WHEN EXPRESSIVITY DOES NOT HELP: A STUDY OF LINEAR AND NONLINEAR RNNS FOR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series forecasting requires models to capture temporal dependencies across diverse time scales and dynamics. Recent work has explored more expressive recurrent architectures for forecasting, including mechanisms with input-dependent transitions and explicit state-tracking capabilities. Several studies have reported performance improvements over simpler recurrent models. However, whether these gains systematically arise from greater recurrence expressivity, and whether capabilities emphasized in general sequence modeling are relevant to forecasting, remains unclear. We present a systematic empirical study of recurrent models with different levels of recurrence expressivity for time series forecasting. Overall, by using bootstrap hypothesis testing, we find that greater recurrence expressivity does not significantly improve forecasting accuracy. On controlled synthetic forecasting tasks, models with nonlinearity or explicit state-tracking mechanisms do not show a systematic advantage over less expressive architectures. Furthermore, on standard long-term forecasting and diverse real-world benchmarks, LTSF and GIFT-Eval, competitive linear models such as mLSTM can match or outperform several non-linear architectures. Moreover, within linear-recurrence models, added expressivity through input-dependent transitions or explicit state-tracking mechanisms does not reliably improve performance. Our results suggest that recurrence expressivity alone is *not* a reliable predictor of forecasting performance, highlighting the importance of dataset characteristics and architectural inductive biases beyond expressivity.

1 INTRODUCTION

Among neural architectures for time series forecasting, linear recurrent neural networks (LRNNs) and state-space models (SSMs), have recently gained increasing attention as a scalable alternative to Transformers and "non-linear" recurrent neural networks (RNNs). Compared to Transformers, they avoid the quadratic scaling of attention with sequence length, while their fixed-size hidden state may limit their memory capacity. Compared to RNNs [1], their structured linear recurrence enables more efficient parallel evaluation of sequences while retaining competitive forecasting performance.

A defining characteristic of SSMs and related models is their reliance on linear recurrence, in contrast to classical recurrent architectures such as RNNs (Elman, 1990) and LSTMs (Hochreiter & Schmidhuber, 1997), which incorporate nonlinear state updates. While linear recurrence is, in principle, more limited in expressivity, empirical results show that stacking SSM layers with interleaved MLP blocks, which introduce position-wise nonlinearities, can match Transformer-level performance across several language modeling benchmarks (Gu & Dao, 2024b; De et al., 2024; Peng et al., 2023). Beyond such nonlinear readouts, several architectural choices have been explored to increase the expressivity of linear recurrence itself. In particular, input-dependent (selective) transitions, popularized by Mamba (Gu & Dao, 2024b), allow the transition dynamics to vary with the input. At the same time, linear recurrent models often rely on structured transition parameterizations, such as diagonal or nearly diagonal matrices, to enable efficient computation and provide stability guarantees, which can impose additional constraints on their expressivity. Thus, the expressivity of

[1] We use the terms RNNs and non-linear RNNs interchangeably.

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.

linear recurrent models depends not only on the linearity of their recurrence, but also on the specific structural choices used to implement it. These expressivity considerations are particularly evident in an important class of synthetic tasks called state-tracking (Merrill et al., 2024; Sarrof et al., 2024; Grazzi et al., 2025; Shakerinava et al., 2026; Chung et al., 2026). In these tasks, a model must track the state of a finite-state system by updating its internal state based only on the current state and input, effectively simulating a finite-state automaton (FSA).

While originally studied primarily in the context of language modeling, state-tracking capabilities have recently been proposed as an important factor in time series forecasting. TiRex (Auer et al., 2025) argues that the state-tracking capability of xLSTM is critical for long-horizon forecasting, while TempoPFN (Moroshan et al., 2025) further argues that nonlinear recurrence is not necessary as long as the model supports state tracking, motivating its use of GatedDeltaProduct. This hypothesis has also been supported by xLSTM-Mixer (Kraus et al., 2025) and by the controlled architectural comparisons of Hartl et al. (2026), which report strong forecasting performance for state-tracking-capable recurrent architectures. At the same time, other recent models based on simpler linear recurrence, including FlowState (Graf et al., 2025), Reverso (Fu et al., 2026), and Mamba4Cast (Bhethanabhotla et al., 2024), achieve competitive forecasting performance without relying on recurrence mechanisms designed for state tracking.

Despite these contrasting results, direct evidence that state-tracking capability itself improves time series forecasting remains limited. In particular, Hartl et al. (2026) investigate the relationship between state-tracking capability and forecasting performance, but rely on standard synthetic state-tracking tasks such as parity, whose relevance to the temporal dependencies encountered in time series forecasting remains unclear. This leaves open whether the greater recurrence expressivity associated with state tracking provides a systematic advantage for forecasting. In this work, we find that it does not consistently translate into improved forecasting performance. We establish this through a systematic evaluation of 12 recurrent models on synthetic and real-world forecasting tasks. On SynTSBench (Tan et al., 2026), we evaluate models on controlled datasets designed to isolate distinct temporal dynamics, including trend and periodicity, dependency range, and robustness to noise. We then extend the evaluation to real-world time-series forecasting using the LTSF (Wu et al., 2021) and GIFT-Eval (Aksu et al., 2024) benchmarks, examining whether the same relationship between recurrence expressivity and forecasting performance persists across diverse datasets.

To sum up, we provide a systematic evaluation of nonlinear and linear RNNs/SSMs in time series forecasting, covering 12 recurrent models, with different state representations and transition mechanisms, across both controlled synthetic tasks from SynTSBench and real-world benchmarks, including LTSF and GIFT-Eval. We show that greater recurrent expressivity is not significantly associated with better forecasting performance. In particular, models with state-tracking capabilities do not show a systematic advantage over simpler architectures. These results suggest that the temporal patterns captured by these forecasting tasks do not generally require highly expressive recurrence mechanisms.

2 BACKGROUND

Time series forecasting A time series is a sequence of observations indexed by time. In this work, we focus on *regulary sampled* time series, represented as sequences $(x_t)_{t=1}^{T}$, where the index t denotes the sequential order of equally spaced observations and $x_t \in \mathbb{R}^d$ corresponds to the variates collected at time t. When $d = 1$, the time series is univariate and when $d > 1$, it is multivariate. The goal of time series forecasting is to predict future time steps given a window of historical data. More fomally, for a time series $(x_t)_{t=1}^{T}$, a context length l, and a horizon length h, time series forecasting consists of predicting $x_{t+1:t+h}$ given $x_{t-l:t}$.

Recurrent Models Recurrent neural networks (RNNs) (Elman, 1990) model sequences by recursively updating a hidden state at time t, denoted as h_t, based on the previous state and current input, x_t, $h_t = \sigma(W_h h_{t-1} + W_x x_t + b), \quad y_t = Ch_t$, where σ is a nonlinear activation. Despite their high expressive power, RNNs suffer from vanishing or exploding gradients (Bengio et al., 1994; Hochreiter, 1998; Pascanu et al., 2013), making long-range dependencies difficult to learn. Gated architectures such as LSTMs (Hochreiter & Schmidhuber, 1997) and GRUs (Cho et al., 2014) alleviate this by improving gradient propagation. Recent architectures further modify recurrent memory: xLSTM (Beck et al., 2024) introduces new memory and activation mechanisms, while M^2RNN (Mishra et al., 2026) increases state capacity through a matrix-valued hidden state.

Linear RNNs (LRNNs), and SSMs in their discretized formulations, model sequences through latent linear dynamics, $h_t = A(x_t)h_{t-1} + B(x_t), \quad y_t = C(x_t)h_t$, with x_t denoting the input at time t, h_t, the corresponding hidden state, and y_t the output. While SSMs in their most general form allow input-dependent coefficients, earlier architectures such as S4 (Gu et al., 2022c) and S5 (Smith et al., 2023) use time-invariant recurrences with constant A, B, and C matrices. Later architectures such as Mamba (Gu & Dao, 2024b; Dao & Gu, 2024) introduce input-dependent transition dynamics to increase expressivity. On the other hand, as discussed in Section A, structured parameterizations of the transition matrix A enable stable and efficient sequence modeling (Gu et al., 2022b;a; Smith et al., 2023), with implications for SSM-based time series foundation models discussed in Section 3. These models therefore range from time-invariant systems to input-dependent linear recurrent architectures; the specific models considered in our study are described in Section 4.1.

State-Tracking We use *state tracking* to refer to sequence-processing tasks in which each input specifies an operation that updates an underlying state, and the model must maintain this state as the sequence unfolds. Examples include parity and modular counting, where the state is updated by successive arithmetic operations, and permutation tasks, where inputs specify transformations of an underlying state. These tasks require the model to compose a sequence of state transitions and recover the resulting state at the end of the sequence. Properties of linear RNNs such as input dependence, the range of eigenvalues, and the structure of the transition matrix (e.g., diagonal versus full-matrix) directly affect their ability to perform these tasks (Merrill et al., 2024; Sarrof et al., 2024; Grazzi et al., 2025; Khavari et al., 2025; Shakerinava et al., 2026), which we refer to as their expressivity. Among our models, the classical nonlinear RNNs, sLSTM, and to some extent DeltaProduct, are the ones that can track state. DeltaProduct (Siems et al., 2025) extends DeltaNet for that purpose. An earlier step, DeltaNet$[-1, 1]$ (Grazzi et al., 2025), widened the eigenvalue range of DeltaNet to include negative eigenvalues and solved only a limited set of state-tracking tasks.

3 Related Work

Several approaches to time series foundation models have emphasized the importance of state tracking and recurrent dynamics. TiRex (Auer et al., 2025), a foundation model for univariate time series, incorporates nonlinear recurrent blocks based on xLSTM and argues that the state tracking capability of this model is a critical property for long-horizon forecasting and its strong performance on GIFT-Eval and Chronos benchmarks. TempoPFN (Moroshan et al., 2025) further argues that recurrence nonlinearity is not necessarily required for time series forecasting, as long as the model is capable of state tracking. They report comparable results on the evaluation benchmarks mentioned above. Their results suggest that a linear recurrence with appropriate state-tracking capabilities, such as GatedDeltaProduct (Siems et al., 2025), can be sufficient for the forecasting task. xLSTM-Mixer (Kraus et al., 2025) proposes a multivariate forecasting model based on xLSTM and argues for the importance of state tracking in sequence models. Based on this argument, their architecture uses only the sLSTM module of xLSTM, excluding the mLSTM component, which does not support state tracking. However, their ablation studies show that replacing sLSTM with not only mLSTM, but also other recurrent layers such as GRU and LSTM, leads to worse performance. Since GRU and LSTM are also capable of state tracking, these results leave open whether the observed benefits can be attributed specifically to state tracking. A recent study (Hartl et al., 2026) provides further support for the state-tracking hypothesis through head-to-head comparisons of several recurrent architectures, including GatedDeltaNet, Mamba-2, and xLSTM, reporting the strongest overall forecasting performance for xLSTM. The authors complement these experiments with synthetic tasks of parity and majority counting, to identify the capabilities underlying the observed differences, including state tracking.

In contrast, other works suggest that simpler linear recurrent structures may already be sufficient for time series forecasting. FlowState (Graf et al., 2025) builds on S5, a structured time-invariant linear SSM, while Reverso (Fu et al., 2026) shows that a linear DeltaNet backbone can match or outperform more expressive variants such as DeltaProduct, in the task of forecasting on GIFT-Eval. Furthermore, (Hartl et al., 2026) leaves it unclear whether their synthetic tasks directly capture capabilities that are relevant to time series forecasting. In particular, their state-tracking synthetic tasks are primarily used in language and math or code-related settings. This raises the question of whether state tracking, as measured by such tasks, is equally relevant to the temporal dependencies encountered in time series.

Table 1: Comparison of recurrent architectures across key design exprissivity axes. TI/ID: time-invariant/input-dependent. Transition: ND/D/PD = non-diagonal/diagonal/partially diagonal. Bold capitals are matrices, bold lowercase letters are vectors, and italic letters are scalars. $\mathbf{S}$ is a matrix state, $\mathbf{C}$ and $\mathbf{c}$ are mLSTM and sLSTM memory cells, and $\mathbf{h}$ is a vector state. For the nonlinear models, the input-dependence column is marked N/A. The distinction between input dependence and time invariance is used here for linear recurrent models. A nonlinear model can have a constant recurrent matrix and still yield an input-dependent Jacobian after linearization. For this study, that case is left unmarked rather than classified as input-dependent.

Model	Recurrence	TI/ID	Transition	State-Tracking	State update
RNN/LSTM/GRU	Nonlinear	N/A	ND	✓	$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t)$
M2RNN	Nonlinear	N/A	ND	✓	$\mathbf{S}_t = f_t \mathbf{S}_{t-1} + (1 - f_t)\tanh(\mathbf{S}_{t-1}\mathbf{W} + \mathbf{k}_t\mathbf{v}_t^\top)$
sLSTM	Nonlinear	N/A	ND	✓	$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{z}_t,\ \mathbf{z}_t = \tanh(\tilde{\mathbf{z}}_t)$ $\tilde{\mathbf{z}}_t = \text{linear}(\mathbf{x}_t, \mathbf{h}_{t-1})$
(Gated-)DeltaProduct	Linear	ID	PD	✓	$\mathbf{S}' = \alpha_t \mathbf{S}_{t-1}(\mathbf{I} - \beta_{t,1}\mathbf{k}_{t,1}\mathbf{k}_{t,1}^\top) + \beta_{t,1}\mathbf{v}_{t,1}\mathbf{k}_{t,1}^\top$ $\mathbf{S}_t = \mathbf{S}'(\mathbf{I} - \beta_{t,2}\mathbf{k}_{t,2}\mathbf{k}_{t,2}^\top) + \beta_{t,2}\mathbf{v}_{t,2}\mathbf{k}_{t,2}^\top$
Mamba	Linear	ID	D	✗	$\mathbf{S}_t = \mathrm{e}^{-(\boldsymbol{\Delta}_t \mathbf{1}^\top) \odot \mathrm{e}^{\mathbf{A}}} \odot \mathbf{S}_{t-1} + (\boldsymbol{\Delta}_t \odot \mathbf{x}_t)\mathbf{b}_t^\top$
Mamba 2	Linear	ID	D	✗	$\mathbf{S}_t = a_t \mathbf{S}_{t-1} + (\Delta_t \odot \mathbf{x}_t)\mathbf{b}_t^\top, \quad a_t = \mathrm{e}^{-\Delta_t A}$
DeltaNet	Linear	ID	PD	✗	$\mathbf{S}_t = \mathbf{S}_{t-1}(\mathbf{I} - \beta_t \mathbf{k}_t \mathbf{k}_t^\top) + \beta_t \mathbf{v}_t \mathbf{k}_t^\top$
mLSTM	Linear	ID	D	✗	$\mathbf{C}_t = f_t \mathbf{C}_{t-1} + i_t \mathbf{v}_t \mathbf{k}_t^\top$
LRU	Linear	TI	D	✗	$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t$
S5	Linear	TI	D	✗	$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t$

4 Experiments

4.1 Experimental Setup

Models We compare recurrent architectures spanning nonlinear and linear recurrence. The nonlinear models tested are RNN (Elman, 1990), GRU (Cho et al., 2014), LSTM (Hochreiter & Schmidhuber, 1997), sLSTM (Beck et al., 2024), and M^{2}RNN (Mishra et al., 2026). The linear RNN/SSM models benchmarked are S5 (Smith et al., 2023), Mamba (Gu & Dao, 2024a), Mamba-2 (Dao & Gu, 2024), Griffin (De et al., 2024), DeltaNet (Yang et al., 2024), Gated-DeltaProduct (Siems et al., 2025), mLSTM (Beck et al., 2024) and LRU (Orvieto et al., 2023). While the nonlinear models above use full transition matrices, the linear variants rely on structured matrices: diagonal for S5, Mamba, Mamba-2, LRU, and Griffin, and diagonal plus low-rank for DeltaNet and DeltaProduct. Furthermore, S5 and LRU use time-invariant transitions, while the other LRNNs/SSMs use input-dependent transitions and hence, are more expressive in theory. A complete list of the models and their characteristics is provided in table 1.

Architecture All models share the same high-level architecture. Each input sequence is first passed through a linear projection layer that maps the input feature dimension to the model's internal hidden dimension. The projected sequence is then processed by an encoder corresponding to the recurrent architecture under study. Layer normalization is applied to the encoder output, followed by a final linear projection that maps the hidden dimension to `num_features × pred_len`.

Normalization Following the Time-Series-Library protocol (Wang et al., 2026), each series is first normalized using the variance over the time series, using statistics computed from the training split. For LTSF and the synthetic data this is performed per channel; while for GIFT-Eval, it is performed over all training windows. As commonly done, we then apply reversible instance normalization (RevIN) at both training and inference: each context window is normalized using its own mean and standard deviation, the model operates in this normalized space, and forecasts are denormalized using the same window statistics. Metrics are finally computed on the dataset-level using the variance-normalized data.

Optimization : All models are trained with the AdamW optimizer (Loshchilov & Hutter, 2017). For each dataset and forecasting task, we select the hyperparameters with Optuna (Akiba et al., 2019). Each model receives the same budget of 24 trials: 8 random initialization trials followed by 16 guided trials. The search includes the training loss, which is either MSE or MAE, together with the other hyperparameters listed in Section G. The selected configuration is the one with the lowest validation

MSE on the synthetic and LTSF tasks, and the lowest validation MASE on GIFT-Eval. Early stopping uses the same validation score.

Experimental settings For each (model, dataset) pair, we trained multiple model settings spanning different depths (1, 2, 3), context lengths (96, 336, 720), and prediction horizons (96, 192, 336, 720 and short, medium, long). Each version was then retrained four additional times using different random seeds. Consequently, the results reported below are averaged over five independent seeds. Further details on the selected model–dataset combinations and experimental settings are provided in Section G.

4.2 Result Summary

Sections 4.3 and 4.4 present many detailed results; to summarize them we present a bootstrap-based hypothesis testing analysis in this section. A linear regression model was fitted using ordinary least squares (OLS) to predict variance-normalized MSE from the considered expressivity axes and control variables: datasets, context length, horizon and number of layers. We use binary encoding to encode each of the four expressivity axes (Table 1) and control variables. Specifically, we encode Non-linear, Linear-TI and Linear-ID as two binary variables; while, we encode state-tracking and transition a one binary variables each. We resample over datasets. We also note that the models tested do not cover the whole space of possible expressivity class combinations. Therefore, for some expressivity classes the test reduces to a pair-wise comparison between two models, reducing power, a limitation of the analysis. We find that across all three benchmarks considered, no statistically significant association was found between the expressivity classes and forecasting performance. *Thus, under our experimental protocol and after controlling for context and prediction lengths, architectural expressivity is not a statistically significant predictor of the MSE.* It is important to note that we show no significance not significance of absence.

Table 2: OLS coefficients (point estimate, 95% bootstrap CI, $n_{boot} = 10000$) predicting variance-normalized `test_mse` showing difference in performance between the features. For the features column, x vs y means a positive coefficient means y performs better and the features are the same as in table 1. Bold means 95% CI excludes zero. The results show that expressivity classes do not significantly predict forecasting performance.

Feature	SynTSBench	LTSF	GIFT
NonLinear vs Linear (TI)	-0.006 [-0.167, 0.163]	-0.224 [-0.835, 0.218]	-0.004 [-0.089, 0.091]
NonLinear vs Linear (ID)	0.018 [-0.140, 0.177]	0.480 [-0.080, 1.432]	0.021 [-0.044, 0.093]
State: Vector vs Matrix	0.006 [-0.123, 0.123]	0.613 [-0.035, 1.763]	0.005 [-0.038, 0.046]
ST: Yes vs No	0.003 [-0.162, 0.169]	0.023 [-0.208, 0.260]	-0.005 [-0.086, 0.075]
Trans.: ND vs D	-0.003 [-0.144, 0.126]	0.152 [-0.055, 0.516]	-0.016 [-0.086, 0.044]

4.3 Synthetic Experiments

Time series forecasting can be conceptualized as the combination of several underlying components. One key advantage of synthetic datasets is that they allow each of these components to be isolated, so that models can be evaluated under controlled conditions and on datasets with particular characteristics. These characteristics include different trend patterns, seasonal behaviors, and noise levels, as well as more complex properties such as multivariate structure and dependence between variables. *The goal of this section is to highlight that models with greater expressivity or state-tracking capability do not outperform simpler recurrent architectures on any of these isolated temporal dynamics.* Starting from the trend and periodic datasets, we first observe that strong or weak performance on these tasks does not depend strongly on the expressivity of the recurrent model. This result persists when noise is added to the periodic datasets, and again on the tasks that require capturing long-range dependence between variables. The final confirmation of the weak association between state-tracking capability and expressivity, on the one hand, and forecasting performance, on the other, is that the same pattern persists on composite synthetic datasets. These datasets combine the temporal features of the individual synthetic tasks so that they more closely resemble real-world time series.

Performance on Trend and periodic datasets Except for `Exponential-Trend`, `Gaussian-bump`, and `Quadratic-Trend` datasets, all models capture the underlying trends well (Table 3). Similarly, for almost all periodic datasets, the performance of all models is near-perfect, with the exception of the `Square-Wave` dataset, as shown in Table 3. *Notably, S5 and*

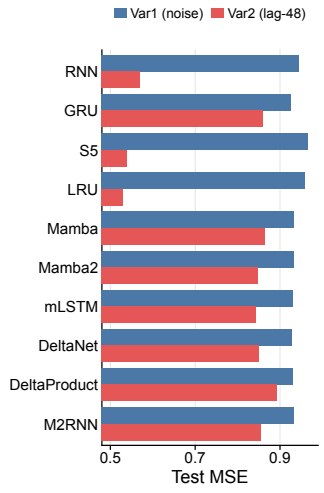
Table 3: Mean test MSE per dataset, grouped by domain,averaged over 3 horizons and 5 seeds. Best result in **bold**, second best underlined. Average ranks are computed over the datasets of each domain. The table indicates, by the lack of no one model having consistently bold results, that expressivity does not predict well the performance of a method and that in our results there is no best recurrent architecture.

Domain	Dataset	GRU	sLSTM	mLSTM	S5	RNN	LRU	LSTM	Mamba-2	Mamba	DeltaNet	DeltaProduct	M²RNN	Optimal
Trend	Exponential-Trend	0.0068	0.0068	0.0069	0.0147	0.0062	0.0131	0.0061	0.0071	**0.0059**	0.0063	0.0063	0.0081	0
	Gaussian-Bump	0.0070	0.0051	0.0133	0.0058	0.0094	0.0056	0.0168	0.0075	**0.0046**	0.0081	0.0051	0.0101	0
	Gompertz-Trend	6.22e-5	4.52e-7	**1.54e-7**	1.01e-6	9.66e-6	1.78e-6	1.06e-5	2.07e-6	2.4e-7	2.7e-7	1.41e-6	8.58e-7	0
	Linear-Trend	6.73e-7	6.44e-7	7.18e-7	1.38e-6	**6.04e-7**	7.76e-7	6.14e-7	7.78e-7	6.17e-7	6.56e-7	6.75e-7	9.39e-7	0
	Logarithmic-Trend	4.78e-5	**2.65e-6**	5.86e-6	8.12e-6	6.73e-5	7.93e-6	4.86e-5	6.4e-6	2.97e-6	3.8e-6	1e-5	5.94e-6	0
	Logistic-Trend	4.96e-4	2.42e-4	2.18e-4	1.05e-4	5.93e-4	3.11e-4	5.67e-4	0.0012	**4.16e-5**	2.12e-4	5.89e-4	5.45e-4	0
	Neg-Expo-Trend	1.29e-8	1.41e-8	1.52e-8	6.06e-8	**1.27e-8**	2.05e-8	1.28e-8	1.55e-8	1.37e-8	1.35e-8	1.35e-8	1.67e-8	0
	Piecewise-Linear	**7.94e-4**	0.0011	0.0064	0.0054	0.0011	0.0035	0.0026	0.0030	0.0023	0.0045	0.0048	0.0034	0
	Power-Law-Trend	8.01e-5	1.04e-5	3.82e-6	1.48e-5	7.99e-5	7.23e-6	8.35e-5	2.05e-5	**2.67e-6**	7.14e-6	1.47e-5	1e-5	0
	Quadratic-Trend	0.0080	0.0014	5.05e-4	4.54e-4	0.0085	9.57e-4	0.0090	0.0018	4.02e-4	2.23e-4	0.0015	**1.73e-4**	0
	Step	2.63e-18	3.8e-15	5.59e-15	7.69e-15	**0.0000**	5.59e-15	1.32e-17	4.23e-15	2.63e-15	6.04e-15	1.35e-15	4.33e-15	0
	Avg. rank ↓	6.77	4.68	6.59	8.36	6.50	7.59	7.36	8.45	**2.64**	5.09	6.68	7.27	–
Periodic	Double-Sine	**7.12e-6**	7.22e-5	1.6e-5	4.85e-4	1.17e-5	2.66e-4	1.53e-5	9.09e-5	1.33e-5	1.21e-5	1.89e-5	1.93e-5	0
	Exp-Sine-Wave	5.55e-5	1.01e-4	2.51e-5	2.56e-4	5.28e-5	2.41e-4	1.28e-5	1.15e-4	**1.01e-5**	4.58e-5	3.51e-5	3.51e-5	0
	Five-Sine	2.34e-5	3.15e-4	2.22e-5	6.36e-4	1.64e-4	7.07e-4	2.92e-4	2.84e-4	1.85e-5	**1.71e-5**	0.0013	3.7e-5	0
	Sawtooth-Wave	4.79e-6	1.88e-5	2.19e-5	3.65e-4	4.58e-6	1.35e-4	**4.51e-6**	8.16e-5	3.48e-5	9.8e-6	2.6e-4	1.86e-5	0
	Sine-Fast	3.55e-5	2.88e-5	1.48e-5	2.17e-4	4.49e-5	1.54e-4	1.58e-5	7.94e-5	**1.06e-5**	1.42e-5	2.52e-5	2.09e-5	0
	Sine-Slow	3.91e-5	8.56e-5	6.27e-5	2.53e-4	1.94e-4	8.25e-4	6.54e-5	7.04e-5	**1.31e-5**	1.39e-5	4.1e-5	4.49e-5	0
	Square-Wave	0.1027	0.1088	0.1111	**0.0901**	0.1025	0.0911	0.1054	0.1359	0.1156	0.1467	0.1120	0.1136	0
	Ten-Sine	6.62e-4	3.81e-4	5.83e-5	5.33e-4	4.4e-4	0.0021	**1.19e-5**	2.71e-4	6.35e-5	1.18e-4	4.2e-4	1.86e-5	0
	Triangle-Wave	9.66e-6	**7.92e-6**	1.17e-5	8.19e-4	2.93e-5	2.27e-4	2.33e-4	9.62e-5	1.98e-5	2.39e-5	3.17e-5	1.27e-5	0
	Triple-Sine	6.01e-5	9.66e-5	1.45e-5	4.88e-4	1.24e-5	1.51e-4	1.53e-5	4.85e-4	**1.17e-5**	1.44e-5	3.39e-5	1.97e-5	0
	Avg. rank ↓	5.20	7.20	4.50	10.40	5.70	10.00	4.90	9.10	**3.70**	4.40	7.55	5.35	–
Noise-Robustness	Five-Sine-Snr-10db	0.9310	0.9557	0.9557	**0.8932**	0.9272	0.9146	0.9252	0.9563	0.9551	0.9577	0.9563	0.9490	0.894
	Five-Sine-Snr0db	0.5185	0.5305	0.6282	0.5127	0.5410	0.5259	**0.5086**	0.6277	0.5798	0.5982	0.6289	0.5238	0.492
	Five-Sine-Snr10db	0.0978	0.1008	0.1053	**0.0942**	0.0982	0.0944	0.0943	0.1139	0.1044	0.1057	0.1117	0.0977	0.090
	Avg. rank ↓	4.33	7.17	9.50	**1.33**	5.67	3.33	2.00	10.83	7.67	10.33	11.17	4.67	–
Cross-Variable	Conditional-Relationship	0.9966	1.0023	1.0006	**0.9822**	0.9838	0.9833	0.9992	0.9994	1.0028	0.9995	1.0029	0.9981	0.902
	Lag-10	0.9590	0.9551	0.9558	0.9577	**0.9499**	0.9543	0.9605	0.9554	0.9592	0.9547	0.9623	0.9562	0.944
	Lag-24	0.9908	0.9909	0.9876	0.9686	**0.9564**	0.9635	0.9924	0.9859	0.9879	0.9867	0.9940	0.9872	0.935
	Lag-48	0.9070	0.9091	0.9051	0.8645	0.8658	**0.8540**	0.9082	0.9067	0.9084	0.9072	0.9139	0.9080	0.801
	Lag-5	1.0533	1.0570	1.0550	1.0593	**1.0512**	1.0543	1.0554	1.0556	1.0551	1.0567	1.0564	1.0548	1.049
	Multivariate-Complex	0.9632	0.9655	0.9568	0.9510	**0.9411**	0.9460	0.9710	0.9543	0.9607	0.9530	0.9631	0.9553	0.918
	Nonlinear-Relationship	0.9407	0.9430	0.9404	**0.8823**	0.8902	0.8886	0.9438	0.9399	0.9430	0.9412	0.9459	0.9406	0.836
	Sin-Noise-Composite	0.4693	0.4702	0.4694	0.4700	0.4692	0.4698	**0.4691**	0.4702	0.4696	0.4704	0.4696	0.4695	0.467
	Avg. rank ↓	6.25	9.62	5.88	4.88	**1.88**	2.75	8.50	6.06	8.62	7.12	10.56	5.88	–
Composite	Economic-Indicator	0.2978	0.2631	0.3320	0.3224	**0.2158**	0.2584	0.2638	0.4298	0.3108	0.5525	0.6368	0.5451	0
	Electricity-Consumption	0.1014	0.1037	0.0850	**0.0540**	0.1751	0.0910	0.0805	0.1075	0.0661	0.1381	0.1003	0.0569	0
	Industrial-Sensor	0.3997	0.2463	0.4524	**0.1040**	0.3703	0.1284	0.4184	0.2468	0.2984	0.3417	0.4736	0.3543	0
	Iot-Sensor-Battery	0.0011	6.67e-5	3.74e-5	4.74e-5	9.89e-4	3.92e-5	0.0011	7.69e-5	**3.49e-5**	3.79e-5	4.8e-5	1.01e-4	0
	Network-Traffic	0.0281	0.0671	0.0330	0.0264	0.0696	0.0449	**0.0249**	0.0434	0.0349	0.0375	0.0409	0.0272	0
	Retail-Sales	0.4914	0.5660	0.5895	**0.3988**	0.5203	0.4874	0.4876	0.5634	0.5523	0.5338	0.5456	0.5090	0
	Stock-Price	1.0900	1.0991	1.0905	0.9968	0.9654	1.4001	**0.9446**	1.1105	1.0839	1.0981	1.0974	1.0058	0
	Temperature-Sensor	0.0669	0.0729	0.0785	0.0765	**0.0656**	0.0721	0.0889	0.0861	0.0726	0.0853	0.0844	0.0796	0
	Website-Traffic	0.0621	0.0684	0.0544	**0.0350**	0.1849	0.0838	0.0470	0.0831	0.0624	0.0536	0.0493	0.0436	0
	Avg. rank ↓	6.28	7.56	7.00	**3.00**	7.11	5.78	5.50	9.11	5.22	7.67	8.22	5.56	–
Overall avg. rank ↓		6.00	7.07	6.24	6.49	5.48	6.52	6.18	8.46	**5.00**	6.27	8.32	5.96	–

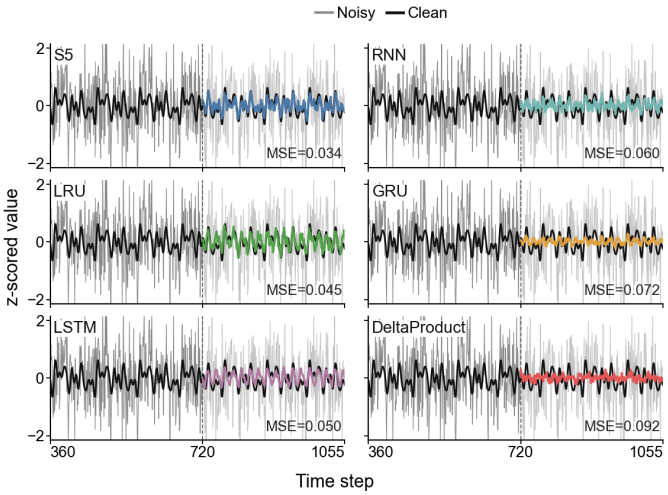
LRU, which are theoretically among the less expressive architectures because their transitions are time-invariant and diagonal, achieve the best results. This is consistent with these synthetic tasks not requiring state-tracking capability.

Tables 6 and 7 show the results for different horizon lengths. It is worth noting that performance on trend tasks degrades with horizon, while periodic tasks stay largely stable. That is expected: once period and phase are identified, the linear head decodes a bounded repeating signal; extra horizon is just more samples of the same cycle. Trends are open trajectories, hence, misspecification of slope or curvature accumulate errors with horizon length, especially for accelerating trends (exponential, quadratic, power-law), as noted in Tan et al. (2026). Linear and most saturating trends remain easy even at long horizon.

Robustness to Noise The third block of Table 3 extends the result of the periodic-data evaluation to signals with Gaussian noise at several levels. *These results show that, as noise is added to the signal, simpler models remain competitive with more expressive ones.* Here, the signal-to-noise ratio, following Tan et al. (2026), runs from -10dB (most noise) to 10dB (least noise). The optimal forecast is the clean signal, so the optimal MSE is the mean squared noise over the horizon (last column), which was zero for clean signals in clean Trend and Periodic datasets. At -10dB, S5 and LRU outperform the nonlinear models, including sLSTM. Figure 1(b) provides an illustrative example of the strong performance of linear recurrent models, showing how S5 and LRU recover the underlying signal in a noisy time series more faithfully than more expressive models.



(a) Lag-48: seed-mean test MSE. Var1 is white noise; var2 is its 48-step delay.



(b) Five-sine, SNR -10 dB, one test window. Gray: noisy observation; black: noiseless signal; color: forecast. Dashed line marks forecast start; MSE is versus the clean signal.

Figure 1: For some temporal structure, S5 recovers the target dynamics more faithfully than more expressive models. (a) Only S5, LRU, and RNN drive lag-48 channel error well below white-noise MSE. (b) On a five-sine series at -10 dB, S5 tracks the clean signal most closely. LRU, LSTM, and RNN keep more of the oscillation than GRU and DeltaProduct.

Cross-Variable Dependency Next we look at the cross-variable dependency datasets, which combine long-range dependence with interactions between variables. Figure 1(a) shows each model's performance in capturing the dependence of one variable (Var2) on another (Var1). Var1 is white noise, and Var2 is a copy of Var1 delayed by 48 steps. The optimal forecast therefore has a lower MSE on Var2 than on Var1, because part of the horizon is determined by the observed lag. This is the case for LRU and S5, as well as for some nonlinear RNNs, as reflected in the corresponding block of Table 3. More expressive models, including those that can track state, do not improve on this result. This is consistent with the lag-48 dependency not requiring state-tracking capability.

Evaluation on Realistic Synthetic Time Series The last block of Table 3, marked Composite, provides the strongest support for our claim of a weak association between model performance and expressivity. It evaluates models on datasets designed to simulate real-world time series. Interestingly, S5 achieves strong overall performance despite belonging to the lowest expressivity class, particularly in state-tracking capability. Moreover, comparing mLSTM with sLSTM, and DeltaNet with DeltaProduct, their respective state-tracking counterparts, we observe comparable performance. Our ablation in Section H, which compares DeltaNet-family models of increasing expressivity, provides further evidence for this observation.

Datasets features Finally, the left panel of Figure 2 presents the overall performance of the models across the different dataset groups. For each model and group, we take the mean test MSE over the group's datasets and forecast horizons (96, 336, 720). Denoising and Anomaly are scored against the clean signal. The score is $-\log_{10}$ of that mean MSE, min–max normalized within each group: 1 (blue) is the best model in the group, 0 (red) the worst. S5 achieves the best performance in 5 of the 9 categories. Long-range dependency and multivariate time series designed to simulate real-world dynamics appear to be the most challenging categories for forecasting.

4.4 Real-World Experiments

To examine if our results on synthetic data hold on real data we perform experiments using two common real-world benchmarks: LTSF and GIFT-Eval. For LTSF, we conduct experiments on all nine datasets: `ETTh1`, `ETTh2`, `ETTm1`, `ETTm2`, `Weather`, `Traffic`, `Electricity`, `Exchange` and `Illness`, taken from (Ailing Zeng & Xu, 2023), which span

Table 4: LTSF, test MSE averaged over horizons. For each dataset, model, and horizon, hyperparameters (context length, n_layers, loss) are chosen by lowest seed-averaged val MSE. Each cell is the mean of those seed-averaged test MSEs over the dataset's available horizons. Bold = best in row, underline = second-best. The results show that mLSTM and sLSTM perform the best across the datasets. However, in terms of expressivity they are quite different, again suggesting expressivity is not a good predictor of forecasting performance.

Dataset	RNN	GRU	LSTM	S5	LRU	Griffin	Mamba	Mamba2	mLSTM	sLSTM	DeltaNet	DeltaProduct	M2RNN
ETTh1	0.5606	0.5759	0.6197	0.5936	**0.5556**	0.6159	0.6001	0.6035	0.5945	0.6294	0.6396	0.6175	0.6348
ETTh2	0.5348	0.4643	0.8046	0.5071	0.5493	0.5057	0.6894	0.5657	**0.4390**	0.4432	0.5738	0.5407	0.6530
ETTm1	0.4694	0.4773	0.4722	**0.4392**	0.4671	0.5287	0.5275	0.5614	0.4855	0.4846	0.5978	0.5630	0.5036
ETTm2	0.3762	0.3457	1.8462	0.3650	0.5134	0.6578	0.7759	0.4981	0.3092	**0.3082**	0.7949	0.3665	1.0399
Weather	0.2723	0.2747	0.2693	0.3234	0.3303	0.3323	0.3278	0.3411	0.2580	**0.2470**	0.3268	0.3302	0.3577
Traffic	0.7332	0.7884	0.9452	0.8037	0.8344	0.7794	0.7040	0.7094	0.6744	**0.6624**	0.7087	0.7335	0.7597
Electricity	0.3808	0.2697	0.4637	0.5606	0.5159	0.4254	0.3941	0.3736	**0.1980**	0.1988	0.4136	0.3743	0.3744
Exchange	0.4125	0.3477	0.3328	0.3642	0.7086	1.1097	0.4089	0.4554	0.3198	**0.3124**	0.3237	0.6163	0.4053
Illness	1.3483	1.3754	1.5054	1.6663	1.7432	1.4474	2.0559	1.5254	**0.9509**	1.0290	1.2532	1.7586	1.0505
Average rank ↓	5.33	4.78	8.78	6.67	8.44	9.00	8.56	8.22	**2.56**	3.00	8.22	8.56	8.89

Table 5: GIFT-Eval test MSE averaged over forecast terms (short, medium, long). Bold: best in row; underline: second-best. The results show that no method consistently outperforms the others. For instance, no model is best for more than 2 datasets.

Domain	Dataset	RNN	GRU	LSTM	S5	LRU	Griffin	Mamba	Mamba2	mLSTM	sLSTM	DeltaNet	DeltaProduct	M2RNN
Energy	ETT1/H	0.7652	**0.6111**	0.7231	0.6648	0.6512	0.6913	0.6836	0.7331	0.7363	0.7033	0.6845	0.6188	0.7252
	ETT1/15T	0.5777	0.5838	0.4921	**0.4088**	0.4843	0.5437	0.5164	0.6228	0.5889	0.5832	0.5470	0.5402	0.4906
	ETT2/H	**0.1933**	0.2077	0.2334	0.2040	0.2392	0.2395	0.1985	0.3928	0.3428	0.3766	0.2552	0.3533	0.2285
	ETT2/15T	**0.1313**	0.1681	0.1444	0.1472	0.1556	0.1519	0.1467	0.1724	0.1649	0.1679	0.1608	0.1728	0.1546
	Electricity/H	0.0426	0.1087	**0.0405**	0.0445	0.0530	0.0509	0.0546	0.0485	0.0462	0.0495	0.0496	0.0738	0.0540
	Solar/H	0.2788	0.3858	0.2989	0.2585	**0.2415**	0.2780	0.2713	0.2814	0.2958	0.2919	0.2856	0.2928	0.2553
Healthcare	US-Births/D	0.2732	0.2127	0.2188	0.1881	0.2179	0.2344	0.2300	0.2373	0.2486	0.2429	0.1997	0.1847	0.2510
	US-Births/W	0.3701	0.3714	0.3318	0.3505	0.3300	0.3665	0.4561	0.3729	0.3845	0.4569	0.4378	0.3528	0.3573
Nature	Jena-Weather/H	0.4672	0.4568	0.5086	0.4359	0.4598	0.5315	0.5268	0.4888	0.4710	0.4516	0.6041	0.4874	0.5623
	Saugeen/D	1.1689	1.1028	1.0566	0.8890	0.9941	0.9749	1.1192	1.0544	0.9999	1.0361	1.0723	1.1926	0.9319
	Saugeen/W	1.0279	1.2149	1.0622	0.9876	0.9173	1.1983	1.2335	1.1787	1.0755	1.0629	1.2006	1.2467	1.0286
Web/CloudOps	BizITObs-App	0.5963	0.3116	0.1438	0.1295	0.1961	0.1212	0.1223	0.1708	0.1212	**0.1169**	0.1767	0.1435	0.1535
	BizITObs-L2C/H	0.6665	0.6842	**0.6652**	0.7645	0.6825	0.7577	0.9446	1.1642	0.9399	1.0563	0.7007	0.7448	0.7569
	BizITObs-Service	0.0156	0.0147	0.0122	0.0148	0.0166	0.0135	0.0184	0.0170	**0.0118**	0.0146	0.0195	0.0165	0.0160
Average rank ↓		6.71	7.86	5.29	**3.36**	4.86	6.46	7.79	9.64	7.89	7.79	8.57	8.07	6.71

a diverse range of domains (electricity, traffic, finance, and environment). The second real-world benchmark is a subset of 14 datasets from GIFT-Eval (Aksu et al., 2024), spanning four domains, ten sampling frequencies, univariate and multivariate series.

The results for LTSF and GIFT-Eval are presented in Tables 4 and 5, respectively. We average the results over four horizon lengths (96, 192, 336, and 720) for LTSF and three (short-, mid-, and long-term) for GIFT-Eval. We always report averaged results over five seeds. Per-horizon results are given in Table 14 for LTSF and in Section D for each GIFT-Eval domain, along with the result for PatchTST and DLinear for some subset of tasks. The main observation from the LTSF and GIFT-Eval results is consistent with the performance on realistic synthetic tasks: recurrent models with lower expressivity achieve overall performance comparable to that of more expressive models. Table 4 shows it, for example, in the test MSE results of mLSTM on LTSF.

Datasets features. Unlike synthetic datasets, real-world datasets cannot be readily categorized into the trend, periodic, and other groups introduced above. Nevertheless, we can characterize each dataset based on its statistical features and compare model performance across these feature-based groups. Following the taxonomy used in Aksu et al. (2024), we consider six time-series features (trend, seasonality, entropy, hurst exponent, stability, and lumpiness, with definitions and formulas provided in Appendix F). We compute these properties for each dataset (Tables 19 and 20) and analyze models performance along the corresponding features. Figure 2 (center and right) shows how each model's performance depends on time series features such as trend and seasonality; more negative values indicate better performance on datasets where that feature is more pronounced. The main observation is that LTSF and GIFT-Eval yield different relationships between model performance and dataset features. For instance, S5 and LRU tend to perform better on LTSF datasets with a strong trend, but this pattern does not consistently hold on GIFT-Eval. Similarly, Mamba, mLSTM, GRU, and M^2RNN tend to perform better on LTSF datasets with strong seasonality, but not on similar GIFT-Eval datasets. These discrepancies show how difficult it is to identify general relationships between dataset characteristics and model performance on noisy real-world datasets.

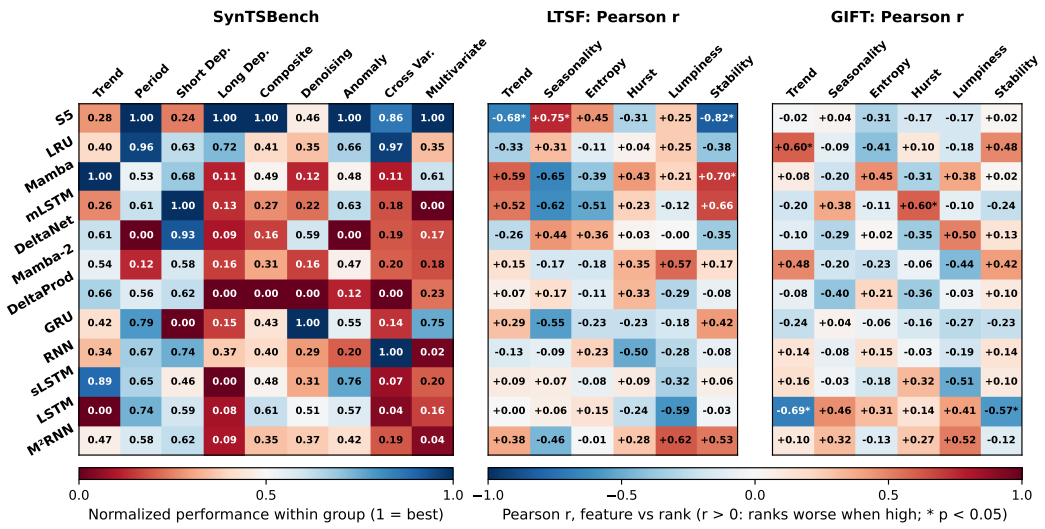


Figure 2: (Left) Performance heatmap of the evaluated models across *synthetic* dataset categories. Values are log-scaled test MSE, min–max normalized across models within each category, so that on each column 1 is the best model and 0 the worst (higher is better). (Center) Pearson correlation between model performance and dataset features on LTSF and (Right) GIFT-Eval. While there is no single best model across different temporal dynamics or features, S5 shows better performance than more expressive models.

5 Conclusion

We presented a controlled empirical comparison of 12 linear and nonlinear recurrent architectures for time-series forecasting. Based on our results on both synthetic and real-world time series datasets we conclude that greater formal expressivity, and state-tracking in particular, does not automatically translate into better forecasting performance. Especially, on synthetic tasks, Linear SSMs such as S5 and LRU, recover delayed structure and remain robust to noise, better than more expressive gated and selective recurrences models. These results underscore the importance of comprehensive benchmarking of recurrent architectures in time-series forecasting, especially as their popularity grows. It suggests that in future works, exploring other architectural differences to improve forecasting could be more fruitful than expressivity.

6 Limitations

Our study is subject to several limitations. First, our hyperparameter search was designed to allocate equal compute budgets across models. This ensures fairness but also means that architectures with larger hyperparameter spaces may not have been explored as thoroughly as those with fewer tunable components. Relaxing this constraint is an important next step for obtaining a more complete picture of each model's best-case performance. Second, while we explore 12 recurrent architectures, there are many more that could be included in our benchmarking. For example, recurrent architectures that specifically target time series forecasting. Third, along the same lines as the previous point, we do not explore all combinations of expressivity axes. This is mainly due to a lack of models which are in some parts of the expressivity space. This impacts our hypothesis testing as it reduces the power at which we can compare some of the expressivity axes. Last, we aimed for a simple hypothesis test to summaries our many experiments. However, we acknowledge that more sophisticated testing methodologies could be used, which could increase the power of our results and potentially allow us to say there significantly no association between expressivity and forecasting performance. We leave this as a key direction of future work.

AI USE STATEMENT

In this work, we did not use generative AI tools for any of the tasks that require disclosure. We have not used generative AI tools to: generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools for: aiding in editing the paper text, creation of figures and generating code for experiments designed by the authors. We have reviewed all AI-assisted work. Specifically, we reviewed any suggested paper edits for hallucinated references or factual errors and tested all generated code to ensure correctness. The oversight of generated text and code was conducted by multiple authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All the benchmarks and models used in this work are open source and we use their public implementations. Additionally, we report hyperparameter search spaces used in section G and experimental details in section 4.1. Therefore, the results in the paper should be able to be easily reproduced.

REFERENCES

Lei Zhang Ailing Zeng, Muxi Chen and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. 2019. doi: 10.48550/ARXIV.1907.10902. URL https://arxiv.org/abs/1907.10902.

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, 2024. URL https://arxiv.org/abs/2410.10393.

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. *arXiv preprint arXiv:2505.23719*, 2025.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, March 1994. ISSN 1941-0093. doi: 10.1109/72.279181. URL http://dx.doi.org/10.1109/72.279181.

Sathya Kamesh Bhethanabhotla, Omar Swelam, Julien Siems, David Salinas, and Frank Hutter. Mamba4cast: Efficient zero-shot time series forecasting with state space models. *arXiv preprint arXiv:2410.09385*, 2024.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, October 2014.

Jiwan Chung, Heechan Choi, and Seon Joo Kim. Rethinking state tracking in recurrent models through error control dynamics. *arXiv preprint arXiv:2605.07755*, 2026.

Federico Danieli, Pau Rodriguez, Miguel Sarabia, Xavier Suau, and Luca Zappella. Pararnn: Unlocking parallel training of nonlinear rnns for large language models. *arXiv preprint arXiv:2510.21450*, 2025.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing gated linear recurrences with local attention for efficient language models, 2024.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

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.

Lars Graf, Thomas Ortner, StanisŁ WoŁşniak, Angeliki Pantazi, et al. Flowstate: Sampling rate invariant time series forecasting. *arXiv preprint arXiv:2508.05287*, 2025.

Riccardo Grazzi, Julien Siems, Jörg K.H. Franke, Arber Zela, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *The Thirteenth International Conference on Learning Representations*, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024a.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024b.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Conference on Advances in Neural Information Processing Systems*, 2022a.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022b.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022c.

Anamaria-Roberta Hartl, Levente Zólyomi, David Stap, Pieter-Jan Hoedt, Niklas Schmidinger, Lukas Hauzenberger, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. On subquadratic architectures: From applications to principles. *arXiv preprint arXiv:2606.12364*, 2026.

Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02): 107–116, 1998.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, nov 1997. ISSN 0899-7667.

Arjun Karuvally, Franz Nowak, Anderson T Keller, Carmen Amo Alonso, Terrence J Sejnowski, and Hava T Siegelmann. Bridging expressivity and scalability with adaptive unitary ssms. *arXiv preprint arXiv:2507.05238*, 2025.

Behnoush Khavari, Mehran Shakerinava, Jayesh Khullar, Jerry Huang, François Rivest, Siamak Ravanbakhsh, and Sarath Chandar. Parity requires unified input dependence and negative eigenvalues in ssms. *arXiv preprint arXiv:2508.07395*, 2025.

Maurice Kraus, Felix Divo, Devendra Singh Dhami, and Kristian Kersting. xlstm-mixer: Multivariate time series forecasting by mixing via scalar memories. In *Advances in Neural Information Processing Systems*, volume 38, pp. 7335–7369, 2025. doi: 10.52202/085713-0228.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning*, 2024.

Mayank Mishra, Shawn Tan, Ion Stoica, Joseph Gonzalez, and Tri Dao. M2 rnn: Non-linear rnns with matrix-valued states for scalable language modeling. *arXiv preprint arXiv:2603.14360*, 2026.

Vladyslav Moroshan, Julien Siems, Arber Zela, Timur Carstensen, and Frank Hutter. Tempopfn: Synthetic pre-training of linear rnns for zero-shot time series forecasting. *arXiv preprint arXiv:2510.25502*, 2025.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, 23–29 Jul 2023.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, 2013.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. Rwkv: Reinventing rnns for the transformer era. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 14048–14077, 2023.

Yash Sarrof, Yana Veitsman, and Michael Hahn. The expressive capacity of state space models: A formal language perspective. In *Conference on Advances in Neural Information Processing Systems*, 2024.

Mehran Shakerinava, Behnoush Khavari, Siamak Ravanbakhsh, and Sarath Chandar. The expressive limits of diagonal ssms for state-tracking. *arXiv preprint arXiv:2603.01959*, 2026.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. Deltaproduct: Improving state-tracking in linear rnns via householder products. *arXiv preprint arXiv:2502.10297*, 2025.

Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023.

Qitai Tan, Yiyun Chen, Mo Li, Ruiwen Gu, Yilin Su, and Xiao-Ping Steven Zhang. Syntsbench: Rethinking temporal pattern learning in deep learning models for time series. *Advances in Neural Information Processing Systems*, 38, 2026.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37: 115491–115522, 2024.

A Background on Expressivity Limits of SSMs and LRNNs

A.1 Limited Expressivity of SSMs and Linear RNNs

Linear RNNs impose structural constraints on the transition matrix to enable stable and efficient sequence modeling. While beneficial computationally, these constraints reduce expressivity on an important class of learning problems known as state tracking. Prior work (Sarrof et al., 2024; Grazzi et al., 2025) shows that models with either time-invariant transitions or time-varying transitions whose eigenvalues lie in a restricted range are incapable of performing arbitrary state tracking. Subsequent work (Karuvally et al., 2025) extends this line by introducing complex-valued diagonal SSMs, which exhibit improved state tracking capabilities. However, Shakerinava et al. (2026) demonstrate that the diagonality of the transition matrix fundamentally limits expressivity, restricting these models to a solvable subset of group tracking tasks. This limitation also applies to recent variants such as Mamba-3 (Mishra et al., 2026), which extends the transition matrix from the real to the complex domain while preserving its diagonal structure. Notably, similar structural constraints arise in recent approaches to parallelizing nonlinear recurrent models, including ParaRNN, ParaLSTM, and ParaGRU (Danieli et al., 2025), implying comparable expressivity limitations in these settings.

A.2 SSMs with Higher Expressivity

Siems et al. (2025) propose DeltaProduct, an SSM that departs from strictly diagonal transitions to improve expressivity on state tracking tasks. This is achieved by introducing a low-rank, non-diagonal correction to the transition matrix, controlled by a rank hyperparameter that trades off expressivity against computational efficiency. While this relaxation enhances modeling capacity, it reduces the efficiency benefits typically associated with diagonal structure. In contrast, xLSTM (Beck et al., 2024) enhances state tracking by reintroducing nonlinearity within its sLSTM modules. This design improves expressivity without modifying the linear transition structure directly, but comes at the cost of losing full parallelizability, as the resulting nonlinear recurrence remains inherently sequential.

B Detailed Synthetic Benchmark Results

This appendix reports full per-dataset test MSE on the synthetic SynTSBench-style suites. For each dataset, model, and forecast horizon $H \in \{96, 336, 720\}$, hyperparameters (context length, n_{layers}, d_{model}, learning rate, and loss) are selected by the lowest seed-averaged validation MSE. Reported numbers are seed-averaged test MSE over available seeds. Horizons are not averaged. In each row, the best model is shown in **bold** and the second-best is underlined. Models: DeltaNet, DeltaProduct, GRU, LRU, LSTM, M2RNN, Mamba, Mamba2, RNN, S5, mLSTM, and sLSTM.

Table 6: Trend: test MSE. For each dataset, model, and horizon, hyperparameters (context length, n_layers, d_model, lr, loss) are chosen by lowest seed-averaged val MSE. Reported values are seed-averaged test MSE (existing seeds only). Bold = best in row, underline = second-best. Horizons are not averaged.

Dataset	H	DeltaNet	DeltaProduct	GRU	LRU	LSTM	M2RNN	Mamba	Mamba2	RNN	S5	mLSTM	sLSTM
Exponential-Trend	96	3.11e-03	2.91e-03	3.06e-03	0.0121	2.80e-03	6.03e-03	2.79e-03	5.13e-03	**2.39e-03**	0.0138	3.70e-03	3.27e-03
	336	2.76e-03	2.72e-03	2.70e-03	4.40e-03	2.85e-03	**2.37e-03**	2.51e-03	2.62e-03	2.58e-03	0.0102	2.85e-03	2.87e-03
	720	0.0131	0.0133	0.0147	0.0234	0.0127	0.0158	**0.0125**	0.0135	0.0136	0.0213	0.0141	0.0142
Gaussian-Trend	96	4.52e-05	3.62e-04	2.75e-04	7.42e-04	2.55e-04	2.95e-05	8.12e-05	3.61e-04	3.05e-04	1.46e-04	1.57e-04	**2.19e-05**
	336	7.86e-04	8.02e-04	6.90e-04	6.76e-04	**5.52e-04**	7.41e-04	7.73e-04	7.51e-04	6.73e-04	6.82e-04	5.67e-04	6.98e-04
	720	0.0235	0.0143	0.0200	0.0155	0.0495	0.0294	**0.0128**	0.0215	0.0226	0.0165	0.0391	0.0147
Gompertz-Trend	96	3.65e-09	3.41e-09	2.06e-06	4.37e-09	3.12e-09	4.73e-09	**1.95e-09**	9.98e-09	2.63e-09	1.28e-08	3.72e-09	5.95e-09
	336	3.40e-08	5.92e-08	6.41e-08	6.59e-08	1.51e-07	4.48e-08	3.28e-08	1.66e-07	5.52e-08	1.59e-07	4.33e-08	**2.10e-08**
	720	7.90e-07	4.19e-06	1.84e-04	5.32e-06	3.18e-05	2.54e-06	6.89e-07	6.12e-06	2.89e-05	2.91e-06	**4.33e-07**	1.33e-06
Linear-Trend	96	2.38e-07	2.46e-07	2.55e-07	4.84e-07	2.30e-07	4.75e-07	2.54e-07	4.15e-07	**2.14e-07**	1.11e-06	2.91e-07	2.57e-07
	336	2.39e-07	2.59e-07	2.27e-07	4.04e-07	2.32e-07	**2.09e-07**	2.27e-07	2.46e-07	2.17e-07	7.21e-07	2.40e-07	2.29e-07
	720	1.49e-06	1.52e-06	1.54e-06	1.50e-06	1.38e-06	2.13e-06	1.37e-06	1.67e-06	**1.20e-06**	2.37e-06	1.62e-06	1.45e-06
Logarithmic-Function	96	2.60e-08	2.91e-08	2.65e-06	3.73e-08	1.29e-08	7.53e-08	2.58e-08	3.68e-08	**9.59e-09**	1.87e-07	2.27e-08	3.20e-08
	336	1.48e-07	5.56e-07	2.66e-07	1.60e-07	2.54e-07	3.11e-07	**1.48e-07**	4.99e-07	2.15e-07	1.36e-06	3.29e-07	4.01e-07
	720	1.13e-05	2.96e-05	1.40e-04	2.36e-05	1.45e-04	1.75e-05	8.73e-06	1.87e-05	1.46e-04	2.34e-05	1.73e-05	**7.46e-06**
Logistic-Trend	96	3.75e-06	1.01e-05	5.37e-05	**2.89e-07**	1.08e-04	1.88e-06	9.75e-07	2.82e-05	6.24e-05	2.58e-06	3.62e-07	9.25e-07
	336	**4.28e-06**	2.42e-05	6.60e-05	9.44e-06	8.89e-05	1.02e-05	8.30e-06	1.71e-05	7.76e-05	1.18e-05	5.52e-06	9.58e-06
	720	6.29e-04	1.75e-03	1.37e-03	9.23e-04	1.50e-03	1.61e-03	**1.21e-04**	3.67e-03	1.64e-03	3.02e-04	6.49e-04	7.18e-04
Negative-Exponential-Trend	96	5.55e-09	5.81e-09	**4.88e-09**	1.29e-08	5.25e-09	6.96e-09	5.87e-09	1.03e-08	4.96e-09	1.41e-08	7.59e-09	5.29e-09
	336	5.48e-09	6.81e-09	5.66e-09	7.05e-09	5.40e-09	6.53e-09	5.80e-09	6.26e-09	**5.33e-09**	1.83e-08	6.42e-09	6.14e-09
	720	2.94e-08	2.80e-08	2.81e-08	4.18e-08	2.78e-08	3.66e-08	2.93e-08	2.99e-08	**2.77e-08**	1.52e-07	3.15e-08	3.09e-08
Piecewise-Linear	96	2.55e-07	**2.49e-07**	2.63e-07	1.01e-05	1.85e-06	2.72e-07	4.35e-07	1.78e-06	6.61e-07	3.32e-06	1.26e-06	4.10e-07
	336	**2.63e-07**	2.99e-07	5.06e-07	7.33e-07	2.70e-07	2.86e-07	2.71e-07	3.25e-07	4.02e-07	1.83e-06	2.77e-07	3.90e-07
	720	0.0134	0.0144	**2.38e-03**	0.0104	7.80e-03	0.0102	6.76e-03	9.04e-03	3.43e-03	0.0161	0.0193	3.38e-03
Power-Law-Trend	96	3.29e-07	1.56e-07	2.54e-06	3.23e-07	2.41e-06	5.05e-07	**9.06e-08**	5.96e-07	2.16e-06	6.43e-07	7.08e-07	1.87e-07
	336	5.75e-07	1.54e-06	7.78e-06	5.69e-07	6.44e-06	8.64e-07	**3.49e-07**	1.65e-06	6.12e-06	2.87e-06	2.53e-06	7.38e-07
	720	2.05e-05	4.28e-05	2.31e-04	2.08e-05	2.45e-04	2.90e-05	**7.51e-06**	5.97e-05	2.02e-04	4.25e-05	9.80e-06	3.04e-05
Quadratic-Function	96	3.14e-06	6.61e-06	4.27e-06	**2.95e-06**	3.09e-04	3.76e-06	3.19e-06	9.41e-06	3.95e-06	2.41e-04	5.19e-06	6.87e-06
	336	1.78e-05	1.50e-04	2.47e-04	1.57e-05	8.00e-05	1.84e-05	**1.24e-05**	1.91e-04	9.84e-05	1.46e-04	5.01e-05	3.17e-05
	720	6.47e-04	4.48e-03	0.0239	2.86e-03	0.0267	**5.02e-04**	1.19e-03	5.22e-03	0.0254	1.09e-03	1.48e-03	4.18e-03
Step-Function	96	6.69e-15	1.15e-15	**0**	7.08e-15	0	1.29e-14	4.20e-15	4.77e-15	0	8.73e-15	6.31e-15	4.20e-15
	336	4.06e-16	7.61e-16	**0**	1.32e-15	0	5.08e-17	7.61e-17	3.34e-15	0	2.32e-15	1.74e-15	1.63e-15
	720	1.13e-14	2.33e-15	7.89e-18	8.37e-15	3.95e-17	1.14e-16	3.66e-15	4.96e-15	**0**	1.20e-14	1.03e-14	5.91e-15

Table 7: Periodic: test MSE. For each dataset, model, and horizon, hyperparameters (context length, n_layers, d_model, lr, loss) are chosen by lowest seed-averaged val MSE. Reported values are seed-averaged test MSE (existing seeds only). Bold = best in row, underline = second-best. Horizons are not averaged.

Dataset	H	DeltaNet	DeltaProduct	GRU	LRU	LSTM	M2RNN	Mamba	Mamba2	RNN	S5	mLSTM	sLSTM
Sin(fre=0.005)	96	1.23e-05	7.48e-05	**1.05e-05**	4.43e-04	6.82e-05	7.25e-05	1.80e-05	7.87e-05	3.02e-05	3.25e-04	1.66e-04	1.03e-04
	336	1.42e-05	2.23e-05	**7.77e-06**	5.36e-04	1.77e-05	1.42e-05	1.13e-05	6.37e-05	3.77e-05	2.50e-04	1.50e-05	7.54e-05
	720	1.62e-05	2.80e-05	9.92e-05	1.85e-03	1.14e-04	4.92e-05	**1.16e-05**	6.90e-05	4.34e-04	2.62e-04	1.30e-05	9.83e-05
Sin(fre=0.05)	96	1.33e-05	4.70e-05	2.26e-05	2.50e-04	**9.73e-06**	3.75e-05	1.10e-05	5.24e-05	9.46e-05	2.14e-04	2.43e-05	3.32e-05
	336	1.40e-05	1.58e-05	1.61e-05	1.17e-04	1.52e-05	1.31e-05	**8.36e-06**	3.12e-05	2.08e-05	2.60e-04	1.67e-05	3.85e-05
	720	1.55e-05	1.63e-05	7.48e-05	1.47e-04	2.76e-05	1.45e-05	1.41e-05	1.61e-04	1.93e-05	2.45e-04	**1.07e-05**	3.27e-05
Double-Sin	96	8.93e-06	1.17e-05	**6.47e-06**	3.10e-04	2.65e-05	3.75e-05	2.75e-05	8.72e-05	8.22e-06	4.65e-04	3.07e-05	1.27e-04
	336	1.36e-05	2.28e-05	**6.13e-06**	2.50e-04	1.27e-05	1.48e-05	6.24e-06	5.11e-05	9.94e-06	5.09e-04	8.31e-06	2.65e-05
	720	1.38e-05	3.02e-05	9.91e-06	2.80e-04	1.27e-05	1.17e-05	**6.40e-06**	1.61e-04	1.68e-05	5.23e-04	9.22e-06	7.18e-05
Triple-Sin	96	1.18e-05	1.68e-05	4.20e-05	1.55e-04	**6.49e-06**	3.47e-05	1.59e-05	1.37e-03	1.24e-05	5.58e-04	1.78e-05	3.12e-05
	336	1.41e-05	1.57e-05	1.29e-05	1.67e-04	**6.11e-06**	1.19e-05	7.17e-06	1.27e-04	8.06e-06	3.72e-04	1.20e-05	3.57e-05
	720	1.84e-05	7.15e-05	1.26e-04	1.70e-04	3.36e-05	1.29e-05	**1.22e-05**	3.16e-05	1.29e-05	5.72e-04	1.73e-05	2.32e-04
Five-Sin	96	**1.39e-05**	8.32e-05	1.75e-05	1.10e-03	1.47e-04	7.16e-05	1.66e-05	4.11e-04	4.62e-05	6.90e-04	3.17e-05	7.42e-05
	336	1.65e-05	2.22e-04	2.28e-05	5.28e-04	2.70e-05	2.12e-05	1.87e-05	**1.33e-05**	3.08e-05	6.43e-04	1.67e-05	2.12e-04
	720	2.07e-05	3.75e-03	2.98e-05	5.05e-04	7.10e-04	2.29e-05	2.42e-05	4.27e-04	4.14e-04	6.23e-04	**1.82e-05**	7.27e-04
Ten-Sin	96	9.41e-05	1.24e-03	2.00e-04	5.42e-03	1.16e-05	2.36e-05	7.31e-05	1.26e-05	**1.15e-05**	6.65e-04	9.62e-05	2.84e-04
	336	1.09e-04	5.20e-04	1.44e-05	5.78e-04	**8.34e-06**	1.28e-05	1.13e-05	1.64e-05	1.06e-05	5.63e-04	3.56e-05	4.23e-04
	720	1.61e-04	**1.16e-05**	1.77e-03	6.03e-04	1.58e-05	1.95e-05	1.06e-04	7.85e-04	1.30e-03	6.93e-04	5.00e-05	6.54e-04
Exp-Sine-Wave	96	2.04e-05	2.35e-05	1.03e-04	2.93e-04	7.46e-06	2.38e-05	**6.34e-06**	2.62e-04	1.77e-05	2.63e-04	2.58e-05	1.72e-04
	336	1.31e-05	2.09e-05	5.23e-05	1.93e-04	1.98e-05	1.19e-05	**9.46e-06**	3.18e-05	2.41e-05	2.38e-04	1.08e-05	5.02e-05
	720	1.04e-04	6.59e-05	3.20e-05	2.42e-04	1.76e-05	7.09e-05	**1.60e-05**	5.73e-05	5.90e-05	2.90e-04	3.93e-05	1.01e-04
Triangle-Wave	96	1.16e-05	9.92e-06	1.12e-05	1.30e-04	9.06e-06	1.61e-05	4.22e-05	2.67e-04	6.60e-06	5.27e-04	1.24e-05	**5.85e-06**
	336	1.68e-05	9.76e-06	6.37e-06	2.06e-04	**5.77e-06**	9.07e-06	6.70e-06	1.33e-05	9.04e-06	6.22e-04	7.01e-06	8.19e-06
	720	4.34e-05	7.70e-05	1.14e-05	3.79e-04	6.85e-04	1.32e-05	1.06e-05	1.21e-05	7.21e-05	1.45e-03	1.56e-05	**9.71e-06**
Sawtooth-Wave	96	1.08e-05	7.59e-04	5.04e-06	4.03e-05	**4.51e-06**	1.49e-05	3.94e-05	2.23e-04	4.55e-06	3.68e-04	3.78e-05	3.35e-05
	336	9.21e-06	1.25e-05	**4.35e-06**	2.83e-05	4.35e-06	1.16e-05	6.81e-06	1.07e-05	4.54e-06	5.47e-04	7.92e-06	6.44e-06
	720	1.18e-05	1.07e-05	4.98e-06	3.36e-04	4.70e-06	3.19e-05	5.89e-05	1.11e-05	**4.60e-06**	2.30e-04	2.06e-05	1.63e-05
Square-Wave	96	0.1525	0.1110	0.1037	**0.0607**	0.1090	0.1212	0.1101	0.1279	0.1038	0.0614	0.1078	0.1116
	336	0.1340	0.1147	0.1027	0.1069	0.1022	0.1076	0.1213	0.1404	**0.1022**	0.1050	0.1161	0.1072
	720	0.1535	0.1105	0.1018	0.1056	0.1050	0.1121	0.1164	0.1447	**0.1016**	0.1039	0.1116	0.1077

Table 8: Noise: test MSE. For each dataset, model, and horizon, hyperparameters (context length, n_layers, d_model, lr, loss) are chosen by lowest seed-averaged val MSE. Reported values are seed-averaged test MSE (existing seeds only). Bold = best in row, underline = second-best. Horizons are not averaged.

Dataset	H	DeltaNet	DeltaProduct	GRU	LRU	LSTM	M2RNN	Mamba	Mamba2	RNN	S5	mLSTM	sLSTM
Exponential-Trend SNR=-10dB	96	**1.0515**	1.0527	1.0519	1.0532	1.0519	1.0523	1.0528	1.0525	1.0516	1.0520	1.0525	1.0526
	336	1.0949	1.0959	1.0950	1.0960	1.0949	**1.0946**	1.0948	1.0953	1.0951	1.0959	1.0958	1.0958
	720	**1.2785**	1.2797	1.2799	1.2800	1.2798	1.2795	1.2788	1.2792	1.2798	1.2791	1.2795	1.2798
Exponential-Trend SNR=0dB	96	**1.6670**	1.6830	1.6912	1.6822	1.6816	1.6861	1.6804	1.6871	1.6802	1.6809	1.6768	1.6818
	336	**4.6080**	4.7538	4.7134	4.7370	4.7346	4.7353	4.7458	4.7288	4.7348	4.7504	4.7393	4.7452
	720	**3.6742**	3.7336	3.7432	3.7317	3.7487	3.7146	3.7348	3.7331	3.7400	3.7296	3.7255	3.7302
Exponential-Trend SNR=10dB	96	**6.0840**	6.6307	6.3380	6.5442	6.4260	6.4944	6.5793	6.4589	6.5203	6.5399	6.4648	6.5185
	336	6.4350	6.4596	6.4290	6.4306	**6.3812**	6.4209	6.4397	6.4378	6.4534	6.4365	6.4178	6.4386
	720	28.9697	28.9994	**28.2324**	29.0571	28.9948	28.9993	29.0655	28.9655	29.0441	28.9955	28.9032	28.9934
Five-Sin SNR=-10dB	96	0.9728	0.9743	0.9421	0.9351	0.9360	0.9767	0.9742	0.9750	0.9311	**0.9088**	0.9737	0.9754
	336	0.9109	0.9112	0.9080	0.8767	0.8858	0.9099	0.9094	0.9111	0.8786	**0.8556**	0.9102	0.9107
	720	0.9894	0.9833	0.9430	0.9321	0.9536	0.9603	0.9816	0.9829	0.9532	**0.9151**	0.9832	0.9811
Five-Sin SNR=0dB	96	0.6254	0.6497	0.5140	0.5284	**0.5097**	0.5337	0.5988	0.6332	0.5181	0.5177	0.6491	0.5398
	336	0.5962	0.5913	0.5254	0.5242	0.5091	0.5119	0.5670	0.6153	0.5252	**0.5063**	0.5814	0.5258
	720	0.5729	0.6457	0.5163	0.5250	**0.5068**	0.5258	0.5737	0.6345	0.5148	0.5140	0.6542	0.5260
Five-Sin SNR=10dB	96	0.1084	0.1172	0.0983	0.0940	**0.0933**	0.0999	0.1049	0.1181	0.0947	0.0939	0.1099	0.1010
	336	0.1059	0.1137	0.0985	0.0953	0.0964	0.0981	0.1064	0.1147	0.0968	**0.0952**	0.1063	0.1022
	720	0.1029	0.1042	0.0965	0.0939	**0.0933**	0.0951	0.1019	0.1089	0.0950	0.0936	0.0995	0.0992

Table 9: Anomaly: test MSE. For each dataset, model, and horizon, hyperparameters (context length, n_layers, d_model, lr, loss) are chosen by lowest seed-averaged val MSE. Reported values are seed-averaged test MSE (existing seeds only). Bold = best in row, underline = second-best. Horizons are not averaged.

Dataset	H	DeltaNet	DeltaProduct	GRU	LRU	LSTM	M2RNN	Mamba	Mamba2	RNN	S5	mLSTM	sLSTM
Quadratic + 0.5% point anomalies	96	0.1002	0.0914	0.0848	0.0870	0.0845	0.0869	0.0883	0.0888	0.1118	0.0790	**0.0780**	0.0830
	336	0.1419	0.1654	0.0980	0.1052	0.1103	0.1266	0.1298	0.1314	0.1764	**0.0969**	0.1181	0.1119
	720	0.3699	0.4590	0.2667	0.2665	0.2723	0.3157	0.3690	0.3249	0.4377	**0.2525**	0.3003	0.3012
Quadratic + 5% point anomalies	96	0.1970	0.1703	0.1313	0.1329	0.1506	0.1496	0.1375	0.1399	0.1579	**0.1096**	0.1263	0.1276
	336	0.4301	0.4883	0.4145	0.3807	0.3516	0.4255	0.3978	0.3500	0.4596	**0.2899**	0.3134	0.3145
	720	1.0632	0.9836	0.7472	0.7863	0.8021	0.8205	0.7900	0.8689	0.9291	**0.5839**	0.8081	0.6943
Quadratic + 20% point anomalies	96	0.1363	0.1276	0.0945	0.0881	0.1080	0.1105	0.1060	0.0986	0.1031	**0.0831**	0.1065	0.0959
	336	0.5333	**0.3109**	0.3919	0.3878	0.4317	0.4390	0.4035	0.3986	0.4585	0.3167	0.4012	0.3786
	720	0.7527	0.7332	0.6832	**0.5333**	0.5740	0.6054	0.5718	0.6036	0.5984	0.5789	0.5578	0.5439

Table 10: Dependency: test MSE. For each dataset, model, and horizon, hyperparameters (context length, n_layers, d_model, lr, loss) are chosen by lowest seed-averaged val MSE. Reported values are seed-averaged test MSE (existing seeds only). Bold = best in row, underline = second-best. Horizons are not averaged.

Dataset	H	DeltaNet	DeltaProduct	GRU	LRU	LSTM	M2RNN	Mamba	Mamba2	RNN	S5	mLSTM	sLSTM
ARMA(1,1)	96	0.8747	**0.8555**	0.9266	0.8726	0.8775	0.8860	0.8722	0.8598	0.8683	0.8956	0.8789	0.8765
	336	**0.8633**	1.0167	1.0740	1.0115	1.0320	1.0114	1.0213	1.0110	1.0259	1.0384	1.0176	1.0163
	720	1.0437	0.9641	0.9427	0.9506	0.9364	0.9312	0.9344	0.9720	0.9430	0.9632	**0.8789**	0.9691
ARMA(2,2)	96	1.0331	1.0290	1.0274	1.0289	1.0298	1.0318	1.0273	1.0287	**1.0270**	1.0297	1.0289	1.0274
	336	1.0611	1.0614	1.0618	1.0617	1.0619	1.0624	1.0616	1.0615	**1.0608**	1.0613	1.0621	1.0615
	720	1.0753	1.0732	1.0688	1.0737	**1.0671**	1.0767	1.0732	1.0739	1.0692	1.0729	1.0727	1.0748
ARMA-Long-Dependency	96	0.9749	0.9889	0.9668	0.8415	0.9847	0.9759	0.9761	0.9650	0.9000	**0.7799**	0.9824	0.9916
	336	0.9754	0.9780	0.9803	0.9717	0.9752	0.9744	0.9725	0.9722	0.9922	0.9756	**0.9678**	0.9828
	720	1.0167	1.0191	1.0070	1.0237	1.0091	1.0171	1.0140	1.0164	**1.0059**	1.0249	1.0088	1.0124
Random-Walk	96	0.1038	0.0896	0.0815	0.0880	0.0970	0.1672	0.1039	0.0947	0.0824	**0.0790**	0.1048	0.0819
	336	**0.4038**	0.4045	0.9926	0.4144	0.9937	0.5324	0.4192	0.4121	0.9836	0.4161	0.4151	0.4094
	720	0.7789	0.7806	0.7389	0.7845	0.7700	0.8070	0.7845	0.7868	**0.7370**	0.7698	0.7892	0.7903
White-Noise	96	0.9765	0.9763	0.9748	0.9761	**0.9747**	0.9769	0.9769	0.9768	0.9751	0.9755	0.9761	0.9765
	336	0.9588	0.9581	0.9572	0.9585	**0.9571**	0.9580	0.9582	0.9584	0.9576	0.9579	0.9582	0.9591
	720	0.9716	0.9719	0.9706	0.9721	**0.9704**	0.9710	0.9708	0.9717	0.9708	0.9714	0.9710	0.9720

Table 11: Cross-variable: test MSE. For each dataset, model, and horizon, hyperparameters (context length, n_layers, d_model, lr, loss) are chosen by lowest seed-averaged val MSE. Reported values are seed-averaged test MSE (existing seeds only). Bold = best in row, underline = second-best. Horizons are not averaged.

Dataset	H	DeltaNet	DeltaProduct	GRU	LRU	LSTM	M2RNN	Mamba	Mamba2	RNN	S5	mLSTM	sLSTM
Feature Lag (5 steps)	96	1.0568	1.0573	1.0510	1.0536	1.0577	1.0542	1.0545	1.0553	**1.0475**	1.0647	1.0537	1.0573
	336	1.0415	1.0397	1.0378	1.0373	1.0377	1.0384	1.0391	1.0393	**1.0356**	1.0401	1.0397	1.0398
	720	1.0717	1.0722	1.0711	1.0721	1.0709	1.0719	1.0717	1.0721	**1.0708**	1.0730	1.0718	1.0739
Feature Lag (10 steps)	96	0.9281	0.9493	0.9418	0.9262	0.9462	0.9306	0.9409	0.9294	**0.9193**	0.9347	0.9298	0.9276
	336	0.9596	0.9625	0.9607	0.9616	0.9610	0.9629	0.9625	0.9620	**0.9564**	0.9629	0.9625	0.9621
	720	0.9763	0.9750	0.9746	0.9752	0.9745	0.9751	0.9743	0.9749	**0.9720**	0.9756	0.9752	0.9755
Feature Lag (24 steps)	96	0.9543	0.9776	0.9701	0.9012	0.9752	0.9582	0.9610	0.9549	**0.8888**	0.9136	0.9592	0.9684
	336	1.0152	1.0141	1.0127	1.0001	1.0128	1.0128	1.0132	1.0127	**0.9931**	1.0010	1.0136	1.0139
	720	0.9905	0.9903	0.9895	0.9892	0.9894	0.9906	0.9895	0.9902	**0.9859**	0.9914	0.9901	0.9904
Feature Lag (48 steps)	96	0.8883	0.9121	0.8931	**0.7442**	0.9004	0.8940	0.8974	0.8907	0.7549	0.7526	0.8863	0.8955
	336	0.8945	0.8921	0.8895	0.8811	0.8893	0.8930	0.8910	0.8919	**0.8763**	0.9030	0.8917	0.8933
	720	0.9389	0.9374	0.9385	0.9365	0.9350	0.9370	0.9368	0.9374	**0.9343**	0.9379	0.9372	0.9386
Sine-Noise Composition	96	0.4692	0.4673	0.4674	0.4673	0.4672	0.4673	0.4671	0.4691	0.4674	0.4675	**0.4670**	0.4678
	336	0.4722	0.4723	0.4716	0.4722	**0.4715**	0.4720	0.4722	0.4721	0.4716	0.4726	0.4721	0.4726
	720	0.4696	0.4692	0.4688	0.4697	**0.4686**	0.4691	0.4693	0.4695	0.4687	0.4700	0.4691	0.4702
Conditional Relationship	96	0.9866	1.0009	0.9890	0.9397	0.9884	0.9875	0.9993	0.9881	0.9398	**0.9395**	0.9938	0.9977
	336	1.0071	1.0032	**0.9995**	1.0051	1.0123	1.0029	1.0028	1.0033	1.0052	1.0027	1.0027	1.0030
	720	1.0049	1.0046	1.0013	1.0051	**0.9968**	1.0040	1.0062	1.0069	1.0056	1.0043	1.0054	1.0062
Nonlinear Relationship	96	0.9134	0.9278	0.9167	0.7807	0.9226	0.9138	0.9217	0.9109	0.7845	**0.7676**	0.9122	0.9180
	336	0.9654	0.9641	0.9614	0.9387	0.9644	0.9639	0.9632	0.9642	0.9465	**0.9339**	0.9646	0.9654
	720	0.9450	0.9458	**0.9438**	0.9465	0.9444	0.9442	0.9441	0.9447	0.9443	0.9454	0.9444	0.9455
Multivariable Complex	96	0.9031	0.9307	0.9355	0.8917	0.9552	0.9135	0.9234	0.9089	**0.8772**	0.8937	0.9145	0.9297
	336	0.9754	0.9807	0.9769	0.9702	0.9805	0.9748	0.9810	0.9747	**0.9687**	0.9809	0.9757	0.9875
	720	0.9806	0.9780	0.9771	**0.9762**	0.9773	0.9776	0.9775	0.9793	0.9769	0.9783	0.9803	0.9793

Table 12: Univariate real-world simulation: test MSE. For each dataset, model, and horizon, hyperparameters (context length, n_layers, d_model, lr, loss) are chosen by lowest seed-averaged val MSE. Reported values are seed-averaged test MSE (existing seeds only). Bold = best in row, underline = second-best. Horizons are not averaged.

Dataset	H	DeltaNet	DeltaProduct	GRU	LRU	LSTM	M2RNN	Mamba	Mamba2	RNN	S5	mLSTM	sLSTM
Economic-Indicator	96	0.0981	0.0959	0.0794	0.0583	0.0643	0.0633	0.0752	0.0723	0.0875	**0.0466**	0.0868	0.0714
	336	0.2212	0.1652	0.2650	**0.1395**	0.2484	0.1924	0.2266	0.2408	0.2198	0.2281	0.2300	0.2260
	720	1.3384	1.6493	0.5488	0.5774	0.4788	1.3798	0.6305	0.9763	**0.3265**	0.6924	0.6791	0.4918
Electricity-Consumption	96	0.0859	0.0657	0.0618	0.0700	0.0482	0.0486	0.0542	0.0775	0.1103	**0.0476**	0.0539	0.0564
	336	0.1118	0.0656	0.0610	0.0582	**0.0458**	0.0490	0.0625	0.0811	0.1370	0.0479	0.0661	0.0579
	720	0.2166	0.1695	0.1814	0.1447	0.1474	0.0733	0.0815	0.1640	0.2647	**0.0665**	0.1352	0.1967
Industrial-Sensor	96	0.0281	0.0354	0.0289	0.0298	**0.0249**	0.0659	0.0295	0.0434	0.0297	0.0297	0.0293	0.0298
	336	0.0253	0.0287	**0.0240**	0.0276	0.0370	0.0493	0.0275	0.0414	0.0618	0.0271	0.0303	0.0322
	720	0.9716	1.3568	1.1462	0.3277	1.1934	0.9479	0.8383	0.6557	0.9924	**0.2553**	1.2977	0.6768
IoT-Sensor-Battery	96	2.44e-05	3.12e-05	2.63e-05	2.58e-05	1.79e-04	2.71e-05	2.62e-05	3.18e-05	2.79e-05	**2.31e-05**	2.75e-05	2.69e-05
	336	2.81e-05	2.94e-05	4.11e-05	2.62e-05	4.20e-04	4.02e-05	2.78e-05	3.66e-05	6.62e-05	**2.60e-05**	3.28e-05	3.54e-05
	720	6.11e-05	8.34e-05	3.29e-03	6.57e-05	2.78e-03	2.35e-04	**5.07e-05**	1.62e-04	2.82e-03	9.31e-05	5.18e-05	1.38e-04
Network-Traffic	96	0.0393	0.0548	0.0291	0.0299	0.0248	0.0277	0.0492	0.0499	0.0757	**0.0219**	0.0350	0.0275
	336	0.0390	0.0334	0.0294	0.0686	0.0281	0.0280	0.0277	0.0407	0.0412	**0.0271**	0.0357	0.0916
	720	0.0343	0.0345	0.0258	0.0362	**0.0219**	0.0259	0.0277	0.0395	0.0694	0.0300	0.0283	0.0821
Retail-Sales	96	0.5051	0.5486	0.5218	0.4665	0.4810	0.4945	0.5775	0.5648	0.4640	**0.3623**	0.6344	0.5795
	336	0.5767	0.5756	0.4907	0.5241	0.4804	0.5080	0.5711	0.6044	0.5752	**0.3877**	0.5878	0.5705
	720	0.5197	0.5126	0.4617	0.4716	0.5015	0.5246	0.5084	0.5208	0.4841	**0.4464**	0.5464	0.5481
Stock-Price	96	0.4659	0.5569	0.5592	0.4906	0.4924	0.4486	0.5469	0.5510	0.5010	**0.4005**	0.6269	0.5004
	336	1.0444	0.9936	1.0099	0.9641	0.9219	1.0272	1.0252	0.9766	**0.9205**	0.9289	1.0920	0.9591
	720	1.7841	1.7416	1.7008	2.7455	**1.4196**	1.5415	1.6797	1.8039	1.4450	1.6609	1.5526	1.8377
Temperature-Sensor	96	0.0622	0.0604	0.0560	0.0536	0.0560	0.0555	0.0580	0.0607	0.0572	**0.0533**	0.0571	0.0570
	336	0.0700	0.0665	0.0644	0.0688	**0.0580**	0.0646	0.0669	0.0749	0.0675	0.0924	0.0700	0.0642
	720	0.1237	0.1262	0.0804	0.0939	0.1527	0.1186	0.0928	0.1226	**0.0711**	0.0838	0.1084	0.0975
Website-Traffic	96	0.0653	0.0479	0.0497	0.0564	0.0560	0.0474	0.0502	0.0821	0.2147	**0.0341**	0.0527	0.0530
	336	0.0408	0.0457	0.0552	0.0477	0.0449	0.0441	0.0633	0.0508	0.0671	**0.0371**	0.0550	0.0535
	720	0.0547	0.0543	0.0814	0.1474	0.0402	0.0395	0.0738	0.1163	0.1537	**0.0337**	0.0556	0.0987

Table 13: Multivariate real-world simulation: test MSE. For each dataset, model, and horizon, hyperparameters (context length, n_layers, d_model, lr, loss) are chosen by lowest seed-averaged val MSE. Reported values are seed-averaged test MSE (existing seeds only). Bold = best in row, underline = second-best. Horizons are not averaged.

Dataset	H	DeltaNet	DeltaProduct	GRU	LRU	LSTM	M2RNN	Mamba	Mamba2	RNN	S5	mLSTM	sLSTM
Ad-Sales	96	0.6049	0.6028	0.6114	0.5999	0.5986	0.5998	**0.5962**	0.5984	0.6141	0.5992	0.6044	0.6015
	336	0.6122	0.6105	0.6014	0.6026	**0.6013**	0.6045	0.6056	0.6069	0.6934	0.6026	0.6075	0.6047
	720	0.6281	0.6145	0.6010	0.6100	**0.5928**	0.5984	0.6118	0.6245	0.6066	0.6037	0.6022	0.6182
Intervention-Effect	96	0.0907	0.0927	0.0884	0.0875	0.0953	0.0874	0.0884	0.0883	0.0919	**0.0870**	0.0885	0.0888
	336	0.0936	0.0935	0.0913	0.0932	0.1107	0.0968	0.0992	0.0958	0.0946	**0.0894**	0.0963	0.0951
	720	0.1135	0.1055	0.1468	0.1133	0.1253	0.0992	0.1576	0.1070	0.1629	**0.0940**	0.1837	0.2325
Lotka-Volterra	96	6.03e-03	6.13e-03	5.16e-03	5.46e-03	7.56e-03	5.23e-03	5.28e-03	5.45e-03	7.30e-03	7.31e-03	**4.24e-03**	5.92e-03
	336	4.56e-03	4.53e-03	4.05e-03	5.90e-03	4.81e-03	4.35e-03	4.10e-03	5.02e-03	7.16e-03	5.79e-03	**3.90e-03**	4.38e-03
	720	4.54e-03	6.28e-03	6.85e-03	6.06e-03	6.25e-03	4.95e-03	**4.53e-03**	5.92e-03	5.82e-03	8.02e-03	4.76e-03	6.94e-03
Macro-Economy	96	0.7759	0.7648	0.7510	0.7536	**0.7493**	0.7636	0.7546	0.7564	0.7552	0.7602	0.7728	0.7572
	336	0.8041	0.7798	0.7734	**0.7641**	0.7748	0.8166	0.7705	0.7809	0.7802	0.7763	0.8485	0.7705
	720	0.8005	0.8619	0.7772	**0.7725**	0.7770	0.8432	0.7765	0.7830	0.7780	0.7799	1.0322	0.7745
SIR-Model	96	0.0843	0.0827	0.0711	0.0873	0.0660	0.0791	0.0799	0.0820	0.0732	0.0800	0.0815	**0.0609**
	336	0.1058	0.0963	0.1095	0.1069	0.1093	0.1010	0.1168	0.1100	0.1988	**0.0957**	0.1068	0.1084
	720	0.5716	0.5072	0.6085	0.6213	0.5872	0.6241	0.6276	0.5945	0.5431	**0.4693**	0.5636	0.5362
Supply-Demand-Price	96	0.2366	0.2594	0.2260	0.2318	0.5908	0.3598	0.2552	0.2909	**0.2138**	0.2307	0.3218	0.2604
	336	0.7092	**0.7014**	0.7263	0.7234	0.7199	0.7252	0.7161	0.7278	0.7240	0.7228	0.7261	0.7146
	720	1.7523	1.7425	1.2302	1.6266	1.5224	1.7487	1.3048	1.7508	1.7486	**1.1717**	1.5587	1.7529
Weather-Sales	96	0.4421	0.4489	0.4321	0.4420	0.4325	0.4325	0.4317	0.4329	0.4352	**0.4316**	0.4321	0.4331
	336	0.4425	0.4387	0.4319	0.4487	**0.4316**	0.4335	0.4328	0.4335	0.4342	0.4332	0.4340	0.4353
	720	0.4384	0.4370	0.4297	0.4331	0.4305	0.4300	**0.4289**	0.4326	0.4338	0.4303	0.4295	0.4301

C LTSF RESULTS

Table 14: LTSF: test MSE. For each dataset, model, and horizon, hyperparameters (context length, n_layers, loss) are chosen by lowest seed-averaged val MSE. Reported values are seed-averaged test MSE (existing seeds only). Bold = best in row, underline = second-best. Horizons are not averaged.

Dataset	H	RNN	GRU	LSTM	S5	LRU	Griffin	Mamba	Mamba2	mLSTM	sLSTM	DeltaNet	DeltaProduct	M2RNN	PatchTST	DLinear
ETTh1	96	0.4656	0.4786	0.5031	0.5098	**0.4429**	0.5276	0.4840	0.5064	0.5139	0.5515	0.5550	0.5309	0.5082	0.390	0.384
	192	0.5425	0.5427	0.5353	0.5771	**0.5221**	0.6012	0.5587	0.5644	0.5554	0.6202	0.6050	0.5763	0.5961		
	336	0.6329	0.6331	0.7314	0.6587	**0.6277**	0.6364	0.7015	0.6832	0.6676	0.6308	0.7064	0.6963	0.7396		
	720	**0.6014**	0.6492	0.7090	0.6286	0.6297	0.6983	0.6563	0.6602	0.6412	0.7150	0.6918	0.6665	0.6954		
ETTh2	96	0.6577	0.3810	1.6726	0.4777	0.4869	0.4773	0.9721	0.6492	0.3689	**0.3661**	0.5537	0.5086	0.8821	0.505	0.333
	192	0.4985	0.5341	0.4573	0.5185	0.6362	0.4974	0.7331	0.5588	**0.4199**	0.4506	0.5200	0.6344	0.5608		
	336	0.4807	0.4553	0.4683	0.4986	0.5226	0.5120	0.5082	0.4989	0.4556	**0.4530**	0.5082	0.4838	0.5063		
	720	0.5023	**0.4865**	0.6202	0.5335	0.5517	0.5360	0.5442	0.5561	0.5117	0.5031	0.7132	0.5360	0.6630		
ETTm1	96	0.4205	0.3913	0.4289	**0.3588**	0.3854	0.4422	0.4160	0.4583	0.4225	0.4088	0.5111	0.4814	0.4216	0.317	0.315
	192	0.4539	0.4409	0.4679	**0.4255**	0.4588	0.4974	0.4787	0.4973	0.4609	0.4554	0.5599	0.5144	0.4789		
	336	0.4711	0.4792	0.4836	**0.4431**	0.4842	0.5901	0.5984	0.6532	0.4942	0.4758	0.6771	0.6568	0.5086		
	720	0.5323	0.5979	**0.5084**	0.5292	0.5399	0.5851	0.6168	0.6366	0.5642	0.5984	0.6429	0.5993	0.6052		
ETTm2	96	0.2828	0.2494	2.5371	0.2452	0.2533	0.7380	0.2639	0.2580	**0.2058**	0.2201	0.2399	0.2553	0.3022	0.168	0.169
	192	0.4967	0.3973	4.0682	0.3415	0.9255	0.3440	0.6705	0.4240	**0.2701**	0.2974	0.3309	0.3502	0.7610		
	336	0.3294	0.3209	0.3542	0.3973	0.3991	1.0769	0.9282	0.4049	0.3133	**0.3109**	1.9368	0.3922	0.3898		
	720	**0.3960**	0.4152	0.4253	0.4762	0.4759	0.4722	1.2411	0.9053	0.4477	0.4042	0.6718	0.4683	2.7066		
Weather	96	**0.1667**	0.1870	0.1706	0.2801	0.2887	0.2970	0.2916	0.3009	0.1743	0.1674	0.2768	0.2702	0.3024	3.772	0.168
	192	0.2878	0.2470	0.2821	0.2896	0.3014	0.3028	0.2959	0.3169	0.2248	**0.2173**	0.3032	0.2820	0.3755		
	336	0.2678	0.3064	0.2852	0.3285	0.3278	0.3278	0.3291	0.3429	0.2784	**0.2583**	0.3292	0.3198	0.3435		
	720	0.3670	0.3583	**0.3395**	0.3953	0.4034	0.4015	0.3947	0.4037	0.3546	0.3450	0.3981	0.4489	0.4096		
Traffic	96	0.6561	0.6374	**0.6150**	0.7676	0.6742	0.8418	0.6341	0.6485	0.6366	0.6178	0.6499	0.7363	0.6247	0.535	0.433
	192	0.7354	**0.6467**	1.0225	0.7592	0.7013	0.7358	0.7046	0.6934	0.6601	0.6514	0.7357	0.6815	0.6961		
	336	0.6634	**0.6523**	0.6571	0.6868	0.7452	0.7901	0.7365	0.7518	0.6807	0.6570	0.6791	0.7486	0.7837		
	720	0.8779	1.2173	1.4862	1.0011	1.2168	0.7499	0.7407	0.7439	**0.7203**	0.7233	0.7699	0.7677	0.9341		
Electricity	96	0.2549	0.1794	0.2775	0.4350	0.2855	0.2698	0.2228	0.2312	**0.1792**	0.1805	0.2605	0.2478	0.2206		
	192	0.1970	0.1937	0.6151	0.5436	0.4654	0.3600	0.3258	0.2823	**0.1932**	0.1935	0.3667	0.2920	0.3032	0.172	0.145
	336	0.1927	0.2011	**0.1894**	0.4861	0.4385	0.3975	0.3018	0.3055	0.1953	0.1928	0.3262	0.2995	0.2891		
	720	0.8786	0.5046	0.7728	0.7779	0.8743	0.6743	0.7259	0.6754	**0.2242**	0.2283	0.7010	0.6580	0.6848		
Exchange	96	0.1557	0.2119	0.1824	**0.1312**	0.1502	0.1546	0.1754	0.1774	0.1840	0.1787	0.1520	0.1564	0.1834	0.097	0.103
	192	0.2952	0.4053	0.3970	**0.2791**	0.3007	0.2874	0.3441	0.3265	0.3116	0.3548	0.3154	0.3365	0.3336		
	336	0.7865	0.4258	0.4192	0.6824	1.6750	2.8873	0.7072	0.8622	0.4638	**0.4036**	0.5036	1.3561	0.6990		
Illness	24	1.3660	1.2742	1.4996	1.6638	1.8401	1.4704	2.2290	1.6979	**0.8529**	1.0527	1.2937	1.7594	1.0404		
	36	1.3306	1.4767	1.5113	1.6688	1.6463	1.4244	1.8827	1.3530	1.0489	**1.0054**	1.2128	1.7577	1.0606		

D GIFT-EVAL RESULTS

E GIFT-EVAL PER-HORIZON RESULTS

The tables below report seed-averaged test MSE for every GIFT-Eval term (short, medium, long) on the datasets used in the main text. Hyperparameters are selected by lowest seed-averaged validation MSE, the same protocol as Table 5. Bold is best in the row; underline is second-best.

E.1 TEST MSE

Table 15: GIFT-Eval Energy: per-term test MSE (short, medium, long). For each dataset, model, and term, hyperparameters (context length, n layers, d_{model}, learning rate, loss) are selected by lowest seed-averaged validation MSE. Values are seed-averaged (existing seeds only). Bold: best in row; underline: second-best.

Dataset	Term	H	RNN	GRU	LSTM	S5	LRU	Griffin	Mamba	Mamba2	mLSTM	sLSTM	DeltaNet	DeltaProduct	M2RNN
ETT1/H	short	48	**0.2873**	0.3519	0.3732	0.3935	0.3218	0.4881	0.4783	0.4397	0.4847	0.4199	0.4833	0.3674	0.3891
	medium	480	1.1111	**0.7469**	0.9535	0.9227	0.8816	0.8132	0.8216	0.9121	0.8935	0.8879	0.8293	0.7576	1.0064
	long	720	0.8973	0.7345	0.8425	**0.6781**	0.7503	0.7725	0.7510	0.8477	0.8308	0.8021	0.7410	0.7315	0.7800
ETT1/15T	short	48	0.4504	0.2639	0.3284	**0.2448**	0.3405	0.2687	0.2882	0.3334	0.3100	0.3386	0.3176	0.2896	0.3529
	medium	480	0.6217	0.5573	0.5707	**0.4514**	0.4811	0.6822	0.6438	0.7418	0.7360	0.6830	0.6373	0.5977	0.5175
	long	720	0.6610	0.9302	0.5772	**0.5300**	0.6314	0.6802	0.6173	0.7933	0.7206	0.7281	0.6860	0.7334	0.6012
ETT2/H	short	48	**0.0835**	0.1052	0.0976	0.1077	0.1047	0.1049	0.0997	0.1120	0.1079	0.1056	0.1010	0.1009	0.0984
	medium	480	0.2380	0.2628	0.2848	0.2445	0.2350	0.2488	0.2254	0.2368	0.2551	**0.2083**	0.2314	0.2174	0.2353
	long	720	0.2583	**0.2553**	0.3177	0.2598	0.3779	0.3647	0.2705	0.8295	0.6655	0.8158	0.4332	0.7415	0.3519
ETT2/15T	short	48	0.1275	0.1204	0.1115	**0.0958**	0.1082	0.1030	0.1121	0.1288	0.1314	0.1241	0.1268	0.1315	0.1242
	medium	480	**0.1408**	0.1763	0.1776	0.1843	0.1894	0.1821	0.1676	0.1704	0.1605	0.1813	0.1688	0.1659	0.1689
	long	720	**0.1257**	0.2077	0.1440	0.1615	0.1692	0.1706	0.1603	0.2180	0.2026	0.1983	0.1868	0.2209	0.1705
Electricity/H	short	48	**9.88e-03**	0.0119	0.0104	0.0101	0.0100	0.0101	0.0105	0.0111	0.0109	0.0109	0.0107	0.0102	0.0112
	medium	480	**0.0390**	0.0523	0.0423	0.0465	0.0433	0.0561	0.0538	0.0522	0.0394	0.0406	0.0471	0.0472	0.0421
	long	720	0.0789	0.2617	**0.0688**	0.0768	0.1057	0.0864	0.0994	0.0822	0.0883	0.0971	0.0911	0.1640	0.1088
Solar/H	short	48	**0.1951**	0.2510	0.2544	0.2493	0.2274	0.2593	0.2174	0.2600	0.2751	0.2249	0.2286	0.2223	0.2269
	medium	480	0.3333	0.3269	0.3104	**0.2237**	0.2282	0.2813	0.2980	0.2767	0.2878	0.3046	0.3204	0.3214	0.2571
	long	720	0.3079	0.5793	0.3319	0.3025	**0.2688**	0.2933	0.2984	0.3076	0.3245	0.3462	0.3079	0.3346	0.2818

Table 16: GIFT-Eval Healthcare: per-term test MSE (short, medium, long). For each dataset, model, and term, hyperparameters (context length, n layers, d_{model}, learning rate, loss) are selected by lowest seed-averaged validation MSE. Values are seed-averaged (existing seeds only). Bold: best in row; underline: second-best.

Dataset	Term	H	RNN	GRU	LSTM	S5	LRU	Griffin	Mamba	Mamba2	mLSTM	sLSTM	DeltaNet	DeltaProduct	M2RNN	DLinear	PatchTST
US-Births/D	short	30	0.2153	0.1764	0.1732	**0.1084**	0.1755	0.1897	0.2149	0.2330	0.2139	0.2473	0.1628	0.1463	0.2289	0.1912	0.1502
	medium	300	0.2715	0.1945	0.2038	0.1991	0.2256	0.2151	0.1969	0.2498	0.2314	0.2128	0.2087	0.1887	0.2452	0.1632	**0.1181**
	long	450	0.3328	0.2672	0.2793	0.2567	0.2527	0.2984	0.2783	0.2289	0.3004	0.2686	0.2275	0.2191	0.2788	0.1785	**0.1771**
US-Births/W	short	8	0.1869	0.2413	0.1515	**0.0920**	0.1453	0.1884	0.1690	0.1983	0.1773	0.2114	0.1578	0.1504	0.1578	0.0995	0.1017
	medium	80	0.3029	0.2199	0.2516	0.3517	0.2794	0.2867	0.3735	0.2740	0.3153	0.4735	0.1818	0.2747	0.2129	0.1609	**0.1438**
	long	120	0.6204	0.6528	0.5922	0.6076	0.5652	0.6244	0.8258	0.6466	0.6608	0.6856	0.9737	0.6333	0.7013	**0.1845**	0.1904

Table 17: GIFT-Eval Nature: per-term test MSE (short, medium, long). For each dataset, model, and term, hyperparameters (context length, n layers, d_{model}, learning rate, loss) are selected by lowest seed-averaged validation MSE. Values are seed-averaged (existing seeds only). Bold: best in row; underline: second-best.

Dataset	Term	H	RNN	GRU	LSTM	S5	LRU	Griffin	Mamba	Mamba2	mLSTM	sLSTM	DeltaNet	DeltaProduct	M2RNN	DLinear	PatchTST
Jena-Weather/H	short	48	0.2802	0.2859	0.3210	0.2514	0.2457	0.3719	0.3426	0.3083	0.3151	0.3045	0.3053	0.2921	0.3312	**0.2117**	0.2334
	medium	480	0.4409	0.4739	0.5353	**0.4042**	0.5006	0.5337	0.5503	0.5613	0.4740	0.4468	0.7291	0.5600	0.6281	0.4281	0.5466
	long	720	0.6804	0.6105	0.6694	0.6522	0.6331	0.6889	0.6874	0.5968	0.6239	0.6034	0.7779	0.6102	0.7275	**0.5501**	0.5785
Saugeen/D	short	30	0.8523	0.8603	0.9196	**0.5947**	0.7170	0.6426	0.8359	0.7660	0.7269	0.7893	0.6070	0.8265	0.7112	0.6008	0.7300
	medium	300	1.3223	1.1645	1.1700	1.0904	1.1873	1.2715	1.3359	1.1963	1.1381	1.1753	1.5760	1.5190	1.1068	**1.0367**	1.1854
	long	450	1.3321	1.2835	1.0801	0.9820	1.0780	1.0105	1.1858	1.2008	1.1347	1.1437	1.0340	1.2324	**0.9776**	0.9797	1.0014
Saugeen/W	short	8	1.0307	1.1982	0.9018	1.1446	0.8757	0.9894	1.0811	0.9286	1.0287	1.0471	0.9342	0.9275	0.9633	0.9042	**0.7920**
	medium	80	0.9497	1.3333	1.0667	0.8458	0.9840	1.0857	1.0045	1.0591	0.9280	0.9141	1.3068	1.0984	1.1827	**0.8030**	0.8693
	long	120	1.1034	1.1131	1.2181	0.9723	**0.8922**	1.5197	1.6150	1.5482	1.2699	1.2275	1.3609	1.7141	0.9399	0.9597	0.9635

Table 18: GIFT-Eval Web/CloudOps: per-term test MSE (short, medium, long). For each dataset, model, and term, hyperparameters (context length, n layers, d_{model}, learning rate, loss) are selected by lowest seed-averaged validation MSE. Values are seed-averaged (existing seeds only). Bold: best in row; underline: second-best.

Dataset	Term	H	RNN	GRU	LSTM	S5	LRU	Griffin	Mamba	Mamba2	mLSTM	sLSTM	DeltaNet	DeltaProduct	M2RNN	DLinear	PatchTST
BizITObs-App	short	60	0.2256	0.1544	0.1500	**0.0350**	0.0748	0.0450	0.0625	0.0681	0.0497	0.0430	0.0790	0.0533	0.0585	0.0680	0.0587
	medium	600	0.6423	0.1495	**0.1175**	0.1830	0.1618	0.1527	0.1293	0.2223	0.1494	0.1359	0.1783	0.1482	0.1936	0.1215	0.1386
	long	900	0.9209	0.6308	**0.1640**	0.1704	0.3517	0.1660	0.1750	0.2222	0.1645	0.1718	0.2727	0.2290	0.2084	0.1681	0.2119
BizITObs-L2C/5T	short	48	**0.1927**	0.4241	0.2035	1.1909	0.9434	0.8099	2.4365	2.6319	2.9723	3.2416	0.9154	2.3636	0.9586	0.2972	413.0
	medium	480	1.3671	0.7046	0.7859	0.8352	1.1611	**0.6336**	0.7213	0.7771	0.7202	1.0750	0.7926	0.6524	1.2069	1.3043	379.2
	long	720	1.4998	2.0382	1.5802	1.9517	2.0778	1.8399	1.8582	1.8371	1.2112	**1.1679**	1.8499	1.9043	1.9494	1.9180	-
BizITObs-L2C/H	short	48	**0.3871**	0.5855	0.5346	0.7216	0.5042	0.6551	0.6613	0.6620	0.6147	0.5143	0.5141	0.5964	0.6425	0.4321	0.6488
	medium	480	0.6591	0.6820	**0.6481**	0.7642	0.7393	0.8466	1.3759	2.0057	1.4297	1.8258	0.7509	0.8658	0.8439	0.8140	0.7526
	long	720	0.9533	0.7852	0.8128	0.8077	0.8040	**0.7715**	0.7967	0.8250	0.7752	0.8287	0.8372	0.7722	0.7842	0.8580	0.8081
BizITObs-Service	short	60	3.41e-03	5.49e-03	3.63e-03	3.37e-03	3.46e-03	4.17e-03	4.06e-03	3.67e-03	3.19e-03	5.01e-03	6.65e-03	4.92e-03	4.45e-03	4.66e-03	**2.70e-03**
	medium	600	0.0141	0.0138	0.0142	0.0129	0.0164	0.0145	0.0211	0.0181	0.0128	0.0167	0.0237	0.0186	0.0204	0.0137	**0.0107**
	long	900	0.0292	0.0246	**0.0189**	0.0282	0.0299	0.0218	0.0301	0.0292	0.0194	0.0219	0.0282	0.0261	0.0231	0.0197	1.2511

F DATASETS FEATURES

We adopt the feature definitions used in (Aksu et al., 2024) and reuse their code to compute these quantities on all datasets considered in this study. These features describe complementary aspects of time-series behavior, including temporal structure, forecastability, and regularity.

Temporal structure: Trend and Seasonality. We first consider two features derived from the STL (Seasonality and Trend decomposiiton using Loess) decomposition of the series, namely trend and seasonality strength. Let a time series be decomposed as

$$y_t = T_t + \sum_{m=1}^{M} S_t^{(m)} + R_t,$$

where T_t is the trend component, $S_t^{(m)}$ is the seasonal component associated with period m, and R_t is the remainder. Trend captures the long-term evolution of the series, for example whether it is globally increasing, decreasing, or approximately constant. Following (Aksu et al., 2024), trend strength is defined as

$$F_{\text{trend}} = 1 - \frac{\text{Var}(R_t)}{\text{Var}(T_t + R_t)}.$$

Seasonality captures the presence of repeating temporal patterns. (Aksu et al., 2024) computes one seasonal-strength value for each seasonal period:

$$F_{\text{season}}^{(m)} = 1 - \frac{\text{Var}(R_t)}{\text{Var}(S_t^{(m)} + R_t)}, \qquad m = 1, \ldots, M.$$

Thus, for multi-seasonal series there are multiple seasonal-strength values, one per period. In our experiments, we use the 'seasonal_strength' quantity returned by the (Aksu et al., 2024) pipeline, leaving any aggregation behavior to the underlying implementation.

The values of F_{trend} and F_{season} are bounded to $[0, 1]$ with negative values set to 0 and values greater than 1 set to 1. Higher F_{trend} values indicate a stronger trend component while higher F_{season} indicate more pronounced and regular periodic structure. Together, trend and seasonality characterize the temporal organization of the series.

Forecastability: Entropy and Hurst exponent. We next consider entropy and the Hurst exponent, two features related to predictability. Entropy measures the amount of information in the series through the Shannon entropy of its normalized spectral density. Let $f(\lambda)$ denote the spectral density estimate and let

$$p_k = \frac{f(\lambda_k)}{\sum_j f(\lambda_j)}$$

be the normalized spectral density. The spectral entropy is then

$$F_{\text{entropy}} = -\sum_k p_k \log p_k.$$

Lower entropy indicates a more concentrated spectrum and typically an easier forecasting problem, whereas higher entropy suggests more irregular dynamics.

The Hurst exponent measures long-range dependence and persistence in the series. In (Aksu et al., 2024), it is obtained from the maximum-likelihood estimate of the fractional differencing parameter d, and is defined as

$$F_{\text{Hurst}} = d + \frac{1}{2}.$$

Entropy and the Hurst exponent provide complementary views of forecastability.

Regularity: Stability and Lumpiness. Finally, we consider stability and lumpiness, which quantify how much the local statistics of the series vary over time. Let the time series be partitioned into B non-overlapping segments. For each segment b, let μ_b and σ_b^2 denote its sample mean and variance. Stability is defined as the variance of the segment means,

$$F_{\text{stability}} = \text{Var}(\mu_1, \mu_2, \ldots, \mu_B),$$

while lumpiness is defined as the variance of the segment variances,

$$F_{\text{lumpiness}} = \text{Var}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_B^2).$$

High stability or lumpiness values indicate substantial distributional changes across segments, suggesting lower regularity and potentially greater forecasting difficulty. These two measures therefore capture complementary aspects of local non-stationarity. Tables 19 and 20 show the features values for the datasets.

Table 19: Statistical features and characteristics of the LTSF datasets. Feature values are per-channel means over the training split, z-scored with training statistics Variables = channels; Timesteps = series length. Seas. Str. = seasonal strength.

Dataset	Domain	Freq	Variables	Timesteps	Trend	Seas. Str.	Entropy	Hurst	Lumpiness	Stability
ETTh1	Energy	1 hr	7	17420	0.85	0.61	0.43	0.96	0.24	0.66
ETTh2	Energy	1 hr	7	17420	0.94	0.50	0.40	0.99	0.17	0.86
ETTm1	Energy	15 min	7	69680	0.99	0.19	0.38	0.96	0.01	0.98
ETTm2	Energy	15 min	7	69680	0.99	0.19	0.35	0.99	0.01	0.98
Electricity	Energy	1 hr	321	26304	0.80	0.88	0.34	1.04	0.20	0.25
Weather	Nature	10 min	21	52696	0.91	0.18	0.46	0.88	307.17	0.88
Traffic	Transport	1 hr	862	17544	0.35	0.74	0.48	0.69	1.13	0.11
Exchange Rate	Finance	1 day	8	7588	1.00	0.17	0.25	0.99	0.00	1.00
Illness	Healthcare	1 week	7	966	0.99	0.17	0.47	0.81	0.11	0.94

Table 20: Statistical features and characteristics of the GIFT-Eval datasets. Feature values are per-channel means over the training split, z-scored with training statistics Variables = channels (series $\times$ target dimensions); Timesteps = series length (median over series). Seas. Str. = seasonal strength.

Dataset	Domain	Freq	Variables	Timesteps	Trend	Seas. Str.	Entropy	Hurst	Lumpiness	Stability
us_births/D	Healthcare	1 day	1	7305	0.91	0.91	0.40	1.13	0.08	0.47
us_births/W	Healthcare	1 week	1	1043	0.98	0.26	0.30	1.04	0.00	0.95
jena_weather/H	Nature	1 hr	21	8784	0.74	0.53	0.49	0.91	22.72	0.56
saugeenday/D	Nature	1 day	1	23741	0.91	0.18	0.67	0.61	1.19	0.82
saugeenday/W	Nature	1 week	1	3391	0.83	0.18	0.73	0.62	2.08	0.66
bizitobs_application	Web/CloudOps	10 s	2	8834	1.00	0.26	0.28	0.66	0.06	0.99
bizitobs_l2c/H	Web/CloudOps	1 hr	7	2664	0.88	0.21	0.35	0.38	0.26	0.76
bizitobs_service	Web/CloudOps	10 s	42	8835	0.99	0.26	0.35	0.65	0.08	0.98
ett1/H	Energy	1 hr	7	17420	0.85	0.67	0.46	0.89	0.24	0.63
ett2/H	Energy	1 hr	7	17420	0.97	0.55	0.34	1.00	0.06	0.92
ett1/15T	Energy	15 min	7	69680	0.99	0.19	0.42	0.89	0.01	0.97
ett2/15T	Energy	15 min	7	69680	0.99	0.19	0.30	1.02	0.00	0.99
solar/H	Energy	1 hr	137	8760	0.44	0.90	0.27	0.83	0.22	0.06
electricity/H	Energy	1 hr	370	35064	0.85	0.87	0.32	0.84	0.27	0.36

G Experimental Setup and Details

For each model–dataset pair, the experimental design targeted three model depths (1, 2, 3), three context lengths (96, 336, 720), and four prediction horizons (96, 196, 336, 720). Because hyperparameter optimization was additionally performed on datasets, a naive exhaustive design would require $3 \times 3 \times 4 = 36$ hyperparameter optimization runs for each model–dataset pair. Given the available computational resources, a subset of experimental cells was selected such that all three depths, three context lengths, and four prediction horizons were covered while limiting the total number of hyperparameter optimization runs (see Table 21). For the real-world LTSF and GIFT-Eval datasets, hyperparameter optimization was not performed independently on every dataset. Instead, a subset of representative datasets was selected. For each model and each selected experimental cell, hyperparameter optimization was performed on these representative datasets (see table 22). The best-performing hyperparameter configuration identified for each model and cell was then reused across the remaining datasets in its group.

Table 21: The eight experimental cells. Each cell fixes a forecast horizon H, a context window L and a number of stacked layers N; its identifier reads `p`H`_l`L`_n`N. Starting from the reference cell (H=336, L=336, N=3), each group varies a single factor while keeping the other two fixed. On GIFT-Eval, horizons follow the benchmark's short/medium/long terms ($1\times$, $10\times$ and $15\times$ a frequency-dependent base length, e.g. 48 steps for hourly data), with the medium term playing the role of the reference horizon; the H=192 cell has no GIFT-Eval counterpart, giving seven cells there.

Factor varied	Cell	Horizon H	Context L	Layers N	GIFT-Eval cell
Horizon	p96_l336_n3	96	336	3	pshort_l336_n3
	p192_l336_n3	192	336	3	–
	p336_l336_n3†	336	336	3	pmed_l336_n3
	p720_l336_n3	720	336	3	plong_l336_n3
Context	p336_l96_n3	336	96	3	pmed_l96_n3
	p336_l720_n3	336	720	3	pmed_l720_n3
Depth	p336_l336_n1	336	336	1	pmed_l336_n1
	p336_l336_n2	336	336	2	pmed_l336_n2

†Reference cell, shared by all three groups. Illness is too short for these settings and instead uses L=36, $H \in \{24, 36\}$ and $N \in \{1, 2, 3\}$.

Table 22: Representative datasets and their groups. Hyperparameters (learning rate, model width, loss and model-specific options) are tuned only on the representative dataset of each group and then reused, unchanged, for every other dataset in that group. GIFT-Eval groups correspond to the dataset domains of Table 20.

Benchmark	Group	Representative	Datasets in group
LTSF	ETT-like	ETTh1	ETTh1, ETTh2, ETTm1, ETTm2, Exchange Rate
	Weather	Weather	Weather
	High-dimensional	Traffic	Traffic, Electricity
GIFT-Eval	Energy	ett1/H	ett1/H, ett2/H, ett1/15T, ett2/15T, solar/H, electricity/H
	Healthcare	us_births/D	us_births/D, us_births/W
	Nature	jena_weather/H	jena_weather/H, saugeenday/D, saugeenday/W
	Web/CloudOps	bizitobs_l2c/H	bizitobs_l2c/H, bizitobs_application, bizitobs_service

Table 23: Common hyperparameters, searched identically for every model.

Hyperparameter	**Search space**	**Models**
lr	log-uniform on $[10^{-5}, 3 \times 10^{-3}]$	all 10
loss_fn	{mse, mae}	all 10
d_model	{128, 256, 512}	GRU
	{128, 256}	sLSTM, mLSTM, S5, LRU, Mamba-2, Mamba, DeltaNet, DeltaProduct, M^{2}RNN

Table 24: Architecture-specific hyperparameters axes for the 10 models Common axes above (table 23) are searched identically for every model.

Hyperparameter	Search space	Models
`d_state`	$\{16, 32, 64\}$	Mamba
	$\{32, 64, 128\}$	LRU, S5, Mamba-2
`n_heads`	$\{2, 4, 8\}$	sLSTM, mLSTM, Mamba-2, DeltaNet, DeltaProduct, M^2RNN
`n_householder`	$\{1, 2, 4\}$	DeltaProduct

H DELTANET FAMILY

Here we compare performance of models from DeltaNet family with increasing expressivity level. Columns are DeltaNet, Neg-DeltaNet, DeltaProduct.

Table 25: DeltaNet family: mean test MSE over datasets in each category (after averaging horizons). Bold = best, underline = second-best.

Category	DeltaNet	Neg-DeltaNet	DeltaProduct
Trend	1.76e-03	**1.46e-03**	<u>1.68e-03</u>
Periodic	<u>0.0147</u>	0.0151	**0.0114**
Noise	**3.3249**	<u>3.3458</u>	3.3766
Anomaly	0.4138	**0.3707**	<u>0.3922</u>
Dependency	**0.8741**	0.8882	<u>0.8778</u>
Cross-variable	**0.9087**	<u>0.9111</u>	0.9135
Univariate real-world simulation	**0.3156**	<u>0.3166</u>	0.3365
Multivariate real-world simulation	<u>0.4439</u>	0.4466	**0.4408**
Mean	<u>0.7872</u>	**0.7869**	0.7938

Table 26: DeltaNet family: number of datasets (out of 59) where each model is best / second-best after averaging over $H \in \{96, 336, 720\}$.

	DeltaNet	Neg-DeltaNet	DeltaProduct
# best (horizon-avg)	29	17	13
# second-best	21	21	17

Table 27: Trend (DeltaNet family): test MSE averaged over $H \in \{96, 336, 720\}$. HP selected independently per horizon. Bold = best, underline = second-best.

Dataset	DeltaNet	Neg-DeltaNet	DeltaProduct
Exponential-Trend	<u>6.31e-03</u>	**5.70e-03**	6.32e-03
Gaussian-Trend	8.12e-03	<u>6.08e-03</u>	**5.14e-03**
Gompertz-Trend	**2.76e-07**	<u>5.14e-07</u>	1.42e-06
Linear-Trend	<u>6.56e-07</u>	**6.41e-07**	6.75e-07
Logarithmic-Function	**3.81e-06**	<u>6.52e-06</u>	1.01e-05
Logistic-Trend	<u>2.13e-04</u>	**1.66e-04**	5.93e-04
Negative-Exponential-Trend	**1.35e-08**	1.40e-08	<u>1.35e-08</u>
Piecewise-Linear	<u>4.47e-03</u>	**3.06e-03**	4.82e-03
Power-Law-Trend	**7.13e-06**	<u>9.28e-06</u>	1.48e-05
Quadratic-Function	**2.23e-04**	<u>1.04e-03</u>	1.55e-03
Step-Function	<u>6.13e-15</u>	1.28e-14	**1.41e-15**
Mean	1.76e-03	**1.46e-03**	<u>1.68e-03</u>

Table 28: Periodic (DeltaNet family): test MSE averaged over $H \in \{96, 336, 720\}$. HP selected independently per horizon. Bold = best, underline = second-best.

Dataset	DeltaNet	Neg-DeltaNet	DeltaProduct
Sin(fre=0.005)	**1.42e-05**	6.56e-05	4.17e-05
Sin(fre=0.05)	**1.43e-05**	4.03e-05	2.64e-05
Double-Sin	1.21e-05	**1.05e-05**	2.15e-05
Triple-Sin	**1.48e-05**	5.12e-05	3.47e-05
Five-Sin	**1.71e-05**	4.42e-05	1.35e-03
Ten-Sin	**1.21e-04**	3.69e-04	5.90e-04
Exp-Sine-Wave	4.58e-05	**1.20e-05**	3.68e-05
Triangle-Wave	2.39e-05	**1.22e-05**	3.22e-05
Sawtooth-Wave	**1.06e-05**	4.39e-05	2.61e-04
Square-Wave	0.1467	0.1499	**0.1121**
Mean	0.0147	0.0151	**0.0114**

Table 29: Noise (DeltaNet family): test MSE averaged over $H \in \{96, 336, 720\}$. HP selected independently per horizon. Bold = best, underline = second-best.

Dataset	DeltaNet	Neg-DeltaNet	DeltaProduct
Exponential-Trend SNR=-10dB	1.1417	**1.1287**	1.1428
Exponential-Trend SNR=0dB	**3.3164**	3.3217	3.3901
Exponential-Trend SNR=10dB	**13.8296**	13.8749	14.0299
Five-Sin SNR=-10dB	0.9577	0.9578	**0.9563**
Five-Sin SNR=0dB	**0.5982**	0.6682	0.6289
Five-Sin SNR=10dB	**0.1057**	0.1238	0.1117
Mean	**3.3249**	3.3458	3.3766

Table 30: Anomaly (DeltaNet family): test MSE averaged over $H \in \{96, 336, 720\}$. HP selected independently per horizon. Bold = best, underline = second-best.

Dataset	DeltaNet	Neg-DeltaNet	DeltaProduct
Quadratic + 0.5% point anomalies	**0.2040**	0.2370	0.2386
Quadratic + 5% point anomalies	0.5634	**0.5312**	0.5474
Quadratic + 20% point anomalies	0.4741	**0.3439**	0.3905
Mean	0.4138	**0.3707**	0.3922

Table 31: Dependency (DeltaNet family): test MSE averaged over $H \in \{96, 336, 720\}$. HP selected independently per horizon. Bold = best, underline = second-best.

Dataset	DeltaNet	Neg-DeltaNet	DeltaProduct
ARMA(1,1)	**0.9272**	0.9907	0.9454
ARMA(2,2)	1.0565	1.0603	**1.0545**
ARMA-Long-Dependency	**0.9890**	0.9945	0.9953
Random-Walk	0.4289	**0.4211**	0.4249
White-Noise	0.9690	0.9746	**0.9688**
Mean	**0.8741**	0.8882	0.8778

Table 32: Cross-variable (DeltaNet family): test MSE averaged over $H \in \{96, 336, 720\}$. HP selected independently per horizon. Bold = best, underline = second-best.

Dataset	DeltaNet	Neg-DeltaNet	DeltaProduct
Feature Lag (5 steps)	1.0567	**1.0542**	1.0564
Feature Lag (10 steps)	**0.9547**	0.9622	0.9623
Feature Lag (24 steps)	**0.9867**	0.9887	0.9940
Feature Lag (48 steps)	**0.9072**	0.9112	0.9139
Sine-Noise Composition	0.4704	0.4700	**0.4696**
Conditional Relationship	**0.9995**	1.0028	1.0029
Nonlinear Relationship	**0.9412**	0.9441	0.9459
Multivariable Complex	**0.9530**	0.9552	0.9631
Mean	**0.9087**	0.9111	0.9135

Table 33: Univariate real-world simulation (DeltaNet family): test MSE averaged over $H \in \{96, 336, 720\}$. HP selected independently per horizon. Bold = best, underline = second-best.

Dataset	DeltaNet	Neg-DeltaNet	DeltaProduct
Economic-Indicator	0.5525	**0.4822**	0.6368
Electricity-Consumption	0.1381	0.1550	**0.1003**
Industrial-Sensor	**0.3417**	0.3995	0.4736
IoT-Sensor-Battery	3.79e-05	**3.19e-05**	4.80e-05
Network-Traffic	**0.0375**	0.0558	0.0409
Retail-Sales	0.5338	**0.5310**	0.5456
Stock-Price	1.0981	**1.0467**	1.0974
Temperature-Sensor	0.0853	0.1131	**0.0844**
Website-Traffic	0.0536	0.0658	**0.0493**
Mean	**0.3156**	0.3166	0.3365

Table 34: Multivariate real-world simulation (DeltaNet family): test MSE averaged over $H \in \{96, 336, 720\}$. HP selected independently per horizon. Bold = best, underline = second-best.

Dataset	DeltaNet	Neg-DeltaNet	DeltaProduct
Ad-Sales	0.6151	0.6244	**0.6093**
Intervention-Effect	0.0992	0.1051	**0.0973**
Lotka-Volterra	**5.05e-03**	6.20e-03	5.65e-03
Macro-Economy	0.7935	**0.7918**	0.8022
SIR-Model	0.2539	0.2365	**0.2287**
Supply-Demand-Price	**0.8994**	0.9208	0.9011
Weather-Sales	**0.4410**	0.4417	0.4415
Mean	0.4439	0.4466	**0.4408**