing entries indicate unavailable forecasts under this panel's original interface.

Model	ETT1 target 0	Jena target 1
Chronos-2-Synth	0.32 [0.06, 0.75]	0.14 [-0.02, 0.41]
PatchTST-FM	0.51 [0.24, 0.77]	0.15 [0.06, 0.25]
Granite-PatchTST	0.47 [0.27, 0.68]	0.08 [-0.09, 0.26]
YingLong-110M	-0.05 [-0.22, 0.12]	0.89 [-0.03, 2.40]
YingLong-300M	0.13 [0.03, 0.27]	-1.40 [-2.80, -0.02]
TimesFM-3	0.36 [0.09, 0.76]	0.31 [-0.02, 0.88]
Toto-2	0.25 [0.07, 0.50]	0.01 [-0.03, 0.07]
TimesFM-2.5	–	–
PatchFM	0.20 [-0.13, 0.66]	0.22 [-0.11, 0.71]
Toto-1	0.19 [-0.04, 0.52]	0.26 [0.09, 0.42]
Sundial-128M	0.38 [0.15, 0.63]	0.29 [0.02, 0.70]
TiRex	0.53 [0.20, 0.96]	0.66 [0.14, 1.19]
Kairos-10M	-0.39 [-0.60, -0.21]	0.54 [-0.14, 1.44]
Kairos-23M	-0.56 [-0.78, -0.35]	1.38 [0.50, 2.33]
Kairos-50M	-1.27 [-1.62, -0.98]	0.10 [-0.42, 0.67]
FlowState-9.1M	0.40 [0.25, 0.64]	0.08 [-0.05, 0.20]
FlowState-r1.1	0.04 [-0.03, 0.11]	-0.04 [-0.41, 0.33]
Granite-FlowState	0.05 [-0.05, 0.17]	-0.14 [-0.37, 0.09]
Reverso-Small	0.64 [0.29, 1.01]	1.07 [0.86, 1.29]
Goia-nano-v0	0.61 [0.29, 0.90]	0.18 [-0.02, 0.51]
CleanTS-65M	-0.79 [-1.29, -0.39]	-0.05 [-0.56, 0.45]

The responses are heterogeneous. PatchTST-FM, Granite-PatchTST, and TiRex have positive ETT1 effects with intervals above zero, whereas the three Kairos checkpoints and CleanTS improve after the shuffle on that target. Jena's four change windows give much less coverage of possible conditions. The results establish sensitivity to old-history content in particular windows; they do not justify a dataset-wide ordering of retrieval ability. TimesFM-2.5's missing entries reflect the original evaluation interface. Its availability in a later supplemental experiment does not retrospectively fill these matched comparisons.

I.3 Joint Channels, Absolute Error, and Dataset Dependence

The original channel comparison uses $C = 512, V = 4, H = 96$ and the same five models with supported joint-input interfaces. Target and origin are held fixed, and channel ranking uses development data only. Gains follow Equation 27, with aggregation over 58 ETT1 time blocks and 57 Jena time blocks. The time-block intervals describe the available evaluation sample, rather than a universal benefit from multiple channels.

Table 23: Four-channel versus target-only gains (%) and descriptive 95% time-block intervals, $C = 512, H = 96$. Positive values indicate improvement.

Model	ETT1	Jena
Chronos-2-Synth	1.80 [−6.03, 10.72]	4.63 [−4.63, 13.57]
TimesFM-3	−3.41 [−18.66, 12.41]	5.34 [−10.71, 15.89]
Toto-2	−6.49 [−14.30, −0.32]	0.17 [−6.00, 5.09]
Toto-1	−3.25 [−9.90, 2.49]	4.95 [−4.43, 12.52]
CleanTS-65M	19.58 [7.72, 29.68]	−70.22 [−112.67, −41.33]

CleanTS improves on ETT1 and deteriorates substantially on Jena. Toto-2 has a negative ETT1 effect with an interval below zero, while most other intervals include zero. These differences support a model–data interaction rather than uniform multichannel improvement. A percentage gain must also be read beside absolute error: a large gain can occur when the target-only baseline is unusually poor.

I.4 Weather Follow-Up: Channel Content and Channel Position

The separate Weather follow-up uses targets 4 and 5. It contains ten stable windows for each target and two and zero change windows, respectively, giving 22 of 40 planned combinations. Stable windows occupy 17 time blocks. The one-step comparison below uses $C = 512$ and all 21 observed channels; it therefore differs from the four-channel, full-horizon comparison above.

Table 24: Weather stable-window one-step NMSE, $C = 512$, with target-only and all-21-channel inputs. The last column is percentage error change, $100(\overline{\mathrm{NMSE}}_{\mathrm{all}}/\overline{\mathrm{NMSE}}_{\mathrm{target}} - 1)$, so positive values indicate deterioration. Intervals are descriptive 95% time-block intervals.

Model	Target only	All channels	Error change (%)
Chronos-2-Synth	0.009537	0.009975	4.60 [−35.35, 33.50]
TimesFM-3	0.008193	0.009275	13.21 [−0.47, 17.80]
Toto-2	0.007219	0.008540	18.30 [13.33, 48.38]
Toto-1	0.008470	0.008804	3.94 [−25.23, 23.48]
CleanTS-65M	2.585433	0.989681	−61.72 [−76.15, −0.42]

Toto-2 deteriorates by 18.30% with an interval above zero. CleanTS improves by 61.72%, yet its joint-input NMSE remains far larger than those of the other four models. This example illustrates why the sign of a relative gain is insufficient for comparing absolute forecasting quality.

The channel-selection comparison ranks neighbors using development-period linear relationships. At $V = 21$, neighbor, far, and random selections necessarily contain the same complete channel set, so that endpoint cannot distinguish selection strategies. Across the tested subsets, development-ranked neighbors do not establish a stable advantage.

Channel-set selection is also distinct from channel ordering. Pure permutations give changes below 0.001% for four of the tested models in the corresponding control, and moving the strongest selected neighbor among positions 1, 2, 4, 8, and 16 gives negligible changes for all five. Toto-1's target-position effects must be compared with repeat inference on identical inputs; the relevant comparison intervals include zero. Repeated inference is a stochastic-variation reference, not additional independent data. Content shuffles or time shifts can have larger effects, but some improve forecasts when the original auxiliary information is harmful. These controls do not support a general claim that an early channel position is privileged.

J Model Design Associations and Sensitivity

J.1 Group Construction and Fixed Scores

This analysis relates observed capability profiles to documented model properties. It uses the common induction, retrieval, tracking, and uncertainty panel of 20 eligible models. Channel and exogenous-conditioning scores are excluded from these group means because their supported model sets are much smaller. The rank scores are those computed from the full eligible pools in Appendix A.5; they are not recomputed after selecting a group or a high-scoring subset.

To reduce the weight of families with several checkpoint versions, we first average the versions within each family and then average families within a group. The exploratory core consists of models with a score of at least 60 on at least three of the four common axes. This selection mechanically favors higher scores. A rise from a full-group mean to a core mean is therefore not independent evidence for the corresponding design choice.

Table 25: Family-balanced capability means under fixed full-pool rank scores. Counts give models/families. The exploratory core requires a score of at least 60 on three of four axes. Values are rank-score points, not percentage error reductions.

Group	Models/families	Induction	Retrieval	Tracking	Uncertainty
Encoder full	5/3	43.3	69.6	60.0	58.8
Encoder core	3/2	56.2	86.9	76.2	82.2
Decoder full	6/4	73.1	67.2	76.2	49.3
Decoder core	4/3	87.1	81.7	88.3	71.1
CPM full	8/5	80.2	72.9	76.0	75.0
CPM core	5/4	85.0	82.8	84.4	84.5
NTP full	3/3	68.3	70.8	76.7	52.6
NTP core	2/2	78.8	77.5	80.0	52.6

J.2 Architecture Associations and Exceptions

The encoder and decoder cores retain three of five and four of six models, respectively. The decoder-core averages exceed encoder-core averages by 30.8 induction points and 12.1 tracking points. The encoder-core averages exceed decoder-core averages by 5.2 retrieval points and 11.2 uncertainty points. These are associations within a selected sample, not evidence that one architecture can perform only a subset of capabilities.

The separate within-model contrast asks a different question. Four of five encoders, covering all three encoder families, have retrieval at least ten points above induction. Four of six decoders, covering all four decoder families, have tracking at least five points above uncertainty. Because these rules use different axes and thresholds, their success proportions are not estimates of the same outcome.

Exceptions clarify the interpretation. YingLong-300M satisfies the encoder contrast while remaining weak on several other axes; relative emphasis can coexist with low absolute capability scores. TimesFM-3 and Toto-2 do not satisfy the decoder contrast because their uncertainty scores are also high. An improvement on uncertainty can reduce a tracking-minus-uncertainty contrast without reducing tracking ability. These cases are retained in the complete groups rather than excluded to sharpen a narrative.

J.3 Training Task, Forecast Head, and Metadata Coverage

Contiguous patch masking (CPM) is a training task spanning multiple backbones. Its full group contains eight models from five families, and its core retains five models from four families. The next-token or next-patch prediction (NTP) group contains TimesFM-2.5, PatchFM, and Toto-1 before selection; two remain in the core. FlowState is retained in the CPM category. The observed CPM/NTP differences combine family, data, architecture, and output-interface differences and do not isolate a training-task effect.

Loss and output descriptions also require separate treatment. Pinball and weighted pinball losses, Student-t likelihood, flow matching, and direct point losses are distinct metadata categories. A model's runtime probabilistic output does not by itself establish the loss used throughout pretraining. CleanTS's executed causal backbone is identifiable, while its pretraining objective is unverified in the collected evidence. FlowState's state-space functional representation is distinct from Sundial's flow-matching forecast construction. These differences preclude treating every distributional output or every occurrence of "flow" as one training paradigm.

Appendix F therefore tests uncertainty behavior directly through conditional coverage, interval-width growth, and distributional scores. The resulting patterns can motivate hypotheses about training objectives, but they cannot establish those hypotheses without controlled training comparisons.

J.4 Raw-Probe Clustering and Training-Task Labels

To complement rank profiles, the primary clustering analysis operates on the original probe values within each capability: three NRMSE probes for induction, four first-step relative-MSE probes for retrieval, five transition NMSE probes for tracking, and five scaled-pinball probes for uncertainty. Coverage is 21, 21, 21, and 20 models, respectively. Ward clustering uses Euclidean distances in the original linear scale. The number of clusters is selected from two through six by the silhouette criterion, without using design labels. Raw scaling retains error magnitude but gives more influence to probes with greater dispersion. Probe standardization, alternative linkage rules, and removal of individual probes or families are sensitivity analyses.

The full-pool cluster counts are four, three, three, and two in capability order. On the strict 11-model CPM/NTP subset, independently refitting the clusters selects two clusters for every capability, but those clusters do not recover the two training-task labels. CPM and NTP models share low-error clusters, while the three FlowState versions differ from other CPM models on several point-prediction probes. The label-defined silhouette values are negative on all four capabilities. Thus, the presence of two fitted clusters is not evidence of two training mechanisms.

Table 26: Raw-probe clustering and agreement with CPM/NTP labels. The full-pool analysis uses all eligible models, then evaluates agreement on the 11 labeled models. Strict-subset clustering is independently refitted on those 11 models. ARI denotes adjusted Rand index; it is not accuracy or a significance test.

Capability	Full clusters	Label silhouette	Full ARI	Strict ARI
Structure Induction	4	−0.111	−0.109	−0.109
History Retrieval	3	−0.069	−0.050	−0.109
State Tracking	3	−0.137	−0.109	−0.109
Uncertainty Modeling	2	−0.031	0.000	−0.088

Raw losses also limit the interpretation of rank gaps. With family weighting, Lorenz NRMSE is 0.18312 for CPM and 0.18454 for NTP. Five-system mean scaled pinball loss is 0.25440 versus 0.25781, a difference of about 1.32%, rather than the much larger gap suggested by treating rank-score points as percentage loss reduction. At dependency distance $d = 33$, CPM's relative MSE is 39.8% lower and the direction survives the recorded family-removal comparisons. At $d = 129$, the 24.3% mean advantage can reverse after removing a family. These are distance-specific associations, not universal paradigm rankings.

J.5 Input Representation and Capability Clusters

The verified input-representation groups contain 13 fixed-patch models from nine families, four adaptive multiscale models from two families, and four pointwise-input models from two families. Input and output granularity are distinct: TiRex uses input patches of length 32, while FlowState's decoder block does not determine its input tokenization. Chronos-2-Synth uses continuous patches of length 16. The two YingLong checkpoints use length-32 patches with a Haar transform; this is not a discrete numerical vocabulary. No suitable discrete-versus-continuous vocabulary comparison is present in this panel.

The existing raw-probe clusters are held fixed and representation labels are attached afterward. Their adjusted Rand indices are 0.594 for induction, 0.102 for retrieval, 0.729 for tracking, and -0.096 for uncertainty. In tracking, twelve fixed-patch models share a cluster, the four adaptive and four pointwise models share another, and CleanTS forms a separate cluster. In uncertainty, the representation groups mix. The association therefore depends on the capability and does not extend uniformly to probability modeling.

Within the fixed-patch group, lengths 16, 32, and 64 have four, eight, and one models. These small, family-dependent groups do not establish a monotonic patch-length effect. Adaptive models' maximum configured lengths are not substitutes for their unknown per-input routing choices. Seasonal AR is also a counterexample to a uniform fixed-patch advantage: family-weighted NRMSE is 0.18273 for pointwise input and 0.27680 for fixed patches. Representation labels correlate with backbones and training recipes, so the cluster agreement cannot be interpreted as their independent causal contribution. Its magnitude is also not directly comparable with the CPM/NTP agreement measured on a different, 11-model subset.

J.6 Relative Capabilities Within a Design Category

A complementary summary compares the four capability scores within the same design category. Each row uses the same eligible members on every axis and checkpoint weighting; it asks which capabilities are relatively emphasized by that category, rather than which model wins within it. The full-pool ranks remain unchanged. Categories overlap, so repeated members under different labels do not provide independent evidence.

Table 27: Within-category rank profiles with checkpoint weighting and common four-axis members. Counts are common members/all category members. The family-balanced encoder, decoder, CPM, and NTP summaries above use different weights and answer a complementary question.

Category	Members	Induction	Retrieval	Tracking	Uncertainty
Fixed input patches	13/13	58.1	64.4	63.8	53.2
Adaptive multiscale patches	4/4	16.2	13.8	16.9	49.3
Pointwise input	3/4	60.8	33.3	42.5	36.8
Encoder-only	5/5	40.5	64.5	54.0	53.7
Decoder-only	6/6	75.4	71.2	80.0	60.1
Quantile supervision (broad)	15/15	48.0	48.3	48.0	51.1
Basic multi-quantile pinball	9/9	68.9	66.7	66.4	60.5
Horizon-weighted pinball	3/3	11.7	8.3	11.7	42.1
NTP	3/3	68.3	70.8	76.7	52.6
CPM	8/8	72.2	63.8	65.9	66.1
Direct multi-step	5/6	24.5	30.0	31.5	56.3

Adaptive-patch models and direct multi-step models emphasize uncertainty relative to their own other axes, whereas the pointwise group emphasizes induction. The probabilistic pointwise subset contains only the three FlowState checkpoints because Reverso lacks the required output. Likewise, the horizon-weighted pinball row is the same three Kairos checkpoints as the encoder–decoder group. These labels cannot be treated as repeated confirmation of separate effects. Broad quantile supervision varies by only 3.1 score points across the four axes, and its small relative ordering changes with family weighting. A relative advantage need not be large or imply strong absolute forecasts.

J.7 Checkpoint Scale and Capability-Specific Changes

Within the original panel, Kairos-10M, -23M, and -50M have induction scores 10, 0, and 25. All three remain at or below 25 on induction, retrieval, and tracking. YingLong-110M to -300M increases retrieval from 20 to 50, while induction changes from 20 to 15 and tracking from 40 to 15. The evaluated checkpoint changes are thus not uniformly beneficial across axes.

The additional Toto-2 scale comparison in Appendix C is a more focused comparison within one family. The 313M, 1B, and 2.5B checkpoints are evaluated on the same accepted induction

cases, but their errors are still not monotonically ordered in every case. The arithmetic, geometric, transformed-growth, logarithmic, and sinusoidal tasks expose different changes. Neither these results nor the original cross-generation TimesFM/Toto contrasts establish a general scaling law: parameter count can change alongside training data, optimization, and interface behavior.

J.8 Sensitivity and Evidential Limits

The group conclusions are conditional on the evaluated checkpoints, verified metadata, eligible interfaces, and the scoring protocol. Family balancing reduces duplicate-family influence but does not create independent architecture interventions. Threshold selection changes core membership, and rank-based axes express relative standing within their reference pools. The complete groups, fixed-score convention, and explicit exceptions make those dependencies visible.

The accepted supplementary analyses add evidence on simple-rule extrapolation, periodic decomposition, lag repetition, forecast-horizon uncertainty, informative channel count, and paired exogenous responses. They are reported in their corresponding capability appendices rather than silently replacing the original frozen profiles. Unaccepted runs and experiments awaiting result validation do not enter the reported tables. The resulting design implications are hypotheses supported by behavioral associations, with their scope limited to the systems and comparisons actually evaluated.

J.9 Which Model Designs Are Associated with Capability Differences?

Controlled studies establish that design choices can induce different forecasting biases (Yu et al., 2026). Our cross-model comparisons provide complementary behavioral associations, rather than controlled interventions on an individual architectural component.

In the preceding sections, we found that models with similar aggregate forecasting performance can exhibit different capability profiles. To understand these differences, we examine how model designs organize historical observations, produce future predictions, and incorporate information across channels.

Our structure induction experiments make this distinction visible at output-block boundaries. Models that predict multiple values per block can exhibit abrupt changes in error when moving between blocks, including the patterns observed in the Kairos family. Producing a block from a shared representation does not explicitly enforce continuity with the next block, which is consistent with the observed boundary effects. These results connect the organization of prediction outputs to how models continue otherwise simple temporal structures.

Where information is combined. The comparison between Toto-2 and TimesFM-3 further shows how models with attention-based architectures can differ in their use of auxiliary information. Toto-2 applies variate attention at the final layer in the evaluated configuration, whereas TimesFM-3 alternates temporal and variate attention across multiple layers (Khwaja et al., 2026; Das et al., 2024).[6] The latter arrangement allows information from other channels to enter subsequent temporal processing. The channel capability differences reported in Section 4.5 are consistent with this distinction: providing access to additional channels and integrating their information throughout temporal processing are separate properties of a model.

How predictive distributions are produced. Our uncertainty experiments reveal a related distinction between an output mechanism and the behavior it produces. Sundial generates stochastic predictions, yet its predictive intervals can remain narrower than the target distribution requires. Randomness in generation therefore does not by itself determine whether predictive uncertainty is represented accurately. The learned transformation of that randomness determines the resulting distribution, just as a quantile prediction head must learn the appropriate conditional quantiles.

[6] Configuration and architecture details: https://huggingface.co/Datadog/Toto-2.0-1B/blob/main/config.json and https://research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/.

Table 28: Overview of the six forecasting capabilities and their diagnostic probes.

Capability	Diagnostic question	Probe dimensions	Representative probes
Structure Induction	Can the model infer and continue predictive rules from observations?	Non-chaotic and chaotic dynamics	Seasonal AR, SEIR, Lorenz–63, elementary function continuation, and component mixtures
History Retrieval	Can the model find and use relevant historical information?	Information distance and information density	Lag-copy, Seasonal AR with context-length variation and history interventions
State Tracking	Can the model update its forecasts as the active dynamics change?	State transitions, transition speed, and state recurrence	Switching AR, smooth AB/ABA transitions, old-state information interventions
Uncertainty Modeling	Can the model capture uncertainty across states and forecast horizons?	Conditional uncertainty structures	OU, Gaussian random walks, GARCH, self-exciting counts, queues, HMMs
Channel Modeling	Can the model exploit predictive information from other channels?	Channel count and dependency structure	Cross-copy, VAR, latent-factor groups, graph diffusion
Exogenous Conditioning	Does the model adjust its forecasts appropriately to future external inputs?	Periodic and non-periodic inputs, and condition validity	Pharmacokinetic and spatial systems with controlled future inputs

K Detailed Probe Definitions and Experimental Settings

This appendix contains the detailed probe constructions and experimental settings relocated from Section 3. Capability-specific results and additional analyses remain in Appendices C–H.

K.1 Structure Induction

K.1.1 Non-chaotic Dynamics

We use Seasonal AR and SEIR to test the induction of seasonal dependencies and nonlinear dynamics, respectively.

`Seasonal AR`. This probe requires the model to identify a recurring lag relationship and infer how values evolve across cycles. We construct sequences as

$$x_t = \phi x_{t-24} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, 0.08^2). \tag{37}$$

We use $\phi = 0.9$ as the default and evaluate parameter variations $\phi \in \{0.85, 0.98\}$. We also change the initial seasonal values to test whether predictions remain accurate when the starting pattern changes. The model receives only the observed sequence; the seasonal lag and autoregressive coefficient are not provided.

`SEIR`. (Hethcote, 2000) This probe tests prediction of nonlinear trajectories governed by interactions among susceptible (S), exposed (E), infectious (I), and recovered (R) populations:

$$\begin{aligned} \dot{S} &= -\beta SI, \qquad \dot{E} = \beta SI - \alpha E, \\ \dot{I} &= \alpha E - \gamma I, \quad \dot{R} = \gamma I. \end{aligned} \tag{38}$$

In the main experiment, the progression and recovery rates are $\alpha = 0.2$ and $\gamma = 0.1$. We vary the transmission rate $\beta \in \{0.45, 0.50, 0.80\}$ and the initial exposed and infectious proportions. Each of the four variables is forecast separately from its own history, without access to the other variables. An additional experiment includes active epidemic phases to assess prediction during growth, turning points, and decline. This experiment uses $\alpha = 1/5.2$, $\gamma = 0.1$, and $\beta \in \{0.35, 0.50, 0.70\}$.

Elementary Functions and Composition. We use simple functions and their combinations to examine whether models continue identifiable temporal relationships and preserve them when components are mixed.

`Elementary Function Continuation.` We construct arithmetic, geometric, exponential, and logarithmic sequences, varying their slopes, growth rates, or time scales. The model receives only the observed values, without the function type or parameters. In addition to prediction error, we inspect changes in forecast residuals at native output-block boundaries. Smooth target sequences allow us to examine whether blockwise predictions introduce discontinuities into otherwise regular continuations.

`Component Composition.` We construct sequences as sums of temporal components, $x = f_1 + f_2$. We compare direct forecasts of the summed history with the sum of forecasts obtained separately from the true component histories. Both branches predict the same future sum under a common evaluation scale. The separate-component branch receives the true decomposition and serves as a privileged diagnostic reference.

We additionally evaluate a history-based periodic decomposition motivated by CycleNet (Lin et al., 2024). Given the period, we estimate a periodic template from historical phase averages, subtract it from the observed sequence, forecast the residual with the frozen TSFM, and add the template back to the prediction. We vary the periodic component and residual type to examine when this separation improves forecasting. Function definitions, parameter settings, and decomposition procedures are provided in Appendix C.

K.1.2 Chaotic Dynamics

We use `Lorenz-63` (Lorenz, 2017) to test forecasting under dynamics that can amplify small errors in the estimated state or inferred evolution rule:

$$\begin{aligned} \dot{x} &= 10(y - x), \\ \dot{y} &= x(\rho - z) - y, \\ \dot{z} &= xy - \frac{8}{3}z. \end{aligned} \tag{39}$$

We vary $\rho \in \{26, 28, 40\}$ and the initial state, and sample trajectories at intervals of 0.02. The three coordinates are forecast in separate calls, each using only its own history. This tests whether a model can predict a scalar trajectory from coupled nonlinear dynamics and retain this ability when the initial conditions or parameters change.

Experimental Settings. The main experiments use context lengths $C \in \{128, 256, 512\}$, forecast horizons $H \in \{8, 16\}$, and five seeds. Additional Lorenz and active-phase SEIR experiments extend the horizons to $H \in \{96, 256\}$ with 16 independent seeds. In these additional experiments, context lengths are varied at fixed forecast origins so that the future targets remain unchanged. We compare forecasts with persistence and local linear baselines to assess whether predictive accuracy exceeds simple continuation.

K.2 History Retrieval

Historical reuse is a relevant forecasting strategy (Zhang & Gilpin, 2026). The probes below specify our controlled constructions for testing dependency distance and the availability of evidence.

Information Distance: Construction.

$$x_{t+1} = x_{t-L} + \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim \mathcal{N}(0, \sigma^2), \qquad d = L + 1. \tag{40}$$

Here, L is the lag parameter in the recurrence, while d is the actual distance between the source and target observations. We vary L, the context length C, and the noise level σ. Only configurations satisfying $C - d \geq 256$ are retained, ensuring that at least 256 source–successor pairs are present in the history.

We evaluate the first forecast step to reduce the influence of multi-step error accumulation. The model is not given the true lag. We compare its predictions with a baseline that searches for the lag

using historical observations, and with a reference that uses the known lag to predict the conditional mean. These comparisons assess whether the predictive relationship can be identified from the available history and effectively used by the model.

Information Density: Construction. For a recurring dependency with period L in a context of length C, the number of visible cycles provides a repetition-based proxy:

$$N_{\text{rep}} \approx \frac{C}{L}. \tag{41}$$

At a fixed period, a longer history contains more repetitions from which the relationship can be identified. At a fixed context length, a longer period provides fewer repetitions. Noise, missing observations, and distractors can further change how much of this history constitutes useful evidence.

We use `Seasonal AR` to examine these effects through different context lengths and controlled history interventions:

$$x_t = \phi x_{t-L} + \epsilon_t, \qquad \phi = 0.95, \qquad \epsilon_t \sim \mathcal{N}(0, 0.08^2). \tag{42}$$

For a history indexed from 0 to $C-1$ and a forecast horizon $H < L$, the source observations needed for the future conditional means are located at

$$\mathcal{P} = \{C - L, \ldots, C - L + H - 1\}. \tag{43}$$

We compare the complete history with variants that mask these source observations, mask an equal number of other observations, retain only recent history, or prepend distractor observations. We also include a cropped input that retains the source observations and masks the remaining positions within the retained time span. Masking preserves the original temporal positions.

These interventions vary the amount and relevance of historical evidence. Observations outside $\mathcal{P}$ may provide repeated examples of the same dependency and help identify its lag or coefficient. Retaining the source values alone therefore preserves the values needed for prediction but may remove the evidence needed to recognize how they should be used. Conversely, adding distractors increases input length without necessarily adding evidence for the target relationship. We use C/L to describe repetition in the unmodified history and the interventions to examine the role of effective evidence. All input variants are evaluated against the same future target and use the same reference scale.

Experimental Settings. Lag-copy uses context lengths $C \in \{512, 2048, 8449\}$, lag parameters ranging from 8 to 8192, noise levels $\sigma \in \{0, 0.05, 0.1\}$, and 20 independent seeds. Seasonal AR uses $L \in \{24, 48, 96\}$, $C \in \{128, 256, 512\}$, $H \in \{8, 16\}$, and five seeds. We report prediction error across dependency distances and paired error changes under the history interventions.

K.3 State Tracking

State Transitions. Regime-switching time-series models provide the background (Hamilton, 1989); our transition schedules and diagnostic controls are specified below. We first use `Switching AR`, in which the active state determines the mean and autoregressive coefficient:

$$x_t = \mu_{s_t} + \phi_{s_t}(x_{t-1} - \mu_{s_t}) + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, 0.15^2). \tag{44}$$

The two states have parameters $(\mu_A, \phi_A) = (-2, 0.65)$ and $(\mu_B, \phi_B) = (2, -0.55)$. Changing the active state therefore changes both the level and the temporal dependence of the series.

We additionally construct smooth transitions between five pairs of dynamics: decline to growth, growth to a plateau, a plateau to growth, low to high frequency, and low to high amplitude. For a transition centered at time c, we define

$$w_\tau(r) = \frac{1}{1 + \exp(-r/\tau)}, \qquad r = t - c, \tag{45}$$

where τ controls the transition width. Frequency and amplitude transitions use w_τ to blend the two oscillations. Trend transitions use its integral,

$$S_\tau(r) = \tau \left[\log(1 + \exp(r/\tau)) - \log 2\right], \tag{46}$$

to change the slope smoothly. We evaluate forecasts at multiple origins before, during, and after each transition.

Transition Speed. Switching AR includes abrupt and gradual transitions. For the smooth function probes, we vary τ to control how quickly the new dynamics become dominant. Moving the forecast origin changes how much evidence of the new state the model has observed. We also vary the context length and observation noise to examine how older observations and noisy recent evidence affect adaptation.

State Recurrence. We compare single transitions, $A \to B$, with return sequences, $A \to B \to A$. For the smooth return probes, the weight assigned to state B is

$$w_{\mathrm{ABA}}(r) = w_\tau(r) - w_\tau(r - D), \tag{47}$$

where D separates the two transition centers. The two transitions are evaluated separately, with a sufficiently long interval for state B to become dominant.

For frequency-return probes, we further replace the old A segment with an oscillation of the wrong frequency, a shuffled version of the segment, or an unrelated sequence. These interventions preserve the input length, temporal positions, recent history, and future target. Comparing them with the unchanged history tests whether the content of the earlier state helps the model predict its return.

Controls and Evaluation. We compare forecasts from complete and recent-only histories. Separate pure-state runs assess whether the model can predict each constituent dynamic without a transition. For the smooth function probes, we distinguish directional recovery from accurate recovery. Directional recovery requires the forecast to be closer to the destination-state reference for two consecutive evaluation origins. Accurate recovery additionally requires a clean NMSE of at most 0.05 at both origins.

Recovery delay is measured from the point at which the destination-state weight reaches 0.8. We report recovery rates alongside delays and treat trajectories that do not recover within the evaluation window as right-censored. This separates choosing the correct direction from producing an accurate forecast.

Experimental Settings. Switching AR uses $C \in \{128, 256, 512\}$, $H \in \{8, 16\}$, and five seeds. The initial smooth AB experiments use $C = 256$, a recent-only context of 64 observations, $H = 96$, $\tau \in \{8, 24\}$, and ten seeds. The corresponding ABA experiments use $D = 256$, $\tau = 16$, and five seeds. For these smooth probes, recovery is primarily evaluated over the first 16 forecast steps. Additional experiments use 16 independent seeds and extend the context-length comparison to $C \in \{64, 256, 1024\}$ for the frequency-transition and return probes.

K.4 Uncertainty Modeling

Horizon-dependent Uncertainty. For the underlying Brownian-motion theory, see Wang & Uhlenbeck (1945). We use an `Ornstein-Uhlenbeck (OU)` process with zero long-run mean. Its conditional distribution is

$$x_{T+h} \mid \mathcal{H}_T \sim \mathcal{N}\left(a^h x_T,\ \frac{\sigma^2}{2\kappa}(1 - a^{2h})\right), \qquad a = \exp(-\kappa \Delta t), \tag{48}$$

where κ is the mean-reversion rate and σ controls the noise level. The conditional variance grows with the forecast horizon and approaches $\sigma^2/(2\kappa)$. This probe tests whether predicted uncertainty captures both its magnitude and its evolution over time.

`Gaussian Random Walk`. We evaluate a Gaussian random-walk probe:

$$x_t = x_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \qquad x_{T+h} \mid \mathcal{H}_T \sim \mathcal{N}(x_T, h\sigma^2). \tag{49}$$

This provides a contrast to OU: the conditional variance continues to grow linearly with the horizon, while the conditional mean remains at the last observation. The corresponding interval width should grow in proportion to $\sigma\sqrt{h}$.

The completed experiment uses $C \in \{128, 256, 512\}$, $H \in \{96, 256\}$, $\sigma \in \{0.1, 0.4, 0.8\}$, and ten independent history seeds. A fixed forecast origin at $T = 512$ makes shorter contexts suffixes of longer contexts with the same future; horizon comparisons share future prefixes, and noise levels use paired standardized innovations. The 20 models with native probabilistic outputs each have

180 valid cases, giving 3600 saved forecasts. Models receive neither the true future nor the noise parameter or oracle quantiles.

Evaluation uses the known Gaussian conditional distribution to compute expected nine-quantile loss, analytic interval coverage, interval-width ratios, and width-growth exponents. The known-noise oracle, history-estimated Gaussian reference, and fixed-width negative control are evaluated separately. Q9 is a nine-quantile approximation to CRPS (Gneiting & Raftery, 2007), not exact CRPS. Complete fixed-panel results and horizon diagnostics are reported in Appendix F.3, including Table 19 and Figure 11.

State-dependent Volatility. We use `GARCH(1,1)` (Bollerslev, 1986) to construct sequences whose conditional variance depends on recent observations:

$$x_t = \sqrt{v_t}\, z_t, \qquad v_t = \omega + \alpha x_{t-1}^2 + \beta v_{t-1}. \tag{50}$$

The innovations z_t follow either a standard Gaussian distribution or a Student-t distribution with five degrees of freedom, standardized to unit variance. The main configurations use $\alpha = 0.12$, $\beta = 0.8$, and $\omega = 1 - \alpha - \beta$, giving a long-run variance of one. This probe examines whether forecasts adjust their intervals to recent volatility and distinguish Gaussian from heavy-tailed outcomes.

Intensity-dependent Discrete Uncertainty. Motivated by the principle of self-excitation (Hawkes, 1971), we use a finite-memory, discrete-time `self-exciting count process`, rather than an exact continuous-time Hawkes process:

$$x_t \mid \mathcal{H}_{t-1} \sim \text{Poisson}(\lambda_t), \qquad \lambda_t = \lambda_0 + \sum_{j=1}^{16} w_j x_{t-j}. \tag{51}$$

We set $\lambda_0 = 0.5$ and use weights proportional to $\exp[-0.25(j-1)]$, normalized to sum to either 0.6 or 0.9. Recent events increase subsequent event intensity. The probe tests whether predictive distributions reflect this changing intensity while respecting the nonnegative, discrete support.

Load-dependent Uncertainty. We use an `M/M/c queue` (Shortle et al., 2018), whose customer count follows a continuous-time process with transition rates

$$q_{n,n+1} = \lambda, \qquad q_{n,n-1} = \min(n, c)\mu, \tag{52}$$

where λ is the arrival rate, μ is the per-server service rate, and c is the number of servers. Departures occur only when $n > 0$. We vary the load $\lambda/(c\mu)$ across $\{0.8, 0.9, 0.95, 1.1\}$ and sample the customer count at regular intervals. These configurations test whether forecasts capture uncertainty associated with the current congestion level, including finite-horizon behavior under overload.

Latent-state Mixture Uncertainty. We use a `two-state Gaussian HMM` (Rabiner, 1989) with state means $\mu_1 = -1$ and $\mu_2 = 1$, switching probability 0.08, and observation standard deviation $\sigma \in \{0.8, 1.2\}$. Its marginal predictive distribution is

$$p(x_{T+h} \mid \mathcal{H}_T) = \sum_{j=1}^{2} (\boldsymbol{\pi}_T \mathbf{P}^h)_j \mathcal{N}(x_{T+h}; \mu_j, \sigma^2), \tag{53}$$

where $\mathbf{P}$ is the transition matrix and $\boldsymbol{\pi}_T$ is the filtered posterior over the current hidden state. Reference futures are sampled from this posterior, rather than initialized using the realized hidden state. The probe therefore includes uncertainty about both the current state and subsequent transitions.

Probe Instantiation and Evaluation. For each history in the original five stochastic probes, we generate 100 reference futures using independent future randomness. Models receive the observed sequence and produce native quantiles or forecast samples. These outputs are evaluated at individual forecast steps using probabilistic losses, interval coverage, and interval width.

We examine these measures across forecast horizons and stratify histories by quantities such as recent volatility and current queue length. This assesses whether aggregate performance conceals errors under particular states. Coverage and width are interpreted together, since narrow intervals are useful only when they provide appropriate coverage. All shared comparisons concern marginal distributions rather than the joint distribution of complete future trajectories.

Experimental Settings. The original five-probe experiments use $C \in \{128, 512\}$, $H \in \{96, 256\}$, and three independent history seeds. Supplementary experiments on OU, GARCH, and queues use 16 independent histories, a fixed forecast origin, $C \in \{128, 512\}$, and $H = 96$. Context-length comparisons in these supplementary experiments share the same reference futures and evaluation scale. Statistical aggregation treats histories as the independent units; multiple reference futures improve distributional evaluation within each history.

K.5 Channel Modeling

Number of Observed Channels. We generate a fixed pool of channels and progressively reveal nested subsets while keeping the target history, forecast origin, and future target unchanged. Thus, increasing the observed channel count changes the information available to the model while preserving the underlying dynamics.

The initial experiments use a pool of 64 channels, with observed counts $D \in \{1, 20, 24, 32, 64\}$, where $D = 1$ denotes target-only input. Additional experiments extend the pool to 512 channels and examine shared-state measurements, weak delayed measurements, and predictive information distributed across multiple parent channels. Together, these probes distinguish the effects of adding predictive information, redundant noisy observations, and unrelated distractors.

Dependency Structure. We examine four types of cross-channel relationships.

`Cross-copy`. The target depends on a previous observation from another channel:

$$x_{0,t+1} = x_{1,t} + 0.05\eta_{t+1}, \qquad \eta_{t+1} \sim \mathcal{N}(0, 1). \tag{54}$$

The source channel and distractor channels are independent standard Gaussian sequences. The model must identify the source channel from the observed histories. We evaluate the first forecast step because subsequent target values depend on source observations that are not yet available.

`Fixed-evidence VAR`. (Lütkepohl et al., 2005) The target depends on its own history and two parent channels:

$$x_{0,t} = 0.45x_{0,t-1} + 0.25x_{1,t-1} + 0.20x_{2,t-1} + \epsilon_{0,t}. \tag{55}$$

The remaining channels follow independent AR(1) processes with coefficient 0.65. Both parent channels are included in every joint-input condition, so further increases in channel count add distractors while holding the predictive sources fixed.

`Latent-factor Groups`. Channels provide noisy observations of shared latent dynamics:

$$x_{j,t} = a_j f_{g(j),t} + 0.1 f_{(g(j)+1) \bmod 4,t} + \epsilon_{j,t}, \qquad g(j) = j \bmod 4. \tag{56}$$

The four factors are AR(2) oscillators with periods $\{24, 48, 64, 96\}$ and characteristic-root radius 0.94. Loadings are sampled as $a_j \sim \mathcal{U}(0.8, 1.2)$. Additional channels provide correlated observations with independent measurement noise, allowing us to test whether models benefit from multiple measurements of shared dynamics.

`Graph Diffusion`. Channels interact through local connections on a fixed path graph:

$$\mathbf{x}_t = \left(0.55\mathbf{I} + 0.30\mathbf{D}_G^{-1}\mathbf{A}\right)\mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{57}$$

where $\mathbf{A}$ is the adjacency matrix and $\mathbf{D}_G$ is its degree matrix. Increasing the observed channel count reveals more distant nodes of the same graph. The graph and transition dynamics remain fixed across observation subsets.

Controls and Evaluation. We compare target-only and joint-channel forecasts on the same future target. Supplementary experiments disrupt auxiliary information by shuffling observations within each auxiliary history or shifting those histories in time, while preserving the target history and the number of auxiliary channels.

We report target-only and joint-channel prediction errors over the first eight forecast steps, together with their paired difference:

$$\Delta\mathcal{L}_{1:8} = \mathcal{L}_{\text{target},1:8} - \mathcal{L}_{\text{joint},1:8}. \tag{58}$$

Positive values indicate a benefit from observing additional channels. We additionally report full-horizon errors to examine whether this benefit persists over longer forecast horizons.

Experimental Settings. The initial experiments use $C \in \{128, 512\}$, $H \in \{96, 256\}$, and three seeds. The VAR, latent-factor, and graph-diffusion probes use Gaussian innovation and observation noise, as applicable, with standard deviation 0.2. Supplementary dependency and large-scale channel-count experiments use 16 independent seeds and $H = 96$. We evaluate both short forecast windows and the full horizon to examine how channel benefits change with prediction distance.

K.6 Exogenous Conditioning

Periodic Inputs. We provide nonnegative sinusoidal inputs spanning either two or eight cycles over the forecast horizon:

$$u_h = \frac{1}{2}\left[1 + \sin\left(\frac{2\pi k h}{H}\right)\right], \qquad k \in \{2, 8\}, \quad h = 0, \ldots, H-1. \tag{59}$$

For PK probes, u_h specifies the input rate; for spatial probes, it specifies the forcing at a selected source location. The model must account for how the dynamics transform this input. For example, the one-compartment PK probe (Gibaldi & Perrier, 1982) follows

$$\frac{dx(t)}{dt} = -0.25x(t) + u(t), \tag{60}$$

so the response depends on both the external input and the current concentration. We examine whether forecasts capture the resulting response amplitude, phase, and lag.

Non-periodic Inputs. We use pulses with controlled amplitude and onset:

$$u_h = a\,\mathbb{I}\{s \le h < s + w\}, \tag{61}$$

where a, s, and w denote the pulse amplitude, onset, and width. In the main PK and spatial experiments, we use $a \in \{0.5, 1, 2\}$, onsets at $H/8$ and $H/2$, and $w = \max(2, \lfloor H/8 \rfloor)$.

We also compare a single pulse with two half-amplitude pulses of the same width, preserving the total input while changing its timing. For spatial probes, we additionally change the source location. These comparisons test whether forecasts reflect the temporal and spatial placement of the input, as well as its magnitude.

Conditioning Controls. We distinguish four input conditions:

- **Correct conditions.** The model receives the future input used to produce the target trajectory.
- **Hidden conditions.** The future input is withheld, while the observed history and target trajectory remain unchanged.
- **Wrong conditions.** The model receives the baseline input instead of the actual future input, and is evaluated against the same target trajectory as in the correct-condition case.
- **Irrelevant conditions.** We add an auxiliary feature that does not affect the target dynamics. Two matched inputs share the same history and effective future conditions, but set this auxiliary feature to different future values.

The hidden and wrong conditions measure the value of accurate future information. The irrelevant-condition comparison examines whether the model changes its forecast in response to an input that has no effect on the target.

Response Evaluation. For two correctly supplied future inputs u_0 and u_1 with a common history, we compare the predicted response with the actual response:

$$\begin{aligned} \Delta\hat{y} &= \hat{y}(u_1) - \hat{y}(u_0), \\ \Delta y &= y(u_1) - y(u_0). \end{aligned} \tag{62}$$

We evaluate the discrepancy between $\Delta\hat{y}$ and Δy, together with response amplitude and timing errors. We also inspect predicted changes before the input takes effect and retain ordinary trajectory errors. Correct-condition response evaluation is kept separate from the hidden, wrong, and irrelevant-condition comparisons.

Experimental Settings. The main experiments use $C \in \{128, 512\}$, $H \in \{96, 256\}$, three seeds, and a sampling interval of 0.2. Historical inputs excite the dynamics before the forecast origin, after which alternative future inputs are applied from the same terminal state. Supplementary pulse experiments use 16 independent seeds, $H = 256$, amplitudes $\{0.5, 1.5\}$, a pulse width of eight steps, and a finer onset grid. In these supplementary experiments, the forecast origin is fixed when comparing context lengths. Input indices denote the intervals ending at the corresponding observations, and timing errors are evaluated using this convention.