STATECOMP: SHARED TRANSFER OF SEASONAL STATES FOR COMPACT FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term time-series forecasting requires preserving fine-grained temporal structure over many future steps, yet learning a fully flexible history-to-future transformation can be parameter intensive. We explore a complementary hypothesis: time-series shapes may be complex, while the rules governing their evolution can be comparatively simple and shared. Based on this idea, we propose StateComp, a compact forecaster that separates the representation of temporal states from their learned transfer. StateComp constructs multiwidth seasonal states directly from each input window and propagates them using a small set of coefficients shared across phases and variables. A low-rank history correction complements these prescribed states by capturing residual dependence that seasonal summaries miss. This design allows rich, input-dependent temporal structure to be retained without assigning learned parameters to every phase or variable. We further analyze the interaction between the two components, showing that shared seasonal transfer can represent multiple predictive directions and that correction rank controls the remaining error relative to the coverage of the state dictionary. Across five long-term forecasting benchmarks, StateComp achieves competitive accuracy with only 61–389 learned parameters, and obtains lower MSE than SparseTSF and MixLinear at every evaluated horizon on Electricity and Traffic.

1 INTRODUCTION

Long term time series forecasting requires predicting many future steps while preserving both large scale temporal evolution and fine grained variation (Wang et al., 2024; Chen et al., 2024). The problem is particularly demanding in multivariate settings, where the output spans many variables and time points and predictive dependencies occur at different temporal scales (Zhou et al., 2021; Nie et al., 2023; Wang et al., 2024). Recent observations describe the current state of a process, while repeated patterns over longer histories provide more stable evidence about its temporal structure (Xu et al., 2026; Han et al., 2025). A forecaster must determine which historical information remains useful far into the future and how local variation, recurring structure, and longer changes should be combined in one forecast (Zhang et al., 2025; Wu et al., 2025).

Existing forecasting architectures address this problem through different assumptions about how predictive structure should be represented. Attention-based methods (Zhou et al., 2021; Nie et al., 2023; Oreshkin et al., 2020; Wang et al., 2024) learn long-range dependencies, patch-based models organize local subsequences, basis-expansion methods reconstruct forecasts from learned components, and multiscale architectures combine information across temporal resolutions. Seasonal decomposition and state-space methods (Bandara et al., 2025; De Livera et al., 2011) instead expose recurring structure more explicitly. These approaches provide flexible mechanisms for transforming historical observations into future trajectories, but such flexibility can require the learned representation to encode both the detailed shape of the signal and the rule by which that shape evolves over time.

Building on this perspective, we ask whether detailed temporal structure can be supplied by the input itself, leaving the model to learn only how that structure evolves. This requires more than periodic alignment. We posit that differences across series and input windows are largely encoded in their current states, whereas the persistence and historical dependence of those states are regular enough to share coefficients across phases and variables. Under this view, the input carries what is specific to each series, while a small set of coefficients governs their common temporal evolution.

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.

This separation raises two practical challenges. First, the relevant state can change over time. Short historical summaries respond quickly to recent change but are sensitive to transient variation, whereas longer summaries are stable but may obscure that change. Multiple temporal widths therefore provide complementary views of the current state. Second, seasonal summaries cannot retain every useful part of the observed history. We assume that once dominant recurring structure is represented explicitly, the remaining predictive dependence can be captured by a small number of historical directions. This motivates a rank constrained correction (Yu et al., 2016; Barratt et al., 2021), with the constraint placed on the residual map rather than on the series or its seasonal profiles.

Based on these principles, we introduce StateComp, which separates state representation from state transfer. A multiwidth seasonal bank computes base cycle profiles at several historical scales and position specific structure across longer cycles directly from each input window. A compact readout transfers entire states through coefficients shared across phases and variables, so adding temporal detail does not require a coefficient for every phase. A rank constrained history map retains access to individual observed cycles and corrects what the seasonal states miss. Expanding the state bank broadens the structure represented explicitly, while raising the correction rank adds learned directions for departures from those states. The resulting forecast remains detailed and input dependent while the learned temporal operator stays small. Our contributions are summarized as follows:

- We propose StateComp, a compact forecasting framework that computes multiwidth seasonal states from the input and transfers them with coefficients shared across phases and variables. The seasonal readout size is independent of seasonal resolution and variable count.
- We add a rank-constrained history correction and show how shared seasonal transfer preserves several predictive directions while the correction reduces the remaining error.
- Across five benchmarks, StateComp achieves competitive accuracy with 61 to 389 parameters and outperforms SparseTSF and MixLinear on Electricity and Traffic at every evaluated horizon.

2 PRELIMINARIES

2.1 PROBLEM SETUP

Given a multivariate history $X \in \mathbb{R}^{L\times C}$, long-horizon forecasting predicts the next H observations $\widehat{Y} \in \mathbb{R}^{H\times C}$, with H potentially comparable to L. Following channel-independent forecasting models (Zeng et al., 2023; Nie et al., 2023), we apply the same scalar forecaster to every variable and suppress the variable index below. To reduce sensitivity to nonstationarity at the window level, we center each input by its own mean before constructing its temporal states (Kim et al., 2022).

2.2 CYCLES, PHASES, AND SEASONAL STATES

Let p denote a base period that divides both L and H, and define $n = L/p$ observed cycles and $m = H/p$ forecast cycles. Periodic forecasting methods commonly organize observations by their position within such cycles (Lin et al., 2024b;a; Niu et al., 2026). For one centered input series $Z \in \mathbb{R}^L$, we form the cycle–phase array $\mathsf{T} \in \mathbb{R}^{n\times p}$,

$$\mathsf{T}_{ij} = Z_{ip+j}, \qquad 0 \le i < n, \quad 0 \le j < p,$$

where rows index successive cycles and columns index phases within a cycle. For a width w, the base-cycle profile $F_w \in \mathbb{R}^p$ is the vector of phase-wise averages over the latest w cycles,

$$F_w(j) = \frac{1}{w}\sum_{i=n-w}^{n-1} \mathsf{T}_{ij}, \qquad \ell_w = \frac{1}{p}\sum_{j=0}^{p-1} F_w(j).$$

We call the phase average ℓ_w its level. Both states are recomputed from each input window rather than learned, so their temporal resolution is determined by p rather than by the number of model parameters. Different widths w summarize the same phases over different amounts of history.

We also allow $g \ge 2$ base cycles to form a longer cycle. Averaging all cycles into F_w removes variation among the g positions of this longer cycle. To retain it, we average observations only at matching longer-cycle positions. For the latest v complete groups, define

$$R_v(d, j) = \frac{1}{v}\sum_{a=1}^{v} \mathsf{T}_{n-ag+d,\, j}, \qquad d = 0, \dots, g-1.$$

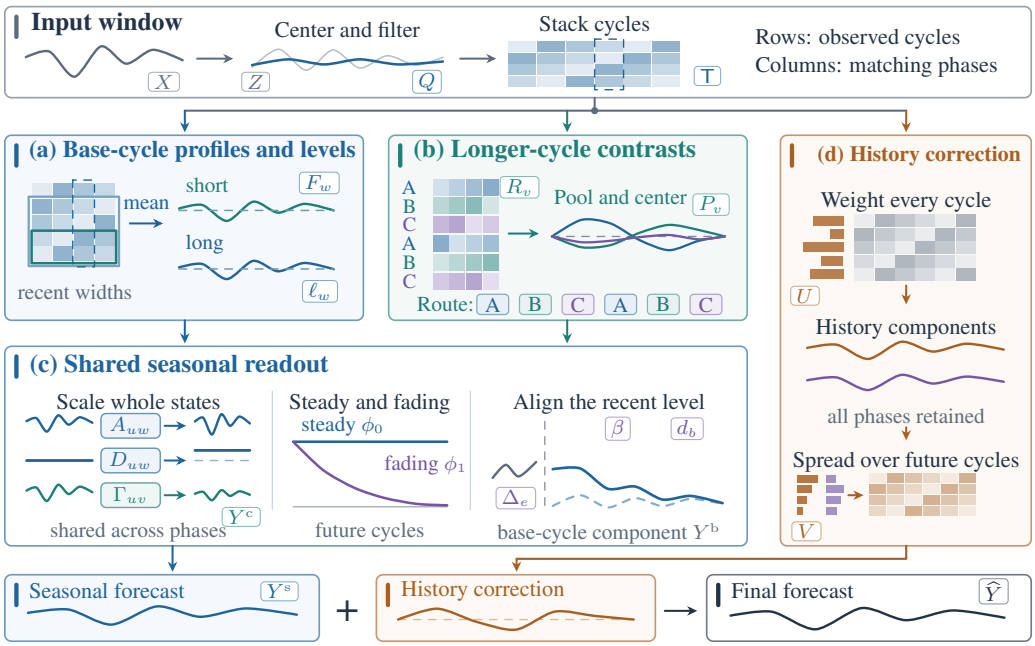


Figure 1: **StateComp transfers observed seasonal detail and adds a learned history correction.** The cycle grid feeds both state families and the history branch. **(a)** Matching-phase averages give base-cycle profiles and levels. **(b)** Longer-cycle positions are pooled and centered into contrasts. **(c)** Shared gains transfer whole states and align the first cycle. **(d)** A low-rank map reads all observed cycles and corrects that forecast. Coefficients are shared across variables; drawings are schematic.

Thus R_v contains one position-specific profile for each d in the longer cycle. Classical seasonal models similarly separate observation-derived seasonal states from a few parameters governing their evolution (Hyndman et al., 2002; De Livera et al., 2011; RB, 1990; Bandara et al., 2025).

3 Methodology

We provide StateComp that forecasts from the seasonal states using two components, shown in Figure 1. A multiwidth bank collects base-cycle and longer-cycle states directly from each input window, and a shared readout transfers them across the forecast horizon. A low-rank temporal map then corrects the predictable variation that those states miss while preserving their seasonal detail.

3.1 Seasonal state transfer

A seasonal state is not static. Levels and shapes drift, so no single width is correct. A short average tracks the current state, a long one is less sensitive to noise, and keeping both lets the readout weigh them. A state is computed from the window rather than learned, so several widths cost no parameters, and we prescribe a whole bank rather than one. The base widths $\mathcal{W} \subseteq \{1, \ldots, n\}$ and longer-cycle widths $\mathcal{G} \subseteq \{v \geq 1 : gv \leq n\}$ are both fixed before fitting. We first subtract the input mean μ, apply a filter with two shared scalars s_1, s_2, and form T from the filtered window Q (Appendix A).

Transferring base-cycle profiles. A base-cycle profile contains both its phase-wise shape and its overall level. Scaling it with one coefficient would tie the two, since lowering the forecast level would also flatten its phase contrasts. We therefore pair each base-cycle profile with its level. The widths also remain informative for different lengths of time, so we transfer every pair through two time bases indexed by u, one constant and one decaying over the forecast cycles, $\phi_0(k) = 1$ and

$\phi_1(k) = \exp[-k/\max(m/2, 1)]$. Summing over widths gives the base-cycle component Y^{b},

$$Y^{\mathrm{b}}_{kp+j} = \mu + \sum_{u=0}^{1} \phi_u(k) \sum_{w \in \mathcal{W}} [A_{uw} F_w(j) + D_{uw} \ell_w]. \tag{1}$$

The base-profile coefficient A_{uw} scales shape and mean together, while the level coefficient D_{uw} shifts the mean and leaves phase contrasts unchanged. Four coefficients per width therefore control shape and level separately, regardless of the phase resolution or number of input variables.

Longer-cycle contrasts. Each base-cycle profile provides one phase vector that is transferred to every future base cycle. It cannot assign distinct base-cycle profiles to different positions in a longer cycle. With days as the base cycles and a week as the longer cycle, for example, the same base-cycle profile would serve Monday through Sunday even when their patterns differ systematically. Changing w changes the observations being averaged but still returns only one vector. The position-specific profiles R_v retain a separate vector for each of the g positions. They cannot be added directly because each also contains the common base-cycle shape already supplied by Eq. (1). We remove that common shape by subtracting the mean over positions, then route position $k \bmod g$ to forecast cycle k to obtain the contrast

$$P_v(k, j) = R_v(k \bmod g,\, j) - \frac{1}{g} \sum_{d=0}^{g-1} R_v(d, j), \tag{2}$$

aligned so that forecast cycle $k = 0$ follows the latest complete group. By construction P_v sums to zero over any g consecutive forecast cycles, so each state carries only what the other does not, and $P_v(k, j) = P_v(k + g, j)$ repeats with the longer cycle. The same two time bases transfer it:

$$Y^{\mathrm{c}}_{kp+j} = \sum_{u=0}^{1} \phi_u(k) \sum_{v \in \mathcal{G}} \Gamma_{uv} P_v(k, j).$$

Like A_{uw} for a base-cycle profile, each Γ_{uv} scales a whole longer-cycle state, so two coefficients per width transfer all gp contrast values. Centering removes the common shape captured by ℓ_w, so no level coefficient is needed. Constructing the state adds no learned coefficients. For Eq. (2), the implementation forms R_v with the median blend and phase smoothing specified in Appendix A.

Boundary alignment and assembly. A forecast assembled from averages need not meet the end of the observed window, and a visible step at the boundary costs accuracy exactly where the forecast is most reliable. Let $\bar{X}_e$ average the final e input values, $\bar{Y}^{\mathrm{b}}_e$ the final e phases of the first cycle of Y^{b}, and $\Delta_e = \bar{X}_e - \bar{Y}^{\mathrm{b}}_e$ the gap between them. A learned gain β transfers that gap, and because a local offset should not persist over the whole horizon it decays through d_b, whose time constant is learned through ζ. The seasonal branch is

$$Y^{s}_{kp+j} = Y^{\mathrm{b}}_{kp+j} + Y^{\mathrm{c}}_{kp+j} + \beta\, \Delta_e\, d_b(kp + j), \tag{3}$$

with $d_b(h) = \exp[-(h + 1)/(p \exp(\mathrm{clip}(\zeta)))]$. See Appendix A for implementation details.

3.2 Low-rank temporal correction

The seasonal branch can only combine the widths it was given, and the right averaging scale need not be among them. The history branch answers the third question by learning weights over all filtered cycles, under an assumption of its own: once the periodic structure above has been accounted for, the predictable variation that remains can be described by a few directions over the observed cycles. The constraint is therefore on the correction map alone; the series and its seasonal profiles may be as rich as they please. Rank-constrained temporal maps are themselves established (Yu et al., 2016; Barratt et al., 2021); what differs here is their role, since this one corrects a forecast that the prescribed states already supply rather than carrying the forecast on its own. With factors $U \in \mathbb{R}^{r \times n}$ and $V \in \mathbb{R}^{m \times r}$ of rank $0 \le r \le \min(n, m)$, it adds this correction to phase j of each future cycle k:

$$\widehat{Y}_{kp+j} = Y^{s}_{kp+j} + (VU\mathsf{T})_{kj}. \tag{4}$$

For $r = 1$, weights over the n historical cycles produce one history component across the p phases, and a learned vector scales it across the m forecast cycles. Each additional rank adds another component and its future weights. Because these weights range over every historical cycle, the branch can both adjust the prescribed transfers and represent departures from them in the observed history.

3.3 PARAMETER SHARING AND MODEL CAPACITY

State resolution and learned capacity enter StateComp separately. A base cycle profile contains p values and a longer cycle contrast contains gp values, yet each learned coefficient transfers an entire state. The same coefficients are also reused for every variable. Writing $J = |\mathcal{W}|$ and $K = |\mathcal{G}|$, the resulting number of learned scalars is

$$N_\theta = 4 + 4J + 2K + \mathbf{1}\{v_{\mathrm{rob}} \in \mathcal{G}\} + r(n + m). \tag{5}$$

The four global scalars are s_1, s_2, β, and ζ. Each base cycle width adds two coefficients for its profile and two for its level, while each longer cycle width adds two contrast coefficients. A designated robust width adds the optional median gain described in Appendix A. None of these terms depends on the number of phases p or variables C. Increasing p gives each state more values without enlarging its shared readout. The history factors contribute $r(n + m)$ coefficients, since each rank adds n weights over observed cycles and m weights over forecast cycles for one learned history direction.

The width sets and the correction rank therefore control different forms of capacity. Adding a width exposes another input-derived averaging scale under the same transfer rule. Increasing r adds a learned history component for variation outside those states. Setting $r = 0$ isolates the prescribed seasonal transfer. This separation lets the representation retain detailed seasonal structure without assigning learned coefficients to individual phases or variables represented in the input window.

3.4 EXPRESSIVITY OF SHARED STATE TRANSFER

The seasonal readout applies shared coefficients to directions already encoded by prescribed states, while the history correction learns additional directions through a rank-constrained map. To compare these two forms of capacity, we introduce a stylized Gaussian construction in which one prescribed seasonal state spans several predictive directions. Its conditional mean is available in closed form, allowing us to relate state coverage and correction rank directly to population prediction error.

Let a longer cycle have g positions and $q = g - 1$ orthonormal, zero mean contrasts, collected in $S \in \mathbb{R}^{g \times q}$. In the latest v complete groups, observe $S(a + \eta_j)$, where $a \sim N(0, I_q)$ and the independent noises satisfy $\eta_j \sim N(0, \sigma^2 I_q)$ for $\sigma^2 > 0$. Other historical cycles are zero, phases within a cycle are identical, and future groups have target Sa. Averaging the observed groups and multiplying by $\gamma = v/(v + \sigma^2)$ gives the conditional mean. The constant readout of the width v longer cycle state realizes this map with one active gain, although its cycle to cycle operator has rank q. For $m = tg$ forecast cycles, the posterior variance gives $\mathrm{MSE}_{\mathrm{Bayes}} = q\sigma^2/[g(v + \sigma^2)]$. A pure rank r temporal map retains at most r of the q predictive directions, so excess over the Bayes MSE is

$$\mathrm{MSE}_{\text{rank-r}} - \mathrm{MSE}_{\mathrm{Bayes}} = \frac{(q - r)_+}{g} \frac{v}{v + \sigma^2}. \tag{6}$$

After covariance whitening, the conditional mean operator has q equal nonzero singular values. Fixed seasonal routing specifies all q directions, while their transfer shares one learned gain.

More generally, let q centered, pairwise uncorrelated direction histories z_i have covariance $G \succ 0$, and let their conditional future means be $hb^\top z_i$, with shared temporal coefficient b and future shape h. Let $A \in \mathbb{R}^{J \times v}$ collect the fixed temporal weighting rules used to form J history summaries. For direction i, the vector Az_i contains the summaries available to the shared readout, whose output basis represents h. The part of b that these summaries cannot represent is

$$\delta_A^2 = \min_{w \in \mathrm{row}\, A} (b - w)^\top G (b - w).$$

For an arbitrary linear correction of rank $0 \le r < q$, the unpenalized population optimum has total excess squared error $(q - r)\|h\|_2^2 \delta_A^2$ (Proposition 1). Adding linearly independent summary rows expands $\mathrm{row}\, A$ and can only reduce δ_A^2. The same unrepresented component occurs in each of the q predictive directions, and a rank-r correction can recover at most r of them. This result motivates combining several seasonal summaries with a small history map. Appendix B proves it by best-rank approximation (Eckart & Young, 1936), and Appendix C gives deterministic bounds for this setting.

4 RELATED WORK

Forecasting architectures. Neural forecasters differ mainly in how they mix information across time and across variables. Transformer variants adapt attention to long series: Informer uses sparse atten-

tion, PatchTST attends over patches with shared channel-wise weights, and iTransformer over variate tokens, while Autoformer and FEDformer build decomposition and frequency-domain attention into the architecture (Zhou et al., 2021; Nie et al., 2023; Liu et al., 2024; Wu et al., 2021; Zhou et al., 2022). TimesNet reshapes the series by its dominant periods and applies convolution (Wu et al., 2023). N-BEATS and NHITS forecast through basis expansions and hierarchical interpolation (Oreshkin et al., 2020; Challu et al., 2023), while TSMixer, TiDE, TimeMixer, and FreTS replace attention with dense, multiscale, or frequency-domain mixing (Chen et al., 2023; Das et al., 2023; Wang et al., 2024; Yi et al., 2023). HiPPO-KAN instead applies a nonlinear map to polynomial history summaries (Lee et al., 2024). A further line reuses a pretrained model rather than training one: language models adapted to forecasting (Chang et al., 2023; Dey et al., 2026), lightweight heads with pruned low-rank adapters (Zheng & Jin, 2026), conditioning on textual cues (Wei et al., 2026), and calibration modules updated at test time (Medeiros et al., 2025). Across these families, instance normalization is a common preprocessing step for nonstationary inputs (Kim et al., 2022).

Seasonal representations and compactness. Classical decomposition and exponential smoothing extract or update seasonal states from the observations (RB, 1990; Bandara et al., 2025; Hyndman et al., 2002; De Livera et al., 2011), and such states can instead be learned, either as per-series state-space parameters (Rangapuram et al., 2018) or as a recurrent cycle added back to the forecast (Lin et al., 2024a). Periodic structure is also what makes small temporal maps viable. Compact forecasters decompose the series (Zeng et al., 2023), truncate a spectral representation (Xu et al., 2024), share a map across the positions of a period (Lin et al., 2024b), mix segment and frequency maps (Ma et al., 2026), or organize the map by basis, time-frequency decomposition, or phase (Huang et al., 2025; Zhao et al., 2026; Niu et al., 2026); Toner & Darlow (2024) express such linear forecasts as augmented-feature regression. A related line constrains the temporal operator itself, learning dynamics in compressed representations (Yu et al., 2016; Barratt et al., 2021) or factorizing past and future separately alongside a learned periodic prototype (Lin et al., 2026).

5 Experiments

5.1 Experimental setup

Datasets and metrics. We evaluate ETTh1, ETTh2, ETTm2, Electricity, and Traffic. ETTm2 uses 15-minute observations; the other datasets are sampled hourly. StateComp uses $L = 720$ and $H \in \{96, 192, 336, 720\}$. Chronological splits are 12/4/4 months for ETT and 70/10/20% for Electricity and Traffic. Standardization uses the training segment. We fit by minimizing mean squared error over training windows and select models by validation MSE. We average MSE and MAE over all forecast values and valid test windows and report parameter counts, which include only the real scalars fitted on forecasting data: structural choices and a few constants held fixed across datasets are excluded, and Appendix I.2 quantifies their influence. Appendix D gives details.

Baselines. We compare with FEDformer (Zhou et al., 2022), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), DLinear (Zeng et al., 2023), FITS (Xu et al., 2024), SparseTSF (Lin et al., 2024b), TimeBase (Huang et al., 2025), MixLinear (Ma et al., 2026), DiPE-Linear (Zhao et al., 2026), PhaseFormer (Niu et al., 2026), and CycleNet/Linear (Lin et al., 2024a). We use published baseline configurations (Appendix D.3) and report StateComp's means and sample standard deviations over five seeds. Table 7 lists MSE for all datasets and horizons, using $L = 720$ for StateComp.

5.2 Accuracy and parameter efficiency

StateComp uses 61 to 389 learned scalars, the fewest among the models with reported counts in Figure 2. The figure compares MSE across twenty forecasting conditions; Figure 4 gives MAE. All coefficients are shared across variables, so the counts also apply to Traffic's 862 variables.

Accuracy with a shared, compact model. StateComp has lower MSE than SparseTSF and MixLinear at every horizon on Electricity and Traffic. At horizon 720 on Traffic, it obtains MSE 0.444 with 87 parameters, compared with 0.448 and 925 parameters for SparseTSF, and 0.452 and 293 parameters for MixLinear. CycleNet/Linear reaches MSE 0.442 with 663,936 parameters, while PhaseFormer reports 0.428. StateComp uses fewer than 100 parameters for competitive accuracy. On ETTh2 at horizon 336, StateComp matches TimeBase's MSE of 0.358 with 71 parameters versus 284. On ETTh1 at the same horizon, the two models obtain MSE 0.422 and 0.408, respectively.

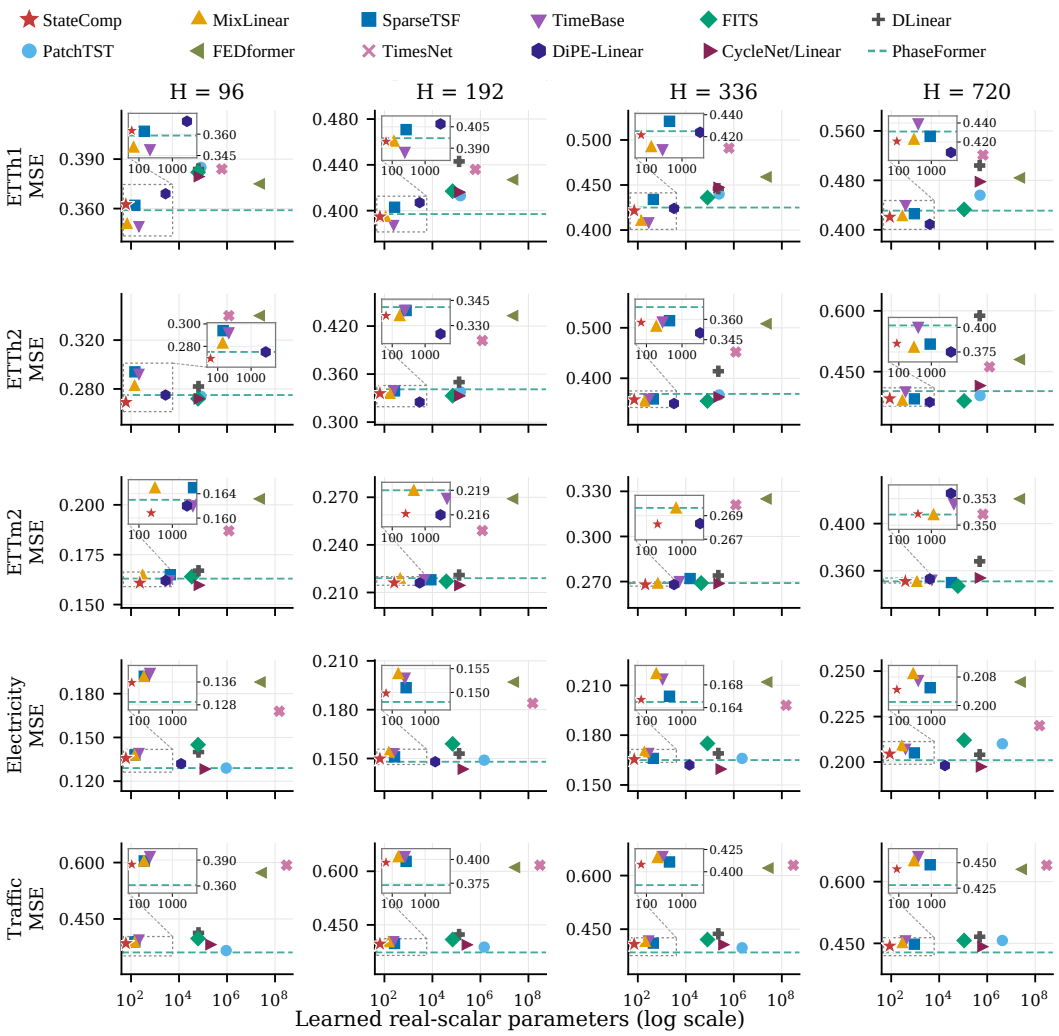


Figure 2: MSE and learned parameter count across datasets (rows) and horizons (columns). Lower and leftward points indicate smaller errors and models; insets enlarge the region below 5K parameters on the log axis. StateComp points are five-seed means. Dashed lines show PhaseFormer's reported MSE without a parameter-axis position: counts for its published accuracy configurations are unspecified (Appendix D.3). DiPE-Linear reports results for the ETT datasets and Electricity in its paper.

Minute-level forecasting. On ETTm2's 15-minute observations, validation selects the period and capacity using the same model formulation (Appendix D.2). At horizon 336, StateComp reaches MSE 0.268 with 205 parameters; MixLinear uses 681 parameters for MSE 0.269. At horizon 720, both report MSE 0.351, with 389 parameters for StateComp and 1,177 for MixLinear. Together, these comparisons support the separation between forecast detail and learned capacity. The input states retain the shape and scale specific to each series, while the learned coefficients determine how those states propagate into the forecast. The same compact transfer rule can therefore serve datasets with very different numbers of variables, and adapting the period extends this principle to finer sampling rates while preserving the same shared forecasting mechanism across temporal resolutions.

5.3 Ablation Study

Table 2 compares seasonal averages with fixed and learned readouts. The *fixed-coefficient* variant assigns unit gains to the widest base profile and longer-cycle contrast and zero to all other coefficients. The hourly *state-only* model has 27 coefficients for readout, filtering, and boundary alignment.

Table 1: Seasonal representations at $L = 720$. The shared map uses 8, 13, 14, and 28 parameters by horizon; state-only uses 27 and the full model uses 61 to 87. Five-seed means; state-only omits SD.

	H	State-only		Shared weekly map		Full model	
		MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	96	0.273	0.331	0.292 ± 0.001	0.354 ± 0.001	0.269 ± 0.001	0.331 ± 0.000
	192	0.341	0.375	0.351 ± 0.002	0.390 ± 0.001	0.336 ± 0.003	0.373 ± 0.001
	336	0.371	0.402	0.377 ± 0.000	0.411 ± 0.000	0.358 ± 0.002	0.395 ± 0.001
	720	0.387	0.424	0.400 ± 0.002	0.435 ± 0.001	0.384 ± 0.001	0.422 ± 0.000
Electricity	96	0.137	0.233	0.165 ± 0.000	0.259 ± 0.000	0.136 ± 0.000	0.232 ± 0.000
	192	0.151	0.245	0.170 ± 0.000	0.263 ± 0.000	0.150 ± 0.000	0.244 ± 0.000
	336	0.166	0.261	0.182 ± 0.000	0.275 ± 0.001	0.165 ± 0.000	0.260 ± 0.000
	720	0.206	0.295	0.220 ± 0.000	0.308 ± 0.001	0.204 ± 0.000	0.294 ± 0.001

Table 2: MSE at $L = 720$. B7 and BW are base cycle means over seven cycles and the widest window, respectively; CM is the longer cycle mean, Fix the fixed coefficient readout, SO the learned seasonal readout, and Full SO with history correction. Fixed predictors use 0 parameters, SO uses 27, and N_θ counts Full parameters. Learned entries are five seed means; sample SD is shown for Full, and dashes mark settings not evaluated.

Dataset	H	N_θ	B7	BW	CM	Fix	SO	Full
ETTh2	96	61	0.350	0.350	0.402	0.399	0.273	$\mathbf{0.269} \pm 0.001$
	192	65	0.401	0.363	0.399	0.397	0.341	$\mathbf{0.336} \pm 0.003$
	336	71	0.410	0.370	0.402	0.399	0.371	$\mathbf{0.358} \pm 0.002$
	720	87	0.420	0.399	0.429	0.427	0.387	$\mathbf{0.384} \pm 0.001$
ETTm2	96	233	0.234	0.316	0.281	0.286	–	$\mathbf{0.161} \pm 0.001$
	192	257	0.302	0.340	0.308	0.312	–	$\mathbf{0.216} \pm 0.001$
	336	205	0.343	0.371	0.341	0.344	–	$\mathbf{0.268} \pm 0.000$
	720	389	0.473	0.423	0.397	0.396	–	$\mathbf{0.351} \pm 0.001$
Traffic	96	61	0.623	0.622	0.440	0.432	0.387	$\mathbf{0.385} \pm 0.000$
	192	65	0.646	0.631	0.437	0.429	0.399	$\mathbf{0.397} \pm 0.001$
	336	71	0.664	0.643	0.444	0.437	0.410	$\mathbf{0.408} \pm 0.000$
	720	87	0.689	0.678	0.489	0.480	0.446	$\mathbf{0.444} \pm 0.000$

Learning the shared seasonal coefficients improves on fixed gains at every horizon on ETTh2 and Traffic, and the full model does so on ETTm2. At $H = 336$ on Traffic, the state-only model reaches 0.410 with 27 parameters, compared with SparseTSF's 0.411 with 445. Full results appear in Appendix G.1. At $H = 336$, the history correction lowers ETTh2 MSE from 0.371 to 0.358. Adding seasonal states to the history-only model lowers Traffic MSE from 0.632 to 0.408 for that horizon (Appendix G.2). These results assign distinct roles to the two branches. Seasonal averaging supplies a useful forecast before any coefficient is fitted, and the small learned readout adjusts how much the different widths and longer-cycle contrasts persist. The history correction then improves this forecast without relearning its seasonal shape. Its modest gain over the state-only model, together with the larger gain from adding states to the history-only model on Traffic, identifies the seasonal branch as the main representation and the history branch as a residual correction.

5.4 Capacity analysis and hyperparameter study

History rank. With the complete seasonal bank, rank one already reaches MSE 0.358 on ETTh2 at $H = 336$, the same value as ranks two and three (Figure 3). On Traffic, increasing rank from one to four changes MSE from 0.408 to 0.407 while increasing the parameter count from 71 to 203. Appendix E reports the full rank sweep and additional bank variants for all four horizons.

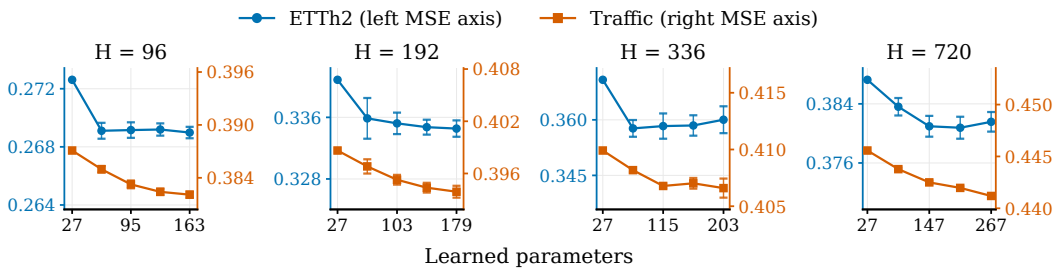


Figure 3: History-rank analysis with the complete seasonal bank. Five points per curve correspond to $r = 0, 1, 2, 3, 4$. Independently scaled left and right axes give ETTh2 and Traffic MSE, respectively. Points show five-seed means for each horizon; error bars show sample SD except at $r = 0$.

Seasonal representation. We compare the seasonal readout with a shared linear map over 168-position cycles, including partial cycles and four preprocessing scalars. The seasonal readout improves both metrics on both datasets, while history factors further lower MSE without degrading MAE.

Table 3: Base-period comparison on ETTm2 with seven-day histories ($L = 672$). Both settings use the same windows; entries are five-seed means $\pm$ sample SD, with lower means in bold.

Forecast	p	Scalars	MSE	MAE
1 day	24	59	0.177 ± 0.000	0.269 ± 0.001
	96	26	$\mathbf{0.165 \pm 0.000}$	$\mathbf{0.255 \pm 0.001}$
2 days	24	63	0.234 ± 0.001	0.308 ± 0.000
	96	27	$\mathbf{0.220 \pm 0.000}$	$\mathbf{0.291 \pm 0.000}$

Period adaptation. Matching ETTm2's daily period lowers MSE and MAE for one- and two-day forecasts while reducing the models from 59 to 26 and 63 to 27 parameters, respectively (Table 3).

Sensitivity to fixed hyperparameters. At $H = 336$, we vary phase smoothing, filter support, temporal decay, and initialization ridge strengths at fixed model size and learning rate. Across the tested values, mean MSE differs from the default by at most 0.002 and mean MAE by at most 0.001. Appendix I.2 gives the settings and five-seed error curves for each tested hyperparameter setting.

Alignment with a fitted temporal predictor. At $H = 336$, a rank correction reduces the held-out error of the mean-based seasonal approximation on all five datasets (Table 4). Both models target the same fitted cycle-to-cycle predictor, so this diagnostic measures representation error; Appendix J relates it to forecast MSE. Together with the rank, period, and weekly-map controls, the result shows that matching the temporal structure matters more than adding correction rank: rank one captures most residual benefit, and performance remains stable across the tested fixed constants.

Table 4: Temporal-predictor approximation at $H = 336$ (units: 10^{-3}). $\mathcal{E}$ is reference energy outside the constant and decaying span; $\mathcal{D}_A$ and $\mathcal{D}_{A,r}$ are held-out errors before and after rank correction.

Dataset	$\mathcal{E}$	$\mathcal{D}_A$	$\mathcal{D}_{A,r}$
ETTh1	0.650	0.534	0.323
ETTh2	2.495	2.250	1.551
ETTm2	32.368	2.521	1.356
Electricity	38.795	1.405	0.836
Traffic	209.704	8.279	5.774

6 Conclusion

StateComp separates input-derived seasonal detail from learned transfer. Multiwidth states preserve phase and longer-cycle patterns; shared coefficients project them forward, and a low-rank history map corrects residual dependence. In a stylized population model, one shared gain transfers several predictive directions already covered by the states, while correction rank controls the remaining error. Across five benchmarks, StateComp achieves competitive accuracy with 61–389 parameters and lower MSE than SparseTSF and MixLinear at every tested horizon on Electricity and Traffic. Ablations support the distinct roles of both branches. The method relies on informative periods and averaging widths; adapting its state bank to changing or weak periodicity is an important next step.

AI USE STATEMENT

Generative AI tools were used to help refine the conceptual framing and method design, develop mathematical derivations and proofs, implement and edit research code, interpret experimental results, draft and revise manuscript text, and prepare the method schematic. The authors reviewed and revised all AI assisted material included in the submission and take responsibility for the final content, including its claims, proofs, code, results, text, and figures.

REFERENCES

Kasun Bandara, Rob J. Hyndman, and Christoph Bergmeir. MSTL: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns. *International Journal of Operational Research*, 52(1):79–98, 2025. doi: 10.1504/IJOR.2025.143957.

Shane Barratt, Yining Dong, and Stephen Boyd. Low Rank Forecasting. arXiv:2101.12414, 2021. URL https://arxiv.org/abs/2101.12414.

Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 6989–6997, 2023.

Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. LLM4TS: Two-stage fine-tuning for time-series forecasting with pre-trained LLMs. arXiv:2308.08469v2, 2023. URL https://arxiv.org/abs/2308.08469v2.

Peng Chen, Yingying ZHANG, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 10527–10545, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/2be6705de7412adf107900add727a795-Paper-Conference.pdf.

Si-An Chen, Chun-Liang Li, Sercan O Arik, Nathanael Christian Yoder, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecast-ing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=wbpxTuXgm0.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan K Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=pCbC3aQB5W.

Alysha M. De Livera, Rob J. Hyndman, and Ralph D. Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011. doi: 10.1198/jasa.2011.tm09771.

Prasanjit Dey, Soumyabrata Dev, and Bianca Schoen-Phelan. One-for-All: A lightweight stabilized and parameter-efficient pre-trained LLM for time series forecasting. arXiv:2603.29756, 2026. URL https://arxiv.org/abs/2603.29756.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. doi: 10.1007/BF02288367.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=GUDnecJdJU.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Qihe Huang, Zhengyang Zhou, Kuo Yang, Zhongchao Yi, Xu Wang, and Yang Wang. Timebase: The power of minimalism in efficient long-term time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=GhTdNOMfOD.

Rob J. Hyndman, Anne B. Koehler, Ralph D. Snyder, and Simone Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18(3):439–454, 2002. doi: 10.1016/S0169-2070(01)00110-8.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(75)90042-1. URL https://www.sciencedirect.com/science/article/pii/0047259X75900421.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

SangJong Lee, Jin-Kwang Kim, JunHo Kim, TaeHan Kim, and James Lee. HiPPO-KAN: Efficient KAN model for time series analysis. In *ACM ICAIF 2024: From Prototype to Production: Deploying Real-World AI / ML Models in the Financial Industry*, 2024. URL https://openreview.net/forum?id=V2tP460rFc.

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

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. SparseTSF: Modeling long-term time series forecasting with *1k* parameters. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=54NSHO0lFe.

Xiachong Lin, Du Yin, Hao Xue, Wen Hu, Imran Razzak, Arian Prabowo, Matthew Amos, and Flora D. Salim. AsyTO: Asymmetric temporal operator for parameter-efficient multivariate time series forecasting. arXiv:2608.16098, 2026. URL https://arxiv.org/abs/2608.16098.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Aitian Ma, Dongsheng Luo, and Mo Sha. Mixlinear: Extreme low resource multivariate time series forecasting with $0.1k$ parameters. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QUj0KuCumD.

Heitor R. Medeiros, Hossein Sharifi-Noghabi, Gabriel L. Oliveira, and Saghar Irandoust. Accurate parameter-efficient test-time adaptation for time series forecasting. In *Second Workshop on Test-Time Adaptation at ICML*, 2025. URL https://openreview.net/forum?id=uFj4EL4GTB.

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

Yiming Niu, Jinliang Deng, and Yongxin Tong. Phaseformer: From patches to phases for efficient and effective time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Lk9SqMQzhX.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31, 2018.

CLEVELAND RB. Stl: A seasonal-trend decomposition procedure based on loess. *J Off Stat*, 6: 3–73, 1990.

William Toner and Luke Nicholas Darlow. An analysis of linear time series forecasting models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xl82CcbYaT.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Kaibin Wei, Jianqiang Jing, Jiawei Liu, Qing Liu, and Xiannian Xie. Sample-efficient fine-tuning with textual prompts for time series forecasting. *PLOS One*, 21(7):e0353130, 2026. doi: 10.1371/journal.pone.0353130.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=J4gRj6d5Qm.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Xingjian Wu, Xiangfei Qiu, Hanyin Cheng, Zhengyu Li, Jilin Hu, Chenjuan Guo, and Bin Yang. Enhancing time series forecasting through selective representation spaces: A patch perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=BirE0jYKt0.

Longlong Xu, Zeyan Li, Xiao He, Zhaoyang Yu, Changhua Pei, Zhe Xie, Zijun Dou, Tieying Zhang, and Dan Pei. Taming the recent-data bias: Towards robust time series forecasting with global context. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=gDA03Yn3fi.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with \$10k\$ parameters. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=bWcnvZ3qMb.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain MLPs are more effective learners in time series forecasting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=iif9mGCTfy.

Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/85422afb467e9456013a2a51d4dff702-Paper.pdf.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Jiahui Zhang, Zhengyang Zhou, Wenjie Du, and Yang Wang. Enhancing the maximum effective window for long-term time series forecasting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Gmwsy7TlFI.

Yuang Zhao, Tianyu Li, Jiadong Chen, Shenrong Ye, Fuxin Jiang, and Xiaofeng Gao. Disentangled parameter-efficient linear model for long-term time series forecasting. In *International Conference on Database Systems for Advanced Applications*, pp. 222–238. Springer, 2026.

Liyang Zheng and Baisuo Jin. TS-PET: A novel framework for fine-tuning pretrained time-series models. *Big Data*, 2026. doi: 10.1177/2167647X261439005. URL https://journals.sagepub.com/doi/10.1177/2167647X261439005.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

A Model and optimization details

Each variable uses the definitions below with shared learned coefficients. The hyperparameters are the base period p, cycle multiplier g, valid width sets $\mathcal{W}, \mathcal{G}$, and factor width r. Empty width sets or $r = 0$ remove the corresponding branch where defined. Appendix D lists the experimental settings.

A.1 Filtering, profiles, and contrasts

Set $\mu = L^{-1} \sum_{t=0}^{L-1} X_t$ and $Z_t = X_t - \mu$. For an odd positive width κ, define

$$\mathrm{Avg}_\kappa(Z)_t = \frac{1}{\kappa} \sum_{u=-(\kappa-1)/2}^{(\kappa-1)/2} \bar{Z}_{t+u}, \quad \bar{Z}_t = \begin{cases} Z_t, & 0 \le t < L, \\ 0, & \text{otherwise.} \end{cases}$$

The denominator remains κ at the boundary. The filtered history is

$$Q = Z + s_1 \,\mathrm{Avg}_{\kappa_1}(Z) + s_2(\mathrm{Avg}_{\kappa_2}(Z) - Z), \qquad \mathsf{T}_{ij} = Q_{ip+j}.$$

The two filter widths are fixed; s_1, s_2 are learned. For each base width w, the profile and level are

$$F_w(j) = \frac{1}{w} \sum_{i=n-w}^{n-1} \mathsf{T}_{ij}, \qquad \ell_w = \frac{1}{p} \sum_j F_w(j).$$

The base-cycle component Y^{b} follows Eq. (1) with $\phi_0(k) = 1$ and $\phi_1(k) = \exp[-k/\max(m/2, 1)]$.

For each longer width v, use the latest v complete groups of g cycles:

$$R_v(d, j) = \frac{1}{v} \sum_{a=1}^{v} \mathsf{T}_{n-ag+d,j}, \qquad 0 \le d < g, \quad 0 \le j < p.$$

If a designated robust width v_{rob} belongs to the bank, replace its mean by

$$R_{v_{\mathrm{rob}}}(d, j) \leftarrow R_{v_{\mathrm{rob}}}(d, j) + \bar{\rho}\,[\mathrm{med}_{1 \le a \le v_{\mathrm{rob}}} \mathsf{T}_{n-ag+d,j} - R_{v_{\mathrm{rob}}}(d, j)], \quad \bar{\rho} = \mathrm{clip}_{[0,1]}(\rho).$$

The median is componentwise and averages the two central observations for an even number of groups. The learned scalar ρ is present only with this group. For a fixed smoothing strength $\alpha \in [0, 1]$, set

$$\widetilde{R}_v(d, j) = (1 - \alpha) R_v(d, j) + \frac{\alpha}{3}[R_v(d, j-1) + R_v(d, j) + R_v(d, j+1)],$$

with phase indices interpreted modulo p. The routed contrast is

$$P_v(k, j) = \widetilde{R}_v(k \bmod g, j) - \frac{1}{g} \sum_{d=0}^{g-1} \widetilde{R}_v(d, j).$$

Future cycle zero aligns with position zero of the latest complete group. The observed window determines all cycle positions and seasonal states without external calendar covariates.

A.2 Boundary alignment and parameter sharing

Choose $1 \le e \le p$. The quantities $\bar{X}_e$ and $\bar{Y}^{\mathrm{b}}_e$ average the last e observed values and the last e phases of the first cycle of Y^{b}, respectively. For fixed bounds $\zeta_- < \zeta_+$ and learned β, ζ, set

$$d_b(h) = \exp\left[-\frac{h+1}{p \exp(\mathrm{clip}_{[\zeta_-, \zeta_+]}(\zeta))} \right].$$

The boundary reference is calculated before adding the alignment term, longer-cycle contrasts, or history correction. Equations (3) and (4) then give the full forecast. The coefficient shapes are $A, D \in \mathbb{R}^{2 \times J}$, $\Gamma \in \mathbb{R}^{2 \times K}$, $U \in \mathbb{R}^{r \times n}$, and $V \in \mathbb{R}^{m \times r}$; they do not depend on the variable count C.

B A shared-state oracle and a realizable calendar example

A seasonal state represents predictive directions with shared gains. Conditional-mean orthogonality and best-rank approximation give its unpenalized population risk (Eckart & Young, 1936).

B.1 JOINT OPTIMIZATION OF THE SHARED READOUT AND CORRECTION

Let q, v, t be positive integers. For direction $i \in \{1, \dots, q\}$, let $z_i \in \mathbb{R}^v$ be a centered history vector, with $\mathbb{E}[z_i z_j^\top] = \mathbf{1}\{i = j\}G$ and $G \succ 0$. Writing z for the stacked histories, suppose the stacked future satisfies

$$y_i = hb^\top z_i + \varepsilon_i \in \mathbb{R}^t, \quad h \neq 0, \quad \mathbb{E}[\varepsilon \mid z] = 0, \quad R_* = \mathbb{E}\|\varepsilon\|_2^2 < \infty.$$

Conditioning on the entire stacked history makes R_* the Bayes risk. A fixed history dictionary $A \in \mathbb{R}^{J \times v}$ and output basis $\Phi \in \mathbb{R}^{t \times d}$ define the shared-state class

$$\widehat{y} = (I_q \otimes \Phi CA)z + Kz, \quad C \in \mathbb{R}^{d \times J}, \quad K \in \mathbb{R}^{qt \times qv}, \quad \operatorname{rank} K \leq r,$$

where $r \in \{0, \dots, q-1\}$ and $h \in \operatorname{col} \Phi$. The correction K can couple all direction histories through a single matrix. Define

$$\delta_A^2 = \min_{w \in \operatorname{row} A} (b - w)^\top G (b - w), \qquad \kappa = b^\top G b.$$

Proposition 1 (Shared-dictionary oracle)**.** *Jointly optimizing the shared readout and correction gives*

$$\inf_{C,\, \operatorname{rank} K \leq r} \mathbb{E}\|y - \widehat{y}\|_2^2 - R_* = (q - r)\|h\|_2^2 \delta_A^2. \tag{7}$$

The best pure rank-k linear predictor has excess risk $(q-k)_+\|h\|_2^2\kappa$ for integer $k \geq 0$.

Proof. Let $\Sigma = I_q \otimes G$ and $T = I_q \otimes hb^\top$. Orthogonality of the conditional-mean residual gives excess risk $\|[T - I_q \otimes \Phi CA - K]\Sigma^{1/2}\|_F^2$. For fixed C, put $D = (hb^\top - \Phi CA)G^{1/2}$. Each singular value of D occurs q times in $I_q \otimes D$. Right multiplication by the invertible $\Sigma^{1/2}$ preserves rank and maps the feasible K onto all rank-at-most-r matrices. As $r < q$, best-rank approximation eliminates K with value

$$q\|D\|_F^2 - r\sigma_1(D)^2 \geq (q - r)\|D\|_F^2 \geq (q - r)\|h\|_2^2 \delta_A^2.$$

The last inequality projects each row onto $\operatorname{row}(AG^{1/2})$. Let w_A be the G-orthogonal projection of b onto $\operatorname{row} A$. Since $h \in \operatorname{col} \Phi$, some C realizes $\Phi CA = hw_A^\top$. Then D has rank at most one and $\|D\|_F^2 = \|h\|_2^2 \delta_A^2$, giving equality in both inequalities. The whitened pure-class target has q equal nonzero singular values $\|h\|_2\sqrt{\kappa}$ when $b \neq 0$; truncation gives the claim, also for $b = 0$. □

Dividing these total squared-error risks by qt gives coordinate MSE. The result depends on the row space of A and column space of Φ, so dependent rows or columns are allowed. The stated second moments and conditional-mean assumption suffice, including when direction histories are dependent.

If $b \in \operatorname{row} A$, the shared state reaches the Bayes predictor using dJ shared coefficients. When $\delta_A > 0$, a rank-r correction leaves $q - r$ copies of the common omission within this linear comparison class.

B.2 A STATIC GAUSSIAN CALENDAR CONSTRUCTION

Take $g \geq 2$, $q = g - 1$, and an orthonormal contrast matrix $S \in \mathbb{R}^{g \times q}$ with $S^\top \mathbf{1}_g = 0$. Let $a \sim N(0, I_q)$, and observe $z_j = a + \eta_j$ for $j = 1, \dots, v$, with independent $\eta_j \sim N(0, \sigma^2 I_q)$, $\sigma^2 > 0$. Embed Sz_j in the latest v complete aligned groups of g input cycles, setting any earlier cycles to zero. Each phase within a cycle has the same value. For $m = tg$ future cycles, repeat Sa as the target in each group. Each group sums to zero, so input centering leaves the construction unchanged.

For each contrast direction, the history covariance, conditional-mean coefficient, and output direction are

$$G = \mathbf{1}_v \mathbf{1}_v^\top + \sigma^2 I_v, \qquad b = \frac{\mathbf{1}_v}{v + \sigma^2}, \qquad h = \mathbf{1}_t.$$

Joint Gaussianity gives $\mathbb{E}[a \mid z_1, \dots, z_v] = \sum_j z_j/(v + \sigma^2)$. The width-$v$ longer-cycle state equals $S\bar{z}$, where $\bar{z} = v^{-1}\sum_j z_j$. Its constant readout gain $\gamma = v/(v + \sigma^2)$ realizes the conditional mean. Set base, robust-blend, decay, filter, boundary, and history-factor parameters to zero. Zero filter gains leave the identity preprocessing path; circular smoothing leaves phase-constant values unchanged.

The direction coordinates lie within one cycle-to-cycle matrix. Up to a permutation, $I_v \otimes S$ and $I_t \otimes S$, together with historical zero padding, give isometric embeddings E_{in} and E_{out}. A cycle correction

W induces $\widetilde{W} = E_{\text{out}}^\top W E_{\text{in}}$ with $\operatorname{rank}\widetilde{W} \le \operatorname{rank} W$. Conversely, $W = E_{\text{out}}\widetilde{W}E_{\text{in}}^\top$ lifts any coordinate correction with the same rank. Thus all contrast directions occupy a single cycle-to-cycle operator within each variable and phase, over the historical and future cycle coordinates.

Here $\delta_A = 0$ when the dictionary contains the width-v average, and $\kappa = v/(v+\sigma^2)$. The target lies entirely in the retained calendar subspace. Total Bayes risk is $tq\sigma^2/(v+\sigma^2)$; the pure cycle rank-r class has additional risk $(q-r)_+ tv/(v+\sigma^2)$. Returning to the original tg cycle coordinates gives

$$\text{MSE}_{\text{Bayes}} = \frac{q}{g}\frac{\sigma^2}{v+\sigma^2}, \qquad \text{MSE}_{\text{rank-r}} - \text{MSE}_{\text{Bayes}} = \frac{(q-r)_+}{g}\frac{v}{v+\sigma^2}.$$

Replicating p identical phases multiplies both total error and coordinate count by p, preserving MSE. The formula applies to $m = tg$ forecast cycles covering complete future groups of the longer cycle.

B.3 A DYNAMIC GAUSSIAN EXAMPLE

A non-static Gaussian example takes independent latent AR(1) directions with stationary variance one, correlation $0 < |\rho| < 1$, Gaussian innovations of variance $1-\rho^2$, and independent Gaussian observation noise of variance one. For v observed states, $G_{ab} = \rho^{|a-b|} + \mathbf{1}\{a=b\}$, $b = G^{-1}(\rho^{v-1}, \dots, 1)^\top$, and $h = (\rho, \dots, \rho^t)^\top$. Gaussianity identifies this best linear predictor with the conditional mean. At $\rho = 1$, the stationary boundary is a zero-innovation random constant Gaussian process across time.

For $v = 4, q = 6, t = 2, r = 1$, $\rho = 1/2$, and averaging widths $\{1, 2, 4\}$, one obtains $b = (2, 7, 26, 97)^\top/209$, $\delta_A^2 = 25/58938$, and total excess $625/943008$. The coordinate excess is $625/11316096$ for this linear shared-dictionary class. StateComp adds median blending, learned filtering, and boundary alignment to the seasonal readout and rank correction used in this comparison.

C STRUCTURED TRANSFER AND LOW-RANK HISTORY

We compare prescribed transfer with learned rank in a deterministic setting (Section 3.4). Throughout this section, $B \in \mathbb{R}^{m\times n}$ is a prescribed transfer operator, $Q_0 \in \mathbb{R}^{m\times d}$ spans a prescribed output space, and $P_\perp = I - Q_0 Q_0^\dagger$ is the orthogonal projector onto its complement. The coefficient matrix $C \in \mathbb{R}^{d\times n}$ is unrestricted. Write $t_k(M) = (\sum_{j>k}\sigma_j(M)^2)^{1/2}$. By the best-rank approximation identity of Eckart & Young (1936), this is M's Frobenius distance to matrices of rank at most k.

Proposition 2 (Separation of a prescribed transfer)**.** *Let $A_i = Q_0 C_i + \gamma_i B + W_i$, with $\operatorname{rank}(W_i) \le r$. Then*

$$|\gamma_1 - \gamma_0|\, t_{2r}(P_\perp B) \le \|P_\perp(A_1 - A_0)\|_F.$$

For every operator in this class, γ and $P_\perp W$ are uniquely determined by A if and only if $\operatorname{rank}(P_\perp B) > 2r$.

Proof. Projection removes the nuisance term, giving

$$P_\perp(A_1 - A_0) = (\gamma_1 - \gamma_0)P_\perp B + P_\perp(W_1 - W_0).$$

The second term has rank at most $2r$. When $\gamma_1 \ne \gamma_0$, dividing by the gain difference and applying the best-rank approximation identity gives the inequality. When the gains agree, the inequality is immediate. If the tail is positive and $A_1 = A_0$, both gains and projected corrections agree.

For necessity, suppose $\operatorname{rank}(P_\perp B) \le 2r$. Partition its singular components into $P_\perp B = B_1 + B_2$, where each part has rank at most r. The operator $A = B_1$ has two representations:

$$(C, \gamma, W) = (0, 0, B_1), \qquad (C, \gamma, W) = (-Q_0^\dagger B, 1, -B_2).$$

Both are feasible and their gains differ. Thus uniform identification fails. □

Let $\kappa = t_{2r}(P_\perp B) > 0$. The same argument and the triangle inequality also give

$$\|P_\perp(W_1 - W_0)\|_F \le \left(1 + \frac{\|P_\perp B\|_F}{\kappa}\right)\|P_\perp(A_1 - A_0)\|_F.$$

For an operator estimate $\widetilde{A}$ satisfying $\|\widetilde{A} - A_0\|_F \le \epsilon$, any feasible fitted operator $\widehat{A}$ with $\|\widetilde{A} - \widehat{A}\|_F \le \eta$ therefore obeys $|\widehat{\gamma} - \gamma_0| \le (\epsilon + \eta)/\kappa$. These bounds identify the gain and projected correction matrix. Its factors retain their usual change-of-basis freedom.

Proposition 3 (Approximation by a learned low-rank map)**.** *For $A_\star = Q_0C_0 + \gamma B + W_0$ with $\mathrm{rank}(W_0) \le r$, every matrix $Q_0C + L$ with $\mathrm{rank}(L) \le \ell$ satisfies*

$$\|A_\star - (Q_0C + L)\|_F \ge |\gamma|\, t_{r+\ell}(P_\perp B).$$

Proof. Projection cannot increase the Frobenius norm. After projection, the error is $\gamma P_\perp B + P_\perp(W_0 - L)$, whose second term has rank at most $r + \ell$. Apply best-rank approximation. □

If a random input $z \in \mathbb{R}^n$ has second moment $G = \mathbb{E}[zz^\top] \succeq \mu I$ for $\mu > 0$, the corresponding prediction discrepancy satisfies

$$\mathbb{E}\|[A_\star - (Q_0C + L)]z\|_2^2 \ge \mu\gamma^2 t_{r+\ell}(P_\perp B)^2.$$

For mean squared error per output coordinate, divide this bound by m. The bound compares the prescribed transfer with predictors that use the same nuisance space and a learned low-rank map.

A calendar-contrast instance. Let $n \ge g$ and $m \ge g + 1$. Let $E_g \in \mathbb{R}^{g\times n}$ select the latest g cycle coordinates and let row k of $R_g \in \mathbb{R}^{m\times g}$ be $e_{k \bmod g}^\top$. Define

$$B = R_g \left(I_g - \frac{\mathbf{1}_g\mathbf{1}_g^\top}{g} \right) E_g, \qquad Q_0 = [\mathbf{1}_m, d], \quad d_k = \exp(-k/\tau), \quad \tau > 0.$$

The first g rows of R_g contain every calendar position, so B has rank $g - 1$. Every vector in its range is g-periodic with zero sum over a full period. If such a vector equals $a\mathbf{1}_m + bd$, comparison of entries 0 and g gives $b(1 - \exp(-g/\tau)) = 0$, hence $b = 0$. Its zero sum then gives $a = 0$. Thus $\mathrm{range}(B)$ intersects $\mathrm{range}(Q_0)$ only at zero, and projection preserves its rank:

$$\mathrm{rank}(P_\perp B) = g - 1.$$

Consequently, $g - 1 > 2r$ gives the separation condition in Proposition 2. This instance connects to the forecasting family by taking phase-constant inputs, zero filter and boundary gains, and one width-one longer-cycle contrast with a constant readout gain. The fixed routing has rank $g - 1$ with one learned gain; a free factorized correction of width r stores $r(n + m)$ scalars. This distinction explains how prescribed seasonal structure can preserve several transfer directions at a small coefficient cost.

D Experimental details

D.1 Data and evaluation

ETTh1, ETTh2, and ETTm2 use chronological 12/4/4-month benchmark partitions. Electricity and Traffic use chronological 70/10/20% partitions. Standardization is fitted to the training segment. MSE and MAE are averaged over all variables, forecast steps, and valid evaluation windows in the standardized scale. Table 5 summarizes the datasets.

Table 5: Benchmark dataset statistics.

Dataset	ETTh1	ETTh2	ETTm2	Electricity	Traffic
Variables	7	7	7	321	862
Observations	17,420	17,420	69,680	26,304	17,544
Frequency	Hourly	Hourly	15 min	Hourly	Hourly

D.2 Model configuration and optimization

All datasets use $L = 720$. Hourly models use $p = 24$, $g = 7$, base widths $(1, 2, 7, 28)$, longer-cycle widths $(1, 2, 4)$, and $r = 1$, so each input window forms a 30×24 cycle–phase array. The robust width is 4, filter widths are 25 and 3, phase smoothing uses $\alpha = 0.250$, the boundary reference uses six values, and the log boundary scale is clipped to $[-6, 6]$. ETTm2 uses a daily longer cycle and validation-selected settings for each horizon (Appendix H.1).

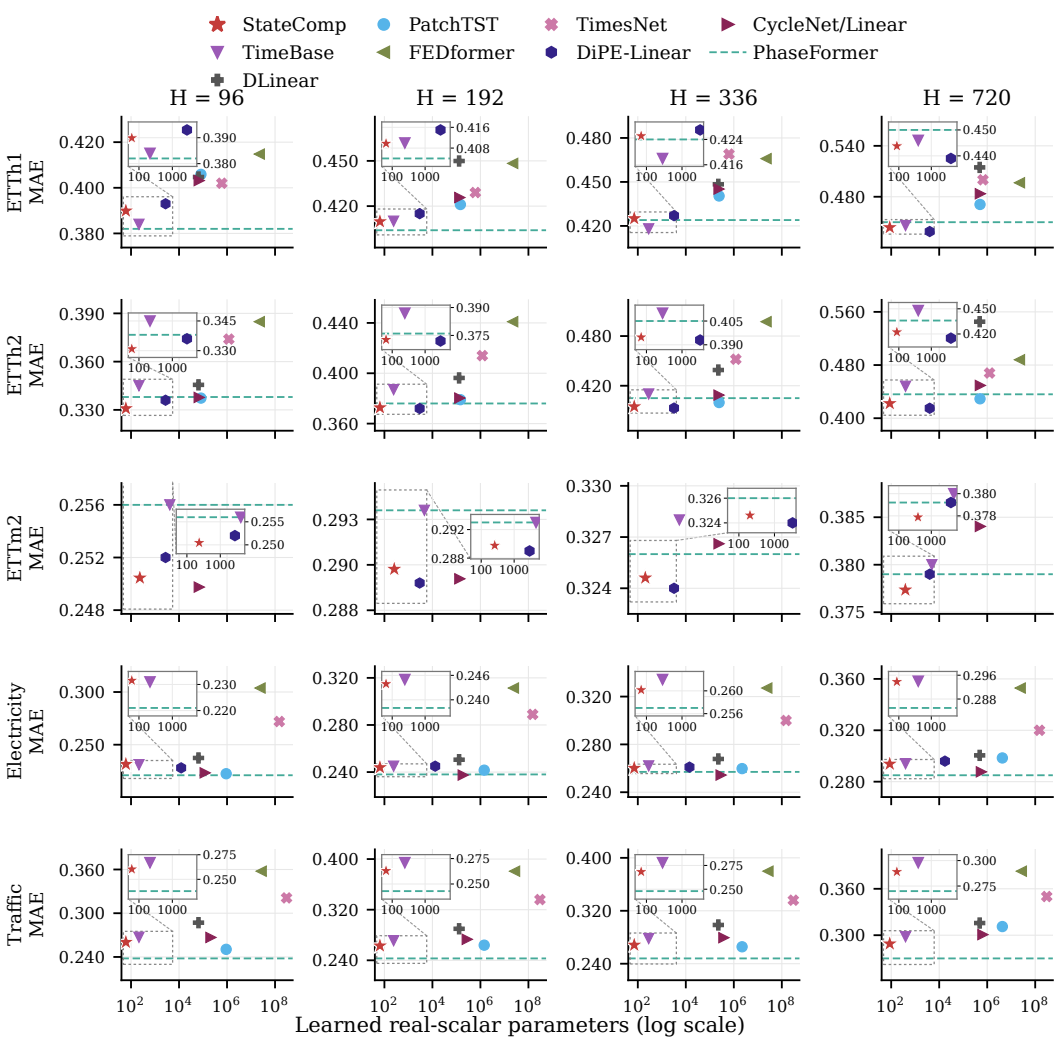


Figure 4: MAE and learned parameter count, using the log axes and insets of Figure 2. StateComp points are five-seed means. Dashed lines give PhaseFormer's MAE; counts for its accuracy configurations are unspecified (Appendix D.3). DiPE-Linear has no reported Traffic results.

We train all coefficients with Adam (Kingma & Ba, 2015) without weight decay, using batch size 64 for at most 30 epochs and early-stopping patience five. Validation MSE selects the learning rate from $\{0.005, 0.020, 0.030\}$ and the final predictor. After epoch four, the learning rate is multiplied by 0.800 per epoch. Table 6 lists the selected hourly rates. Ablations and representation controls use the same procedure; Appendix I.1 compares initializations.

Initialization. We initialize the state readout by ridge regression with the history map set to zero, then fit the history factors by reduced-rank regression on the residuals (Hoerl & Kennard, 1970; Izenman, 1975). These two updates alternate for 1,000 refinement cycles in the reported models. The relative state and history ridge strengths, λ_s^{rel} and λ_h^{rel}, are both 0.010. Filter and boundary coefficients start at zero, and the robust gain starts at 0.500. Validation MSE selects among the initialized and jointly trained predictors. Appendix I.1 reports the initialization ablation.

D.3 Baseline configurations

Baseline MSEs follow the full-test results reported by Lin et al. (2024b); Huang et al. (2025); Ma et al. (2026). For MAE, we use the corresponding published results and released experiment logs for

Table 6: Validation-selected learning rates for StateComp at each horizon.

Dataset	$H = 96$	$H = 192$	$H = 336$	$H = 720$
ETTh1	0.005	0.030	0.020	0.020
ETTh2	0.005	0.030	0.005	0.020
Traffic	0.030	0.030	0.005	0.005
Electricity	0.020	0.020	0.030	0.030

Table 7: Multivariate forecasting MSE. StateComp reports five-seed mean ± sample SD. Baselines use published results or author logs with their reported configurations. Dashes mark unreported DiPE-Linear Traffic results. Bold marks the lowest mean across all methods for each condition.

	H	FEDformer	TimesNet	PatchTST	DLinear	FITS	SparseTSF	TimeBase	MixLinear	DiPE	CycleNet	PhaseFormer	StateComp
ETTh1	96	0.375	0.384	0.385	0.384	0.382	0.362	**0.349**	0.351	0.369	0.379	0.359	0.362 ± 0.001
	192	0.427	0.436	0.413	0.443	0.417	0.403	**0.387**	0.395	0.407	0.416	0.397	0.395 ± 0.001
	336	0.459	0.491	0.440	0.446	0.436	0.434	**0.408**	0.411	0.424	0.447	0.425	0.422 ± 0.002
	720	0.484	0.521	0.456	0.504	0.433	0.426	0.439	0.423	**0.409**	0.478	0.431	0.421 ± 0.002
ETTh2	96	0.340	0.340	0.274	0.282	0.272	0.294	0.292	0.283	0.275	0.272	0.275	**0.269** ± 0.001
	192	0.433	0.402	0.338	0.350	0.333	0.339	0.339	0.336	**0.325**	0.333	0.341	0.336 ± 0.003
	336	0.508	0.452	0.367	0.414	0.355	0.359	0.358	0.355	**0.350**	0.363	0.369	0.358 ± 0.002
	720	0.480	0.462	0.391	0.588	0.378	0.383	0.400	0.380	**0.375**	0.416	0.402	0.384 ± 0.001
ETTm2	96	0.203	0.187	0.166	0.167	0.164	0.165	0.162	0.165	0.162	**0.160**	0.163	0.161 ± 0.001
	192	0.269	0.249	0.223	0.221	0.217	0.218	0.218	0.219	0.216	**0.214**	0.219	0.216 ± 0.001
	336	0.325	0.321	0.274	0.274	0.269	0.272	0.270	0.269	**0.268**	0.269	0.269	**0.268** ± 0.000
	720	0.421	0.408	0.362	0.368	**0.347**	0.350	0.352	0.351	0.353	0.354	0.351	0.351 ± 0.001
Electricity	96	0.188	0.168	0.129	0.140	0.145	0.138	0.139	0.138	0.132	**0.128**	0.129	0.136 ± 0.000
	192	0.197	0.184	0.149	0.153	0.159	0.151	0.153	0.154	0.148	**0.143**	0.148	0.150 ± 0.000
	336	0.212	0.198	0.166	0.169	0.175	0.166	0.169	0.170	0.162	**0.160**	0.165	0.165 ± 0.000
	720	0.244	0.220	0.210	0.204	0.212	0.205	0.207	0.209	0.198	**0.197**	0.201	0.204 ± 0.000
Traffic	96	0.573	0.593	0.366	0.413	0.398	0.389	0.394	0.389	–	0.382	**0.361**	0.385 ± 0.000
	192	0.611	0.617	0.388	0.423	0.409	0.398	0.403	0.403	–	0.394	**0.373**	0.397 ± 0.001
	336	0.621	0.629	0.398	0.437	0.421	0.411	0.417	0.416	–	0.407	**0.385**	0.408 ± 0.000
	720	0.630	0.640	0.457	0.466	0.457	0.448	0.456	0.452	–	0.442	**0.428**	0.444 ± 0.000

DLinear, PatchTST, FEDformer, TimesNet, and TimeBase. DiPE-Linear and PhaseFormer use their published forecasting tables; CycleNet/Linear uses the official five-seed results with input length 720.

We count trainable real scalars in the official implementations using the original papers' configurations. Each complex coefficient counts as two scalars; fixed buffers and non-trainable constants are excluded. DiPE-Linear counts follow its paper-specified configuration and released shared-weight tensors. Its paper does not report Traffic results, hence the missing entries in Table 7 and Figures 2 and 4.

PhaseFormer reports 1.156K parameters at $L = 720$, $H = 96$ in its efficiency table, while its released Traffic configurations contain 38,690 to 275,998 trainable scalars across horizons (Niu et al., 2026). The configurations used for its accuracy table are unspecified, so Figures 2 and 4 plot those errors as horizontal references without a parameter-axis position at each of the four forecast horizons.

Hardware and software. StateComp experiments were run on a single NVIDIA A100 GPU with 40 GB of memory. Each run uses one compute device and one PyTorch CPU thread, and the five seeds are executed sequentially without distributed training. The implementation uses Python and PyTorch. The released environment requires Python 3.10 or newer, PyTorch 2.0 or newer, NumPy 1.24 or newer, pandas 2.0 or newer, and scikit-learn 1.3 or newer; it also supports CPU execution.

D.4 Full forecasting results

Table 7 gives MSE for all twenty conditions, with five-seed means and sample SD for StateComp.

E Bank and rank variants

We vary state membership and history rank with budgets of 32, 64, and 128 learned scalars. At each rank $r \in \{0, 1, 2\}$, greedy search adds state groups according to the reduction in the training ridge objective per additional scalar, within the parameter budget. Each addition is evaluated by alternating state and history updates. The search stops when no feasible addition improves the objective, and the

lowest-objective structure across ranks is retained. Each variant uses the initialization and training procedure of Section D.2. The main hourly model uses seven groups at rank one.

Tables 8 through 9 report MSE and MAE as five-seed means for each parameter budget, with sample SD for nonzero ranks. Budgets with identical model and training configurations share a row.

Table 8: Bank and rank variants on ETTh1 (h1) and ETTh2 (h2). Nonzero-rank entries show five-seed mean ± sample SD; a slash joins identical configurations.

H	Budget	Rank	Scalars		LR		MSE		MAE	
			h1	h2	h1	h2	h1	h2	h1	h2
96	32	0	27	27	0.030	0.005	0.357	0.273	0.390	0.331
	64	1	61	61	0.005	0.005	0.362 ± 0.001	0.269 ± 0.001	0.390 ± 0.001	0.331 ± 0.000
	128	2	95	95	0.005	0.005	0.367 ± 0.001	0.269 ± 0.001	0.392 ± 0.001	0.331 ± 0.000
192	32	0	27	27	0.030	0.030	0.392	0.341	0.410	0.375
	64	1	61	63	0.030	0.030	0.396 ± 0.002	0.335 ± 0.002	0.410 ± 0.001	0.372 ± 0.001
	128	2	103	103	0.030	0.020	0.397 ± 0.002	0.335 ± 0.001	0.411 ± 0.001	0.372 ± 0.000
336	32	0	27	27	0.030	0.005	0.422	0.371	0.426	0.402
	64	1	63	64	0.030	0.005	0.421 ± 0.004	0.357 ± 0.001	0.425 ± 0.002	0.395 ± 0.000
	128	2	115	115	0.030	0.020	0.424 ± 0.004	0.358 ± 0.003	0.427 ± 0.002	0.395 ± 0.002
720	32/64	0	27	27	0.030	0.030	0.424	0.387	0.445	0.424
	128	1	87	87	0.020	0.020	0.421 ± 0.002	0.384 ± 0.001	0.444 ± 0.002	0.422 ± 0.000

Table 9: Bank and rank variants on Electricity and Traffic. Nonzero-rank entries show five-seed mean ± sample SD. Scalar counts are shared across datasets.

H	Budget	Rank	Scalars	LR		MSE		MAE	
				Electricity	Traffic	Electricity	Traffic	Electricity	Traffic
96	32	0	27	0.005	0.030	0.137	0.387	0.233	0.260
	64	1	61	0.020	0.030	0.136 ± 0.000	0.385 ± 0.000	0.232 ± 0.000	0.260 ± 0.001
	128	2	95	0.020	0.020	0.136 ± 0.000	0.383 ± 0.000	0.232 ± 0.001	0.259 ± 0.001
192	32	0	27	0.020	0.030	0.151	0.399	0.245	0.264
	64	1	61	0.020	0.030	0.149 ± 0.000	0.396 ± 0.000	0.243 ± 0.000	0.263 ± 0.001
	128	2	103	0.030	0.030	0.150 ± 0.000	0.395 ± 0.001	0.244 ± 0.000	0.263 ± 0.001
336	32	0	27	0.030	0.030	0.166	0.410	0.261	0.269
	64	1	63	0.020	0.005	0.166 ± 0.000	0.409 ± 0.000	0.261 ± 0.001	0.269 ± 0.001
	128	2	115	0.020	0.005	0.165 ± 0.000	0.407 ± 0.000	0.260 ± 0.001	0.268 ± 0.000
720	32/64	0	27	0.030	0.030	0.206	0.446	0.295	0.290
	128	1	87	0.030	0.005	0.204 ± 0.000	0.444 ± 0.000	0.294 ± 0.001	0.289 ± 0.001

E.1 History rank

We retain the complete seasonal bank and vary history rank from 0 to 4. Validation selects each rank's learning rate. Rank zero is state-only. Tables 10 and 11 give MSE and MAE for Figure 3.

Table 10: History rank on ETTh2 ($L = 720$). LP denotes learned parameters; MSE and MAE are five-seed means with sample SD for $r > 0$.

Rank	$H = 96$			$H = 192$			$H = 336$			$H = 720$		
	LP	MSE	MAE	LP	MSE	MAE	LP	MSE	MAE	LP	MSE	MAE
0	27	0.273	0.331	27	0.341	0.375	27	0.371	0.402	27	0.387	0.424
1	61	0.269 ± 0.001	0.331 ± 0.000	65	0.336 ± 0.003	0.373 ± 0.001	71	0.358 ± 0.002	0.395 ± 0.001	87	0.384 ± 0.001	0.422 ± 0.000
2	95	0.269 ± 0.001	0.331 ± 0.000	103	0.335 ± 0.001	0.372 ± 0.000	115	0.358 ± 0.003	0.395 ± 0.002	147	0.381 ± 0.001	0.421 ± 0.000
3	129	0.269 ± 0.000	0.332 ± 0.000	141	0.335 ± 0.001	0.372 ± 0.000	159	0.358 ± 0.003	0.395 ± 0.001	207	0.381 ± 0.001	0.421 ± 0.000
4	163	0.269 ± 0.000	0.332 ± 0.000	179	0.335 ± 0.001	0.372 ± 0.000	203	0.360 ± 0.004	0.396 ± 0.001	267	0.382 ± 0.001	0.422 ± 0.001

Table 11: History rank on Traffic ($L = 720$). LP denotes learned parameters; MSE and MAE are five-seed means with sample SD for $r > 0$.

Rank	$H = 96$			$H = 192$			$H = 336$			$H = 720$		
	LP	MSE	MAE	LP	MSE	MAE	LP	MSE	MAE	LP	MSE	MAE
0	27	0.387	0.260	27	0.399	0.264	27	0.410	0.269	27	0.446	0.290
1	61	0.385 ± 0.000	0.260 ± 0.001	65	0.397 ± 0.001	0.263 ± 0.001	71	0.408 ± 0.000	0.268 ± 0.000	87	0.444 ± 0.000	0.289 ± 0.001
2	95	0.383 ± 0.000	0.259 ± 0.001	103	0.395 ± 0.001	0.263 ± 0.001	115	0.407 ± 0.000	0.268 ± 0.000	147	0.442 ± 0.000	0.289 ± 0.001
3	129	0.382 ± 0.000	0.259 ± 0.001	141	0.394 ± 0.001	0.263 ± 0.001	159	0.407 ± 0.000	0.269 ± 0.001	207	0.442 ± 0.000	0.289 ± 0.001
4	163	0.382 ± 0.000	0.259 ± 0.001	179	0.394 ± 0.001	0.263 ± 0.001	203	0.407 ± 0.001	0.269 ± 0.001	267	0.441 ± 0.000	0.288 ± 0.001

F Representation controls

F.1 Shared weekly map

The shared weekly map periodizes the filtered history into 168-position coordinates while retaining all 720 input values and all H forecast values. A validity mask handles partial cycles. The effective temporal map has 4, 9, 10, and 24 coefficients for $H = 96, 192, 336, 720$, respectively, plus four shared preprocessing and boundary scalars. Ridge regression initializes this map before joint training. Table 1 compares it with the complete seasonal bank, alone and with a rank-one history correction.

G Seasonal controls and branch ablations

G.1 Fixed seasonal predictors

We compare input-only seasonal predictors on ETTh2, ETTm2, and Traffic at $L = 720$ and $H \in \{96, 192, 336, 720\}$, using the same standardization and forecast windows as the learned models. The mean predictor repeats the window mean. Base-period and longer-cycle naive predictors repeat the latest p and gp observations. Phase means average observations at matching positions over cycles.

ETTh2 and Traffic use $p = 24$, $g = 7$, base widths 7 and 28, and a four-week longer-cycle mean. ETTm2 follows each full model's period configuration: $p = 12$, $g = 8$ at $H = 336$, and $p = 4$, $g = 24$ otherwise. Its base widths are 7 and $720/p$, and its longer-cycle mean uses seven days. The wide base and cycle means in Table 2 use the largest supported widths for each period configuration.

The fixed-coefficient variant is the rank-zero endpoint

$$\widehat{Y}_{kp+j} = \mu + F_{w_*}(j) + P_{v_*}(k, j), \qquad w_* = \max \mathcal{W}, \quad v_* = \max \mathcal{G}.$$

Here $Q = X - \mu$ and $\rho = 0$; smoothing uses $\alpha = 0.25$. The widest base and contrast gains are one; all other coefficients are zero. We use $(w_*, v_*) = (28, 4)$ on ETTh2 and Traffic and $(720/p, 7)$ on ETTm2. Fixed predictors learn no parameters. Tables 12 through 14 report MSE and MAE.

Table 12: Fixed and learned seasonal predictors on ETTh2 at $L = 720$. The full model uses 61, 65, 71, and 87 parameters by horizon; state-only uses 27. Full-model values are five-seed means with sample SD.

Model	$H = 96$		$H = 192$		$H = 336$		$H = 720$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Window mean	0.389	0.425	0.403	0.434	0.406	0.438	0.434	0.461
Base-period naive	0.391	0.380	0.482	0.429	0.532	0.466	0.525	0.474
Longer-cycle naive	0.582	0.485	0.579	0.487	0.567	0.491	0.554	0.491
Base-phase mean (7)	0.350	0.378	0.401	0.413	0.410	0.428	0.420	0.439
Base-phase mean (wide)	0.350	0.397	0.363	0.407	0.370	0.414	0.399	0.443
Longer-cycle mean	0.402	0.430	0.399	0.429	0.402	0.433	0.429	0.459
Fixed coefficients	0.399	0.429	0.397	0.428	0.399	0.432	0.427	0.458
State-only	0.273	0.331	0.341	0.375	0.371	0.402	0.387	0.424
Full model	0.269 ± 0.001	0.331 ± 0.000	0.336 ± 0.003	0.373 ± 0.001	0.358 ± 0.002	0.395 ± 0.001	0.384 ± 0.001	0.422 ± 0.000

Table 13: Fixed and learned seasonal predictors on ETTm2 at $L = 720$. The full model uses 233, 257, 205, and 389 parameters by horizon. Full-model values are five-seed means with sample SD.

Model	$H = 96$		$H = 192$		$H = 336$		$H = 720$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Window mean	0.316	0.369	0.340	0.382	0.371	0.399	0.423	0.433
Base-period naive	0.270	0.331	0.343	0.373	0.420	0.416	0.523	0.467
Longer-cycle naive	0.263	0.301	0.321	0.337	0.376	0.370	0.471	0.422
Base-phase mean (7)	0.234	0.314	0.302	0.352	0.343	0.371	0.473	0.443
Base-phase mean (wide)	0.316	0.369	0.340	0.382	0.371	0.400	0.423	0.434
Longer-cycle mean	0.281	0.336	0.308	0.352	0.341	0.372	0.397	0.409
Fixed coefficients	0.286	0.340	0.312	0.355	0.344	0.374	0.396	0.410
Full model	0.161 ± 0.001	0.250 ± 0.001	0.216 ± 0.001	0.290 ± 0.001	0.268 ± 0.000	0.325 ± 0.000	0.351 ± 0.001	0.377 ± 0.001

Table 14: Fixed and learned seasonal predictors on Traffic at $L = 720$. The full model uses 61, 65, 71, and 87 parameters by horizon; state-only uses 27. Full-model values are five-seed means with sample SD.

Model	$H = 96$		$H = 192$		$H = 336$		$H = 720$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Window mean	1.383	0.795	1.393	0.797	1.409	0.800	1.431	0.804
Base-period naive	1.217	0.497	1.092	0.458	1.150	0.475	1.185	0.489
Longer-cycle naive	0.620	0.296	0.624	0.297	0.632	0.302	0.660	0.319
Base-phase mean (7)	0.623	0.378	0.646	0.387	0.664	0.393	0.689	0.404
Base-phase mean (wide)	0.622	0.383	0.631	0.387	0.643	0.391	0.678	0.404
Longer-cycle mean	0.440	0.272	0.437	0.272	0.444	0.275	0.489	0.296
Fixed coefficients	0.432	0.272	0.429	0.271	0.437	0.275	0.480	0.295
State-only	0.387	0.260	0.399	0.264	0.410	0.269	0.446	0.290
Full model	0.385 ± 0.000	0.260 ± 0.001	0.397 ± 0.001	0.263 ± 0.001	0.408 ± 0.000	0.268 ± 0.000	0.444 ± 0.000	0.289 ± 0.001

G.2 Branch ablations

Both branch ablations keep the complete input and forecast windows. The history-only model retains four filter and boundary scalars and a rank-one history map. The state-only model retains all seven state groups and five shared scalars. Training follows Appendix D. Table 15 reports the five-seed means, paired changes in error relative to the full model, and sample SD except for state-only.

Table 15: Branch ablation at $H = 336$. Five-seed means include sample SD except for state-only. Paired differences are ablation minus full model, in 10^{-3} units; positive values mean higher error.

Dataset	Model	Scalars	MSE	MAE	ΔMSE	ΔMAE
ETTh2	Full model	71	0.358 ± 0.002	0.395 ± 0.001	–	–
	History-only	48	0.363 ± 0.001	0.399 ± 0.001	$+4.848 \pm 2.308$	$+3.754 \pm 1.160$
	State-only	27	0.371	0.402	+13.124	+7.338
Traffic	Full model	71	0.408 ± 0.000	0.268 ± 0.000	–	–
	History-only	48	0.632 ± 0.000	0.387 ± 0.000	$+224.320 \pm 0.315$	$+119.068 \pm 0.303$
	State-only	27	0.410	0.269	+1.731	+0.412

H Period configuration

H.1 ETTm2 model configuration

ETTm2 uses the same forecasting formulation with a daily longer cycle. Model hyperparameters are selected separately for each forecast horizon using validation MSE. Training follows Section D.2.

H.2 Daily-period comparison

For sampling intervals of Δ minutes, a daily period is $p = 1440/\Delta$. On ETTm2, a seven-day history gives $(L, p) = (672, 96)$; we forecast one or two days, $H \in \{96, 192\}$.

Retaining $w \in \{1, 2, 7, 28\}$ with $w \leq L/p$ and $v \in \{1, 2, 4\}$ with $7v \leq L/p$ gives $\mathcal{W} = (1, 2, 7)$, $\mathcal{G} = (1)$, and $g = 7$. Filter widths are $p + 1$ and 3; the boundary reference contains $p/4$ samples. Rank one with $n = 7$ and $m \in \{1, 2\}$ gives 26/27 learned scalars by Equation 5.

Table 3 compares these configurations with $p = 24$ under identical input and output windows. Each retains its supported widths and uses the training procedure of Section D.2. The daily-period configurations lower mean MSE and MAE at both forecast lengths while using fewer learned scalars.

I Initialization and hyperparameter studies

I.1 Initialization variants

Table 16 uses $L = 720$ and $H = 336$. Hourly models have 71 parameters; ETTm2 has 205, with $p = 12$, $g = 8$, base widths $(1, 2, 7, 60)$, longer-cycle widths $(1, 2, 4, 7)$, and rank two.

Conventional initialization assigns equal weights to the base profiles, initializes the robust gain to 0.500 and other readout, filtering, and boundary coefficients to zero. The input factor has independent $\mathcal{N}(0, 1/n)$ entries, and the output factor is zero. *One cycle* and *1,000 cycles* first solve for the state readout with zero history correction, then fit the history factors by reduced-rank regression. They perform one or 1,000 additional complete alternating state and history refinement cycles, respectively. All three variants then use the joint optimization procedure in Section D.2, selecting a learning rate separately for each variant by validation MSE. *Fit only* evaluates the 1,000-cycle regression fit without joint training. One-cycle and 1,000-cycle initializations give similar errors on all five datasets.

Table 16: Initialization and joint training at $L = 720$, $H = 336$. Trained variants report five-seed mean $\pm$ sample SD. Fit only uses deterministic conditional regression after 1,000 refinement cycles.

Dataset	Metric	Conventional	One cycle	1,000 cycles	Fit only
ETTh1	MSE	0.421 ± 0.002	0.421 ± 0.002	0.422 ± 0.002	0.423
	MAE	0.424 ± 0.001	0.425 ± 0.001	0.425 ± 0.001	0.425
ETTh2	MSE	0.357 ± 0.000	0.357 ± 0.002	0.358 ± 0.002	0.358
	MAE	0.395 ± 0.000	0.395 ± 0.001	0.395 ± 0.001	0.396
ETTm2	MSE	0.269 ± 0.001	0.268 ± 0.000	0.268 ± 0.000	0.269
	MAE	0.325 ± 0.001	0.325 ± 0.000	0.325 ± 0.000	0.327
Electricity	MSE	0.165 ± 0.000	0.165 ± 0.000	0.165 ± 0.000	0.166
	MAE	0.260 ± 0.001	0.260 ± 0.001	0.260 ± 0.000	0.260
Traffic	MSE	0.409 ± 0.000	0.408 ± 0.000	0.408 ± 0.000	0.412
	MAE	0.269 ± 0.000	0.268 ± 0.000	0.268 ± 0.000	0.271

I.2 Fixed hyperparameters

The sensitivity study uses the same $L = 720$, $H = 336$ architectures. We vary one factor at a time while fixing the architecture, learning rate, and other hyperparameters (Table 17). The phase-smoothing strength multiplies the circular smoothing term. For a filter-support multiplier c_f, the long moving-average width is $2\lfloor pc_f/2 \rfloor + 1$; the short filter retains width three. The temporal basis is $\phi_1(k) = \exp[-k/\max(mc_d, 1)]$, where c_d is the decay scale. The state and history ridge factors are the relative strengths λ_s^{rel} and λ_h^{rel} in the conditional initialization.

J Empirical alignment with the prescribed transfer dictionary

We fit shared seasonal transfers to a linear predictor estimated from real data, then measure their approximation error on test windows. This tests the dictionary coverage analyzed in Appendix C.

Reference and dictionary. Let $L = np$ and $H = mp$. After training-set standardization, subtract each variable's input-window mean from its history and target. Each variable and phase supplies

Table 17: Values in the fixed-constant study. Each experiment changes one factor while preserving the other default values, architecture, and learning rate.

Factor	Low	Default	High
Phase smoothing	0.125	0.250	0.500
Filter support	0.500	1.000	2.000
Decay scale	0.250	0.500	1.000
State ridge	0.001	0.010	0.100
History ridge	0.001	0.010	0.100

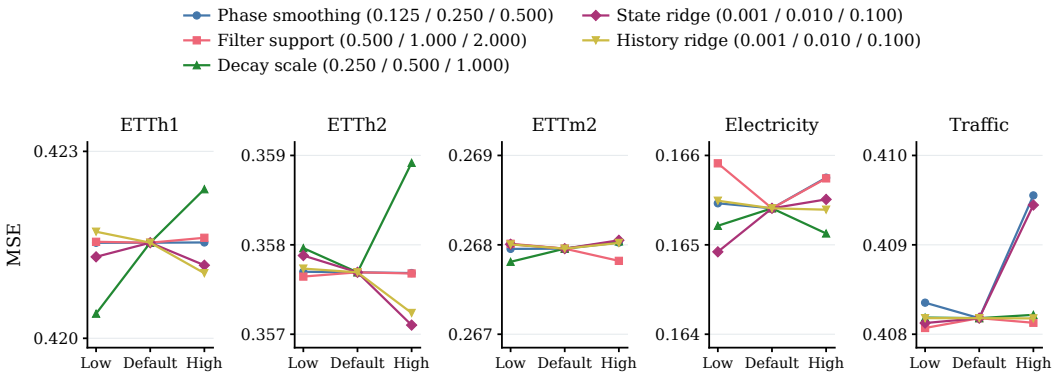


Figure 5: Sensitivity to fixed hyperparameters at $H = 336$. Curves show five-seed mean MSE for low, default, and high values listed in the legend, with model size and learning rate fixed. Vertical axes use the original MSE units with a separate scale for each dataset.

a cycle-history vector $z_i \in \mathbb{R}^n$ and target $y_i \in \mathbb{R}^m$. From the N_{tr} pooled training pairs, form the second moments $G = N_{\mathrm{tr}}^{-1} \sum_i z_i z_i^\top$ and $C = N_{\mathrm{tr}}^{-1} \sum_i y_i z_i^\top$, and fit

$$\widehat{T}_\lambda = C(G + \lambda a I_n)^{-1}, \qquad a = \max\{\mathrm{tr}(G)/n, 10^{-12}\}.$$

Shared across variables and phases, $\widehat{T}$ uses $\lambda \in \{10^{-4}, 10^{-2}, 1\}$ selected by validation MSE.

For a longer cycle of g positions and a width w with $wg \le n$, the mean-based transfer $B_w \in \mathbb{R}^{m \times n}$ is

$$(B_w z)_k = \frac{1}{w} \sum_{j=0}^{w-1} \left[z_{n-wg+jg+(k \bmod g)} - \frac{1}{g} \sum_{\ell=0}^{g-1} z_{n-wg+jg+\ell} \right], \quad 0 \le k < m.$$

It averages matching positions in the latest w groups, removes their group-position mean, and routes them to future cycles. This is the mean-only longer-cycle branch, without median mixing, phase smoothing, learned filtering, or boundary correction. Let $\phi_0(k) = 1$, $\phi_1(k) = \exp[-k/\max(m/2, 1)]$, $Q_0 = [\phi_0, \phi_1]$, $P_0 = Q_0 Q_0^\dagger$, and $P = I_m - P_0$. We retain $P_0 \widehat{T}$ as a relaxed output component: arbitrary $Q_0 F$, $F \in \mathbb{R}^{2 \times n}$, allows more coefficients than the model's compact base and level readouts. For widths $w \in \mathcal{W}$ and $b \in \{0, 1\}$, set $D_{b,w} = P\,\mathrm{diag}(\phi_b) B_w$ and fit

$$\widehat{\theta} \in \arg\min_\theta \left\| \left(P\widehat{T} - \sum_{b,w} \theta_{b,w} D_{b,w} \right) G^{1/2} \right\|_F^2,$$

$$\widehat{T}_A = P_0 \widehat{T} + \sum_{b,w} \widehat{\theta}_{b,w} D_{b,w},$$

$$\widehat{T}_{A,r} = \widehat{T}_A + [(\widehat{T} - \widehat{T}_A) G^{1/2}]_r (G^{1/2})^\dagger.$$

Here $[\cdot]_r$ is a best rank-r singular-value truncation. Both fits use training moments.

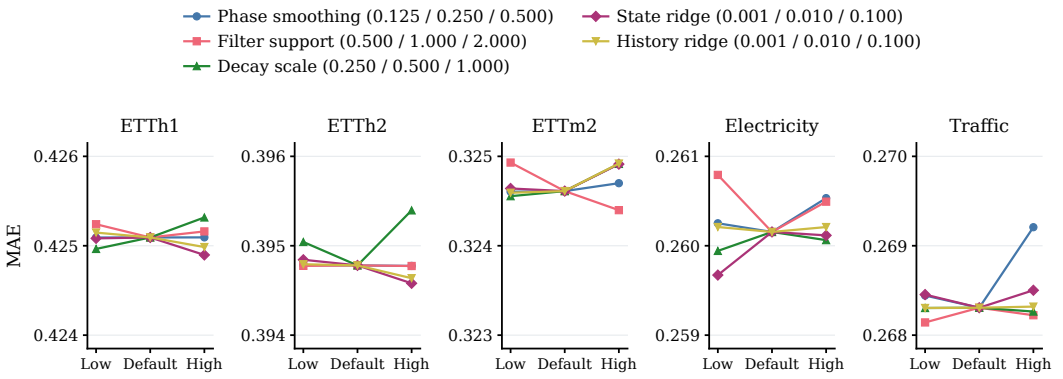


Figure 6: MAE sensitivity to the five fixed hyperparameters at $H = 336$. Curves show five-seed mean MAE in its original units, using the settings of Figure 5.

Table 18: Test-set approximation of the training-fitted reference at $H = 336$. Energy and discrepancy use 10^{-3} units; ratios are percentages. A and A, r denote fits before and after a rank-r correction.

Dataset	λ	$\mathcal{E}$	$\mathcal{D}_A$	$\mathcal{D}_A/\mathcal{E}$ (%)	$\mathcal{D}_{A,r}$	$\mathcal{D}_{A,r}/\mathcal{E}$ (%)
ETTh1	1	0.650	0.534	82.2	0.323	49.7
ETTh2	10^{-4}	2.495	2.250	90.2	1.551	62.2
ETTm2	10^{-2}	32.368	2.521	7.8	1.356	4.2
Electricity	10^{-2}	38.795	1.405	3.6	0.836	2.2
Traffic	10^{-4}	209.704	8.279	3.9	5.774	2.8

Held-out representation error. For N_{te} test pairs, define

$$\mathcal{E} = \frac{1}{N_{\text{te}} m} \sum_i \|P\widehat{T} z_i\|_2^2, \qquad \mathcal{D}(A) = \frac{1}{N_{\text{te}} m} \sum_i \|(\widehat{T} - A) z_i\|_2^2.$$

Thus $\mathcal{D}/\mathcal{E}$ normalizes prediction discrepancy by the reference energy outside the constant-and-decay span. Both quantities are per original forecast scalar, since $N_{\text{te}} m$ equals the number of windows times variables times H. For $\mathcal{R}(A) = (N_{\text{te}} m)^{-1} \sum_i \|y_i - A z_i\|_2^2$, their relation to prediction error includes a residual cross term:

$$\mathcal{R}(A) - \mathcal{R}(\widehat{T}) = \mathcal{D}(A) + \frac{2}{N_{\text{te}} m} \sum_i \langle y_i - \widehat{T} z_i, (\widehat{T} - A) z_i \rangle. \tag{8}$$

Thus the approximation error equals the change in forecast MSE after accounting for the residual cross term. We evaluate the mean-only dictionary with this relaxed output component.

Evaluation. We use $L = 720$ and $H = 336$. The hourly datasets use $p = 24$, $g = 7$, $\mathcal{W} = \{1, 2, 4\}$, and $r = 1$; ETTm2 uses $p = 12$, $g = 8$, $\mathcal{W} = \{1, 2, 4, 7\}$, and $r = 2$. Routing includes partial final longer-cycle groups. Table 18 reports approximation errors over all valid test windows. The rank correction reduces the dictionary's representation discrepancy in all five conditions.