CONFIGURING BEFORE PREDICTION: RETHINKING TIME-SERIES FORECASTING TASK FORMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series forecasting research has devoted substantial effort to improving how models predict, while largely taking for granted what historical input they should predict from. Yet the historical input is itself a key part of the forecasting task formulation: it determines both what temporal structure is exposed to the predictor and how far into the past that structure extends. In common forecasting practice, these aspects are typically predetermined rather than adapted to the time series and forecast horizon at hand. We therefore propose a framework that jointly configures the temporal structure and temporal extent of the historical input for each channel and forecast horizon before prediction, selecting among reversible temporal representations and candidate historical extents. To make heterogeneous input formulations comparable without repeatedly fitting downstream models, we establish a unified risk characterization of their expected multi-step forecasting behavior on the original data scale, enabling principled configuration directly from the training series. Experiments on public benchmarks show that the resulting configurations consistently improve forecasting accuracy over commonly used fixed configurations across diverse forecasting architectures, including DLinear, PatchTST, and Koopman-based models. The improvements also persist across different choices of linear operators, demonstrating that the benefits are not tied to a particular predictor. Notably, even a simple scalar autoregressive forecaster becomes competitive with substantially more complex forecasting models when equipped with the configured historical input. These results highlight historical input formulation as an important yet underexplored dimension of time-series forecasting and show that configuring what a model predicts from can be as consequential as improving how it predicts.

1 INTRODUCTION

Time-series forecasting research has made substantial progress by developing increasingly capable predictors. Yet before any predictor is applied, a more basic choice has already been made: *what historical input should it predict from?* This choice determines not only how far into the past the predictor can observe, but also what temporal structure is presented for prediction. In common forecasting practice, these aspects are typically predetermined as part of the input specification, rather than treated as variables of the forecasting task itself. However, different historical inputs can expose different predictive relationships even when the predictor remains unchanged. This motivates us to ask whether historical input should remain a fixed input convention or instead be regarded as part of the forecasting task formulation.

To examine whether historical input is a genuine task-dependent choice, Figure 1 holds the predictor fixed and varies the two dimensions of historical-input formulation separately. With the identity representation fixed, Fig. 1(a) shows that test error varies non-monotonically with temporal extent and that the post-hoc best fixed lookback ranges from $p = 336$ to $p = 1920$ across the 28 dataset–horizon tasks. With $p = H$ fixed, Fig. 1(b) shows that the preferred temporal structure also varies across tasks: Id, Δ_1, $\mathcal{R}_T$, and Δ_T attain the lowest error on 2, 10, 14, and 2 tasks, respectively. Test outcomes are used here only for post-hoc visualization and do not enter the proposed configuration procedure. Together, these observations show that neither a single temporal extent nor a single temporal structure provides a universal input convention, motivating their joint configuration before prediction. This raises a more fundamental question: rather than committing to a fixed input

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

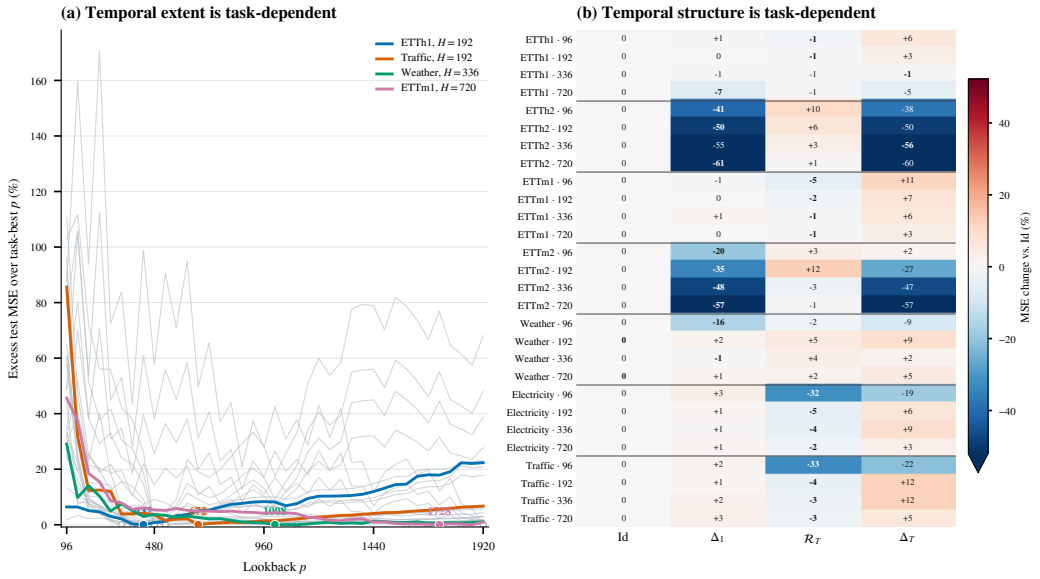

Figure 1: Historical-input formulation is task-dependent even when the predictor is held fixed. (a) With $m = \mathrm{Id}$, each curve reports channel-mean test MSE over $p \in \{96, 144, \ldots, 1920\}$, expressed as excess error relative to that task's post-hoc best fixed p. Gray curves denote all 28 dataset–horizon tasks; colored curves are four pre-specified representative tasks, and markers indicate task-wise minima. (b) With $p = H$, cells report the percentage MSE change of each temporal structure relative to Id; negative values indicate improvement, and bold values denote the best representation within each task. Colors are symmetrically capped at 52.1% for readability, while printed values remain uncapped. Test outcomes are used only for this post-hoc motivation diagnostic and do not enter the proposed configuration procedure.

convention, how should the historical input be formulated for the time series and forecast horizon at hand?

Answering this question requires a common basis for comparing heterogeneous historical-input formulations. Neither a more regular temporal structure nor a longer history is inherently preferable: altering temporal structure can expose useful dynamics but may also remove variation important for prediction (Liu et al., 2022), while additional history may capture useful dependencies but can also introduce redundancy and noise (Zhang et al., 2025a). Yet regardless of how history is represented or how far it extends, the resulting forecast is ultimately evaluated against future values of the same original series. We exploit this shared prediction target to characterize the expected multi-step forecasting behavior of different formulations on the original data scale, providing a unified basis for configuring temporal structure and temporal extent directly from the training series before downstream prediction. Experiments show that the resulting configurations improve forecasting across linear, Transformer-based, and Koopman-based predictors; notably, even a simple autoregressive forecaster becomes competitive with substantially more complex models when equipped with an appropriately formulated historical input. Figure 2 places this configured scalar forecaster in the descriptive accuracy–capacity lower-left region. Because its horizontal coordinate is fitted autoregressive coefficients per channel whereas neural baselines use total trainable parameters, the comparison is not an equal-budget comparison. These results suggest that forecasting performance depends not only on how a model predicts, but also on how the historical evidence it predicts from is formulated. To summarize, our contributions are threefold:

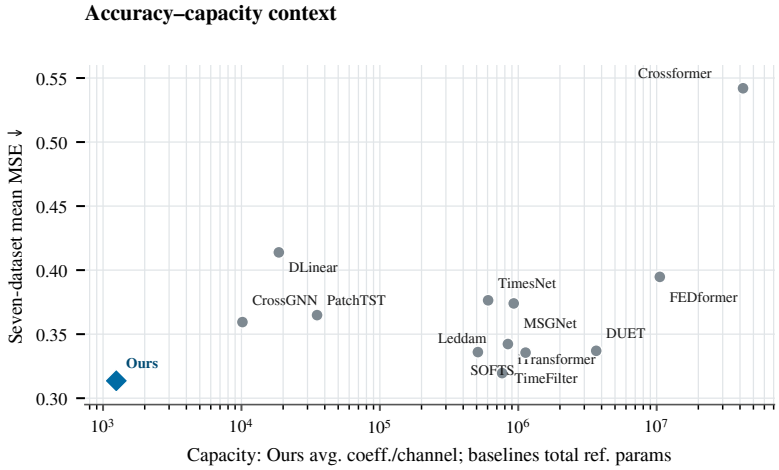

Figure 2: Accuracy–capacity context for the 13 methods in Table 1. The vertical axis is the arithmetic mean MSE over seven datasets and four forecast horizons. For our scalar forecaster, capacity is the average number of fitted autoregressive coefficients per channel over the same 28 tasks; for neural baselines, it is the total trainable parameter count under the reported reference configuration. Because these quantities have different definitions, the plot provides descriptive model-capability context rather than an equal-budget comparison.

1. We rethink time-series forecasting from the perspective of *historical input formulation*, identifying temporal structure and temporal extent as two coupled dimensions that shape the forecasting task before a predictor is applied. This perspective shifts attention from optimizing the predictor alone to jointly considering how the historical evidence presented to it should be formulated for different series and forecast horizons.

2. We establish a unified characterization of expected multi-step forecasting risk that makes heterogeneous historical-input formulations comparable under the same original-scale forecasting objective. This characterization provides a principled basis for configuring temporal structure and temporal extent directly from the training series, without relying on repeated downstream model fitting or held-out tuning.

3. We demonstrate that the benefit of historical input formulation generalizes across forecasting paradigms rather than being tied to a particular predictor. Across linear, Transformer-based, and Koopman-based forecasters, as well as alternative linear-operator constructions, appropriately configured historical inputs yield consistent improvements over fixed formulations; notably, they can make a simple autoregressive forecaster competitive with substantially more complex architectures.

2 Related Work

2.1 Time-Series Forecasting

Long-term forecasting has progressed through several architectural paradigms. DLinear (Zeng et al., 2023) showed that direct linear mappings can be highly competitive; PatchTST (Nie et al., 2023) represents each univariate history as patch tokens for a channel-independent Transformer; and TimeFilter (Hu et al., 2025) adaptively filters temporal and cross-channel dependencies. Koopman-based models instead evolve delay or learned states through linear operators, including Hankel-DMD (Arbabi & Mezić, 2017), KNF (Wang et al., 2023), Koopa (Liu et al., 2023), and SKOLR (Zhang et al., 2025b). These models provide diverse backbones for evaluating whether historical-input configuration transfers beyond a particular predictor.

2.2 Historical Input Formulation

Historical input can be reformulated along both temporal structure and temporal extent. For distribution shift, RevIN (Kim et al., 2022) normalizes each instance and restores its statistics at the output, Dish-TS (Fan et al., 2023) models shifts between historical and forecast windows, and Non-stationary Transformers (Liu et al., 2022) reintroduces information lost during stationarization. Structural methods act differently: Autoformer (Wu et al., 2021) performs progressive trend–seasonal decomposition, while CycleNet (Lin et al., 2024) estimates a recurring cycle and forecasts the residual.

Temporal extent is equally consequential. PatchTST makes longer contexts computationally accessible through patching, whereas work on the maximum effective window (Zhang et al., 2025a) shows how redundancy and attention dispersion can limit their utility. Adaptive Lookback with Wavelets (Tong & Yuan, 2026) learns instance-dependent multi-scale histories. These methods typically prescribe a particular transformation or learn a model-specific lookback mechanism. We instead jointly configure explicit reversible temporal structures and temporal extents under the same original-scale multi-step forecasting objective, allowing the resulting input to be reused across predictors.

2.3 Forecast-Oriented Configuration

Statistical model-selection criteria provide a foundation for configuring predictive models from observed data. Akaike's information criterion (Akaike, 1974) balances fit and complexity; Ing (2004) studies autoregressive order and prediction-method selection under multi-step mean-squared forecast error; and Ing et al. (2009) extend multi-step configuration to non-stationary autoregressions. These works motivate evaluating behavior at the target forecast horizon rather than relying on one-step fit alone. We broaden the configuration object from order within a fixed representation to the coupled pair of temporal structure and temporal extent, characterized directly from the training series on the original prediction scale.

3 Method

Figure 3 provides an overview of our framework. We treat historical input as a task-dependent formulation along two coupled dimensions: temporal structure and temporal extent. Specifically, each channel is represented using a reversible temporal structure m and a history extent p (Section 3.1). To compare heterogeneous formulations on a common basis, we use the same scalar autoregressive reference to characterize their reconstructed multi-step forecasting behavior on the original data scale, accounting for prediction error, finite-data uncertainty, and persistent level variation (Section 3.2). The resulting training-only criterion $Q_{c,H}(m,p)$ determines the formulation $(\widehat{m}_{c,H}, \widehat{p}_{c,H})$ before downstream prediction (Section 3.3). This separates the formulation of historical input from predictor design, allowing the configured input to be used by different forecasting architectures.

3.1 Historical Input Formulation

Let $x_s^{(c)} \in \mathbb{R}$ denote the observed value of channel c at time s. For a channel c and forecast horizon H, we describe a historical input by two coupled choices, $a = (m, p)$: a reversible representation m determines its temporal structure, and a lookback length p determines its temporal extent. At forecast origin t, the formulated input is passed to a forecaster f; its output is then reconstructed using only information available at t:

$$\widehat{x}^{(c)}_{t+1:t+H} = \mathcal{I}_m\Big(f\Big(\Phi_{m,p}\big(x^{(c)}_{t-p+1:t}, \mathcal{A}^{(c)}_{m,p,t}\big)\Big), x^{(c)}_{t-p+1:t}, \mathcal{A}^{(c)}_{m,p,t}\Big). \tag{1}$$

Here $\Phi_{m,p}$ constructs a length-p input in representation m. The auxiliary information $\mathcal{A}^{(c)}_{m,p,t}$ contains any earlier observations or training-derived quantities needed to construct that representation and reconstruct its forecast; it is not passed directly to f. The map $\mathcal{I}_m$ reconstructs the forecast in the original data space. Different formulations change the evidence presented to f, but all are evaluated against the same future values in that space. This common prediction target makes historical input formulation a forecasting-task decision that can be studied across different predictors.

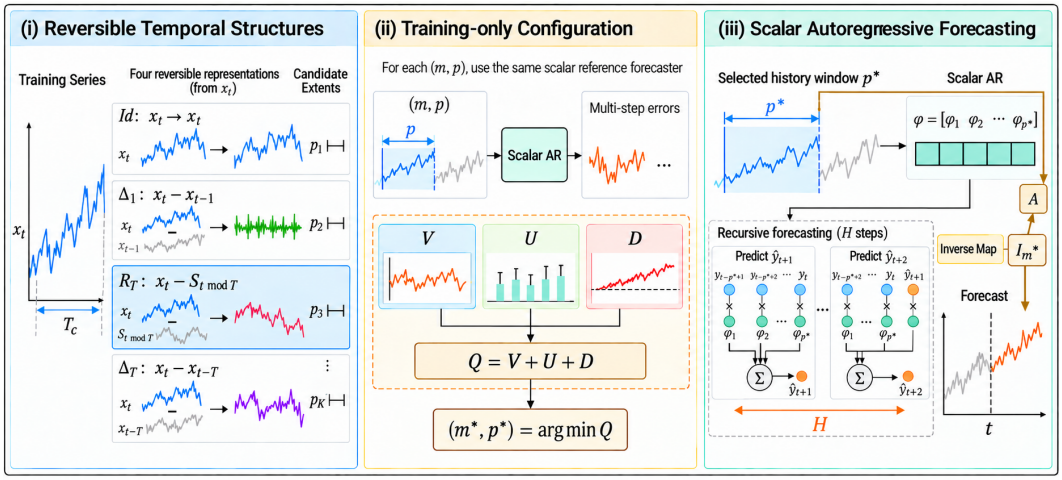

Figure 3: Overview of historical-input formulation. (i) Each channel is expressed through one of four reversible temporal structures and a candidate temporal extent p. (ii) For every (m, p), the same scalar autoregressive reference characterizes reconstructed multi-step error V, coefficient uncertainty U, and persistent level variation D; minimizing $Q = V + U + D$ using the training series yields $(m^\star, p^\star)$. (iii) The configured history and scalar coefficient vector form a dot product at each recursive step, while auxiliary information $\mathcal{A}$ bypasses the predictor and enters $\mathcal{I}_{m^\star}$ to reconstruct the original-scale forecast.

Reversible temporal structures. The representation m determines which temporal relationship the forecaster must predict. For the two seasonal representations, T denotes a training-derived lag measured in observation steps. Suppressing the channel index for clarity, we consider

$$r_t^{(m)} = \begin{cases} x_t, & m = \mathrm{Id}, \\ x_t - x_{t-1}, & m = \Delta_1, \\ x_t - S^{(T)}_{t \bmod T}, & m = \mathcal{R}_T, \\ x_t - x_{t-T}, & m = \Delta_T. \end{cases} \tag{2}$$

Here Id retains the observed level; Δ_1 represents change from the preceding observation; $\mathcal{R}_T$ represents deviation from a recurring pattern; and Δ_T represents change from the observation T steps earlier. These candidates distinguish level, ordinary differencing, recurring-profile removal, and seasonal differencing, whose forecasting implications are classically treated as different structural cases (Hylleberg et al., 1990; Franses, 1991).

For each representation, T, the phase template, and the normalization statistics are estimated from the original training series; the exact adjusted template-fit criterion and normalization are given in Appendix A. The last p values of $y_t^{(m)} = (r_t^{(m)} - \widehat{\mu}_m)/\widehat{\sigma}_m$ are supplied to the forecaster. Its outputs are mapped back and reconstructed with the available historical anchors and, for $\mathcal{R}_T$, the fixed phase template. Thus all four representations forecast the same original series.

3.2 Training-only Risk Characterization

To compare formulations without repeatedly fitting downstream architectures, we use the same scalar linear autoregressive reference for every (m, p). With $s_t^{(m,p)} = (y_t^{(m)}, \ldots, y_{t-p+1}^{(m)})^\top$, its coefficients $\widehat{\phi}_{m,p}$ are estimated from training autocovariances by Yule–Walker equations, and recursively forecast

$$\widehat{y}_{t+j}^{(m)} = e_1^\top \widehat{K}_{m,p}^{\,j} s_t^{(m,p)}, \qquad j = 1, \ldots, H. \tag{3}$$

Here $\widehat{K}_{m,p}$ is the companion operator and e_1 its first coordinate vector. Appendix A gives the linear-projection, Yule–Walker, Levinson, and reconstruction constructions.

Reconstruction combines the transformed future steps appropriate to m. Let $\widehat{a}_{m,h}$, $\widehat{d}_{m,p,h}$, $\widehat{\Gamma}_{m,p}$, and $\widehat{b}_{m,p,h}$ denote respectively the training-estimated aggregate target variance, target–history covariance, history covariance, and recursive reconstruction coefficient (Appendix A). The original-scale plug-in multi-step risk is

$$V_{c,H}^{(m)}(p) = \frac{\widehat{\sigma}_m^2}{H} \sum_{h=1}^{H} \left(\widehat{a}_{m,h} - 2\widehat{b}_{m,p,h}^{\top} \widehat{d}_{m,p,h} + \widehat{b}_{m,p,h}^{\top} \widehat{\Gamma}_{m,p} \widehat{b}_{m,p,h} \right). \tag{4}$$

It characterizes reconstructed H-step error implied by training-estimated temporal dependence and is measured in original-series squared units.

Finite training data also make the reference coefficients uncertain. Let $\widehat{b}_{m,p,h}^{(r)}$ be the recursive reconstruction coefficient under a Gaussian perturbation of $\widehat{\phi}_{m,p}$ drawn from its training-derived covariance (Appendix A). We measure the resulting original-scale uncertainty as

$$U_{c,H}^{(m)}(p) = \frac{\widehat{\sigma}_m^2}{BH} \sum_{r=1}^{B} \sum_{h=1}^{H} \left(\widehat{b}_{m,p,h}^{(r)} - \widehat{b}_{m,p,h} \right)^{\top} \widehat{\Gamma}_{m,p} \left(\widehat{b}_{m,p,h}^{(r)} - \widehat{b}_{m,p,h} \right). \tag{5}$$

The same standard-normal directions are shared across candidates. To account for persistent original-scale level variation, Appendix A constructs a non-alternation gate G_c, local-level variation $\Lambda_{c,H}$, and representation weight W_m. The resulting score is

$$D_{c,H}^{(m)} = G_c W_m \Lambda_{c,H}, \qquad Q_{c,H}(m,p) = V_{c,H}^{(m)}(p) + U_{c,H}^{(m)}(p) + D_{c,H}^{(m)}. \tag{6}$$

All terms have original-series squared units; division by the shared $\sigma_{x,c}^2$ changes neither within-channel ordering nor the configuration rule.

3.3 Configuration and Transfer

Let $\mathcal{M}_{T_c} = \{\mathrm{Id}, \Delta_1, \mathcal{R}_{T_c}, \Delta_{T_c}\}$ denote the candidate temporal structures for channel c, and let $\mathcal{P}$ be a predeclared set of temporal extents. We configure the historical input for channel c and horizon H by

$$(\widehat{m}_{c,H}, \widehat{p}_{c,H}) = \operatorname*{arg\,min}_{m \in \mathcal{M}_{T_c},\, p \in \mathcal{P}} Q_{c,H}(m,p). \tag{7}$$

The lag, phase template, normalization statistics, autocovariances, reference dynamics, and score use only the complete training series. No training observations are held out for configuration, and validation and test outcomes never enter this minimization. Once fixed, a channel-wise downstream forecaster receives its configured structure and extent and is fitted once. For architectures requiring one shared input configuration, we aggregate the channel-wise scores as specified in Appendix A; this separates formulation from architecture-specific input search while retaining a fully training-only protocol.

4 Experiments

4.1 Experimental Setup

Our reference forecaster is a channel-wise scalar linear autoregressive operator computed in closed form; it contains no learned cross-channel interaction. We additionally compare channel-specific and channel-shared coefficient parameterizations: the former estimates one coefficient vector per channel, whereas the latter shares one across channels. We use $H \in \{96, 192, 336, 720\}$ and $p \in \{96, 144, \ldots, 1920\}$, and configure historical input from the complete training segment with $B = 16$ and seed 0. We report channel-normalized MSE/MAE, averaging test origins within channel and then channels within dataset. In the controlled temporal-extent comparison, we fix $m = \mathrm{Id}$ and use the same channel-wise Yule–Walker scalar AR forecaster for every selector over 39 candidate extents, 1,232 channels, and 28 dataset–horizon tasks. Only validation selection reads validation outcomes; the test oracle is used solely as a nondeployable diagnostic. For the full-candidate criterion audit, we separately evaluate all $4 \times 39 = 156$ feasible (m,p) candidates on ETTh1 and ETTm2. Training-derived Q is frozen before computing test errors, and the test oracle is used only as a post-hoc diagnostic. Throughout comparable result tables, red denotes the best result and boldface denotes the second best; ties receive the same style. Additional protocol details are provided in Appendix B.

Table 1: All baselines use fixed input length $L = 96$. Our method uses training-configured historical input. In the capacity row, ours is the average number of fitted coefficients per channel over seven datasets and four horizons.

Dataset	Ours		TimeFilter		DUET		iTransformer		Leddam		MSGNet		SOFTS		CrossGNN		PatchTST		Crossformer		TimesNet		DLinear		FEDformer	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	0.344	0.377	**0.377**	**0.393**	0.390	**0.393**	0.407	0.410	0.386	0.397	0.398	0.411	0.393	0.403	0.393	0.404	0.387	0.400	0.513	0.496	0.400	0.406	0.404	0.408	0.448	0.452
ETTm2	0.243	0.314	**0.272**	**0.321**	0.280	0.324	0.288	0.332	0.281	0.325	0.288	0.330	0.287	0.330	0.282	0.330	0.281	0.326	0.757	0.610	0.291	0.333	0.354	0.402	0.305	0.349
ETTh1	**0.428**	0.432	0.420	0.428	0.443	0.436	0.454	0.447	0.431	**0.429**	0.452	0.452	0.449	0.442	0.437	0.434	0.469	0.454	0.529	0.522	0.458	0.450	0.461	0.457	0.440	0.460
ETTh2	0.349	**0.398**	**0.364**	0.397	0.372	0.397	0.383	0.407	0.373	0.399	0.396	0.417	0.385	0.408	0.373	0.400	0.387	0.407	0.942	0.684	0.414	0.427	0.563	0.519	0.437	0.449
Weather	0.217	0.266	**0.239**	**0.269**	0.251	0.273	0.258	0.279	0.242	0.272	0.249	0.278	0.255	0.278	0.247	0.289	0.259	0.281	0.259	0.315	0.259	0.287	0.265	0.315	0.309	0.360
Electricity	**0.165**	0.262	0.158	0.256	0.172	**0.258**	0.178	0.270	0.169	0.263	0.194	0.300	0.174	0.264	0.201	0.300	0.216	0.304	0.244	0.334	0.193	0.295	0.225	0.319	0.214	0.327
Traffic	0.449	0.307	0.407	**0.268**	0.451	0.269	0.428	0.282	0.467	0.294	0.641	0.370	**0.409**	0.267	0.583	0.323	0.555	0.362	0.550	0.304	0.620	0.336	0.625	0.383	0.610	0.376
Capacity	1.25K/ch		764K		3.66M		842K		1.13M		927K		511K		10.2K		35.2K		42.1M		605K		18.6K		10.5M	

4.2 Comparison with Mainstream Forecasters

Table 1 provides model-capability context rather than a controlled input ablation. Our scalar operator uses training-configured historical input. We report MSE/MAE averaged over $H \in \{96, 192, 336, 720\}$; full per-horizon results are in Appendix B.

Table 1 reports each dataset's MSE and MAE averaged over the four forecast horizons. Averaging the seven displayed MSEs, our method attains 0.3136 with 1.25K fitted autoregressive coefficients per channel on average. For context, the ETTh1 reference configurations of TimeFilter and DUET attain 0.3196 and 0.3370 with 764K and 3.66M total parameters, respectively. Despite using neither learned cross-channel interaction nor iterative optimization, the configured scalar operator is competitive with TimeFilter and substantially stronger than the published PatchTST and DLinear results. Because the capacity units differ, this is descriptive model-capability context; the following same-model ablation isolates historical-input configuration. Capacity provenance is reported in Appendix B.

4.3 Core Ablation

Table 2 isolates temporal-structure configuration, temporal-extent configuration, the scale on which heterogeneous formulations are compared, and the three terms in the training-only criterion under the same reference forecaster and evaluation protocol. "Extent only (Id)" restricts temporal structure to the identity representation while retaining training-only configuration of p; "Representation only, $p = H$" removes temporal-extent configuration while otherwise retaining the full criterion. Representation-space V scores every standardized transformed target directly, whereas original-scale V applies the corresponding reconstruction rule before comparing candidates. The remaining rows use the original-scale objective with U, D, or both.

Table 2: Core ablations averaged over seven datasets and four forecast horizons.

Configuration	MSE ↓	MAE ↓	Full W/T/L	Full gain ↑	95% bootstrap CI
Extent only (Id)	0.3888	0.3746	19/0/9	10.60%	[4.25, 18.19]%
Representation only, $p = H$	**0.3248**	**0.3377**	19/0/9	4.23%	[1.76, 7.01]%
Representation-space V	0.3645	0.3722	25/0/3	10.68%	[6.19, 15.68]%
Original-scale V	0.4151	0.3851	24/0/4	13.49%	[6.60, 21.43]%
$V + U$ (w/o D)	0.4052	0.3746	10/8/10	10.80%	[3.75, 19.00]%
$V + D$ (w/o U)	0.3270	0.3492	24/0/4	3.09%	[1.67, 4.73]%
Joint ($V + U + D$)	0.3135	0.3367	–	–	–

W/T/L is reported from the full method's perspective over the 28 dataset–horizon tasks. Full gain is the paired average of $1 - \mathrm{MSE}_{\mathrm{full}}/\mathrm{MSE}_{\mathrm{variant}}$, and the final column gives its 95% bootstrap confidence interval.

The joint criterion achieves the lowest average MSE and MAE in this ablation. Relative to the extent-only identity selector, temporal-structure adaptation yields a 10.60% paired gain; relative to representation-only configuration at $p = H$, extent adaptation adds 4.23%, with both confidence intervals excluding zero. These contrasts establish the value of adapting both dimensions, not the necessity of simultaneous joint search. Removing U raises MSE from 0.3135 to 0.3270 and gives

Joint 24 wins in 28 tasks; removing D yields a balanced 10/8/10 win–tie–loss count but a large average penalty, consistent with preventing a smaller number of costly level-drift failures.

The two V-only baselines separate measurement scale from the finite-sample and level-variation corrections (Appendix B.10). Representation-space V selects identity in 4,852 of 4,928 cells, whereas original-scale V selects the periodic template in 4,580 cells and wins 19 of 28 tasks; both concentrate at the maximum order. Full $V+U+D$ reduces the median selected order from 1,920 to 1,296, wins 25 tasks, and yields a 10.68% mean paired improvement over representation-space V. Thus original-scale scoring supplies a common basis, while U and D stabilize configuration decisions.

To isolate temporal-extent selection, we fix $m = \mathrm{Id}$ and compare selectors under the same scalar forecaster and grid (Appendix B.6). Here $D = 0$ and frozen Ours-Q reduces to $V + U$; it obtains MSE 0.388830 and 6.78% oracle regret, improving on Fixed-96, Fixed-H, random selection, AIC, and BIC, and is comparable to validation selection (0.404198; 6.64%) without validation outcomes.

An exhaustive audit of all 156 (m, p) candidates on ETTh1 and ETTm2 yields Spearman correlations of 0.235/0.315 and oracle regrets of 9.52%/3.97%, without recovering either exact minimizer (Appendix B.9). Near ties partly limit exact hit (9.9/26.9 candidates lie within 2% of the oracle; Q: 17.9%/39.3%, random: 6.4%/17.2%), but genuine ranking error remains: structure/within-structure extent regrets are 1.23%/8.17% and 1.04%/2.90%. Staged development controls and a separate frozen PEMS08/Solar confirmation appear in Appendices B.7 and B.8.

4.4 Transfer Across Forecasting Architectures

We transfer the training-derived configurations to DLinear, PatchTST, Hankel-OLS, and projected SVD-Hankel on ETTh1, ETTm2, and Weather. To separate configured temporal extent from the effect of simply using a longer input, all four forecasters compare fixed $L = 96$, fixed $L = H$, and configured p while retaining the original temporal structure. DLinear and PatchTST require a shared input, so their configured p is the median channel-wise extent and their configured (m, p) pairs the channel-majority temporal structure with that median extent. The two channel-wise Koopman operators instead receive each channel's configured p or full (m, p) directly; projected SVD-Hankel uses its rank99 projection. Each downstream forecaster is fitted once for every prescribed configuration, without architecture-specific input search.

Table 3: Four-horizon average MSE after transferring the configured historical input across forecasting architectures.

Dataset	Forecaster	$L = 96$	$L = H$	Configured p	Configured (m, p)
ETTh1	DLinear	0.4589	**0.4396**	0.4452	0.4302
	PatchTST	0.4543	0.4298	**0.4221**	0.4195
	Hankel-OLS	0.4610	**0.4345**	0.4429	0.4289
	SVD-Hankel	0.4545	**0.4360**	0.4445	0.4316
ETTm2	DLinear	0.3103	0.2830	**0.2693**	0.2518
	PatchTST	0.2820	0.2586	**0.2624**	0.3120
	Hankel-OLS	0.5061	0.4821	**0.3966**	0.2522
	SVD-Hankel	0.5234	0.5113	**0.4081**	0.2937
Weather	DLinear	0.2670	**0.2500**	0.2350	0.6563
	PatchTST	0.2579	**0.2333**	0.2208	0.6729
	Hankel-OLS	0.2802	0.2382	0.2138	**0.2198**
	SVD-Hankel	0.4195	0.3297	**0.2642**	0.2349

Relative to $L = 96$, configured temporal extent reduces the three-dataset mean MSE by 8.4%, 8.9%, 15.5%, and 20.1% for DLinear, PatchTST, Hankel-OLS, and projected SVD-Hankel, respectively. Full channel-specific (m, p) further helps the two Hankel models on average but is mixed for the deep forecasters.

The more mixed (m, p) results for DLinear and PatchTST suggest that temporal extent is more architecture-agnostic than temporal structure. A plausible explanation is that the temporal structures are scored through a scalar autoregressive reference, whereas different downstream encoders process levels, increments, and recurring-pattern residuals differently. In addition, reducing channel-wise decisions to one majority structure and a median extent can discard useful heterogeneity. This

does not imply that reversible structures are intrinsically unsuitable for deep forecasters; rather, it motivates future training-only criteria that account for encoder structure and architecture-specific aggregation when configuring (m, p).

Additional per-horizon comparisons, channel-level choices, the controlled temporal-extent selector comparison, the full-candidate criterion audit, and capacity provenance are reported in Appendix B.

5 Discussion

The experiments support three conclusions. First, historical-input formulation is consequential even with a fixed predictor: temporal-structure adaptation accounts for most of the gain over identity-based extent selection. Joint is statistically indistinguishable from the pre-registered staged control, so simultaneous joint optimization is not necessary. Under identity, the training-only extent criterion improves on short fixed histories, fixed $p = H$, random selection, AIC, and BIC, and is comparable to validation selection without validation outcomes. The full-candidate audit shows that near-tied test candidates coexist with a remaining gap concentrated largely in within-structure extent selection. Second, frozen PEMS08/Solar confirmation supports configuration over conventional fixed inputs. Third, an appropriately formulated input enables a closed-form scalar forecaster to be competitive with more complex models, and temporal-extent configuration transfers to DLinear, PatchTST, Hankel-OLS, and projected SVD-Hankel; full channel-wise (m, p) transfers particularly well to the explicit Hankel constructions.

Several limitations qualify these findings. The evidence does not establish Joint as superior to staged configuration; the development-period policy comparison and frozen confirmation are therefore reported separately. The risk characterization is derived from a scalar autoregressive reference and therefore need not rank temporal structures optimally for every downstream encoder. The full-candidate ranking audit remains limited to two ETT datasets. The architecture-transfer study currently covers three datasets, and the shared (m, p) rule can lose channel-specific information, as illustrated by the deep-forecaster degradation on Weather. The candidate family is limited to four reversible structures and a finite extent grid rather than all forms of non-stationarity or temporal dependence. Moreover, the capacity comparison is descriptive because fitted coefficients per channel and total neural-network parameters are not identical units.

Future work should develop architecture-aware but still training-only criteria, grouped configurations for shared-input architectures, and richer representations or stability-aware extent search. Evaluation across additional architectures and datasets should clarify when historical-input formulation transfers directly and when it must be adapted to the predictor.

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

Hassan Arbabi and Igor Mezić. Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator. *SIAM Journal on Applied Dynamical Systems*, 16 (4), 2017. doi: 10.1137/17M1125236.

Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, New York, 2 edition, 1991. doi: 10.1007/978-1-4419-0320-4.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-TS: A general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7522–7529, 2023. doi: 10.1609/aaai.v37i6.25914.

Philip Hans Franses. Seasonality, nonstationarity and the forecasting of monthly time series. *International Journal of Forecasting*, pp. 199–208, 1991. doi: 10.1016/0169-2070(91)90054-Y.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. TimeFilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*,

volume 267, pp. 24893–24911, 2025. URL https://proceedings.mlr.press/v267/hu25ac.html.

Svend Hylleberg, Robert F. Engle, Clive W. J. Granger, and B. Sam Yoo. Seasonal integration and cointegration. *Journal of Econometrics*, 44:215–238, 1990. doi: 10.1016/0304-4076(90)90080-D.

Ching-Kang Ing. Selecting optimal multistep predictors for autoregressive processes of unknown order. *The Annals of Statistics*, 32(2):693–722, 2004. doi: 10.1214/009053604000000148.

Ching-Kang Ing, Jin-Lung Lin, and Shu-Hui Yu. Toward optimal multistep forecasts in non-stationary autoregressions. *Bernoulli*, 15(2):402–437, 2009. doi: 10.3150/08-BEJ165.

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.

Norman Levinson. The Wiener (Root Mean Square) error criterion in filter design and prediction. *Journal of Mathematics and Physics*, 25(1–4):261–278, 1946. doi: 10.1002/sapm1946251261.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. CycleNet: Enhancing time series forecasting through modeling periodic patterns. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bfe7998398779dde03cad7a73b1f81b6-Abstract-Conference.html.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with Koopman predictors. In *Advances in Neural Information Processing Systems*, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/28b3dc0970fa4624a63278a4268de997-Abstract-Conference.html.

Andrew W. Lo and A. Craig MacKinlay. Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, 1(1):41–66, 1988. doi: 10.1093/rfs/1.1.41.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/nie2023iclr-time/.

Suxin Tong and Jingling Yuan. Efficiently enhancing long-term series forecasting via adaptive look-back with wavelets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25966–25974, 2026. doi: 10.1609/aaai.v40i31.39797.

Gilbert Thomas Walker. On periodicity in series of related terms. *Proceedings of the Royal Society of London. Series A*, 131(818):518–532, 1931. doi: 10.1098/rspa.1931.0069.

Rui Wang, Yihe Dong, Sercan O. Arik, and Rose Yu. Koopman neural operator forecaster for time-series with temporal distributional shifts. In *International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/wang2023iclr-koopman/.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

G. Udny Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A*, 226:267–298, 1927. doi: 10.1098/rsta.1927.0007.

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. doi: 10.1609/aaai.v37i9.26317.

Jiahui Zhang, Zhengyang Zhou, Wenjie Du, and Yang Wang. Enhancing the maximum effective window for long-term time series forecasting. In *Advances in Neural Information Processing Systems*, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/79345e916c5611bf39138892d6d58a86-Abstract-Conference.html.

Yitian Zhang, Liheng Ma, Antonios Valkanas, Boris N. Oreshkin, and Mark Coates. SKOLR: Structured Koopman operator linear RNN for time-series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 75734–75756, 2025b. URL https://proceedings.mlr.press/v267/zhang25be.html.

A METHOD DETAILS

This appendix gives the complete training-only constructions used by the compact Method description. All quantities below are computed from the training segment only.

A.1 SEASONAL TEMPLATE AND NORMALIZATION

The phase template $S^{(T)}$ and lag T are determined from the original training series $x_{0:N-1}$. Let $\bar{x}$ be its mean, $T_{\max}$ a prespecified search bound, and $\mathcal{T} = \{2, \dots, \min(\lfloor N/10 \rfloor, T_{\max})\}$. For each $k \in \mathcal{T}$, define

$$\begin{aligned} S_j^{(k)} &= \operatorname{median}\{x_t : 0 \le t < N,\ t \bmod k = j\}, \\ q_{\mathrm{adj}}(k) &= 1 - \frac{\displaystyle\sum_{t=0}^{N-1} (x_t - S^{(k)}_{t \bmod k})^2/(N-k)}{\displaystyle\sum_{t=0}^{N-1} (x_t - \bar{x})^2/(N-1)}, \\ T &= \min \arg\max_{k \in \mathcal{T}} q_{\mathrm{adj}}(k). \end{aligned} \tag{8}$$

The phase-wise median reduces the influence of extreme observations, the $N - k$ denominator discourages a longer lag from being favored only because its template has more phases, and ties select the shorter lag. For each representation, reindex its valid transformed training segment as $r_0^{(m)}, \dots, r_{n_m-1}^{(m)}$ and set

$$\widehat{\mu}_m = \frac{1}{n_m} \sum_{t=0}^{n_m-1} r_t^{(m)}, \qquad \widehat{\sigma}_m^2 = \frac{1}{n_m} \sum_{t=0}^{n_m-1} \left(r_t^{(m)} - \widehat{\mu}_m \right)^2. \tag{9}$$

A.2 REFERENCE DYNAMICS AND RECURSIVE RECONSTRUCTION

For $s_t^{(m,p)} = (y_t^{(m)}, \dots, y_{t-p+1}^{(m)})^\top$, the scalar reference is the best linear prediction

$$y_t^{(m)} = \sum_{j=1}^{p} \phi_{m,p,j} y_{t-j}^{(m)} + \varepsilon_t^{(m,p)}, \tag{10}$$

whose residual satisfies $\mathbb{E}[\varepsilon_t^{(m,p)} y_{t-k}^{(m)}] = 0$ for $k = 1, \dots, p$ (the orthogonality principle). Under the weakly stationary working model,

$$\gamma_m(\ell) = \mathbb{E}\left[y_t^{(m)} y_{t-\ell}^{(m)} \right], \qquad \ell \ge 0, \tag{11}$$

and orthogonality gives $\gamma_m(k) = \sum_{j=1}^{p} \phi_{m,p,j} \gamma_m(|k-j|)$ for $k = 1, \dots, p$. Equivalently,

$$\underbrace{\begin{bmatrix} \gamma_m(0) & \gamma_m(1) & \cdots & \gamma_m(p-1) \\ \gamma_m(1) & \gamma_m(0) & \cdots & \gamma_m(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_m(p-1) & \gamma_m(p-2) & \cdots & \gamma_m(0) \end{bmatrix}}_{\Gamma_{m,p}} \underbrace{\begin{bmatrix} \phi_{m,p,1} \\ \phi_{m,p,2} \\ \vdots \\ \phi_{m,p,p} \end{bmatrix}}_{\phi_{m,p}} = \underbrace{\begin{bmatrix} \gamma_m(1) \\ \gamma_m(2) \\ \vdots \\ \gamma_m(p) \end{bmatrix}}_{g_{m,p}}. \tag{12}$$

We estimate autocovariances by

$$\widehat{\gamma}_m(\ell) = \frac{1}{n_m} \sum_{t=\ell}^{n_m-1} \left(y_t^{(m)} - \bar{y}_m \right) \left(y_{t-\ell}^{(m)} - \bar{y}_m \right), \tag{13}$$

substitute them in the Yule–Walker equation, and solve the resulting Toeplitz system by Levinson recursion (Yule, 1927; Walker, 1931; Levinson, 1946). The recursion also returns the remaining

one-step prediction-error variance $\widehat{\sigma}^2_{\varepsilon,m,p}$. The fixed fitted coefficients define

$$\widehat{K}_{m,p} = \begin{bmatrix} \widehat{\phi}_{m,p,1} & \widehat{\phi}_{m,p,2} & \cdots & \widehat{\phi}_{m,p,p} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{p\times p}. \tag{14}$$

Thus Levinson recursion constructs $\widehat{\phi}_{m,p}$ across orders, whereas Equation (3) propagates fixed p-order dynamics through future steps.

The inverse representations combine transformed future values through

$$J_m(h) = \begin{cases} \{h\}, & m \in \{\mathrm{Id}, \mathcal{R}_T\}, \\ \{1,\dots,h\}, & m = \Delta_1, \\ \{h, h-T, h-2T, \dots\} \cap \mathbb{N}_{>0}, & m = \Delta_T, \end{cases} \tag{15}$$

where $\mathbb{N}_{>0}$ denotes the positive integers. Define

$$Z^{(m)}_{t,h} = \sum_{j\in J_m(h)} y^{(m)}_{t+j}, \qquad \widehat{b}_{m,p,h} = \sum_{j\in J_m(h)} (\widehat{K}^j_{m,p})^\top e_1. \tag{16}$$

The aggregate prediction is $\widehat{b}^\top_{m,p,h} s^{(m,p)}_t$, and reconstruction gives

$$x_{t+h} - \widehat{x}_{t+h} = \widehat{\sigma}_m \left(Z^{(m)}_{t,h} - \widehat{b}^\top_{m,p,h} s^{(m,p)}_t \right). \tag{17}$$

The training mean, observed reconstruction anchors, and fixed phase template cancel because they occur identically in target and prediction. The required moments are

$$\begin{aligned} a_{m,h} &:= \mathrm{Var}(Z^{(m)}_{t,h}) = \sum_{j,\ell\in J_m(h)} \gamma_m(|j-\ell|), \\ d_{m,p,h} &:= \mathrm{Cov}(s^{(m,p)}_t, Z^{(m)}_{t,h}) = \left(\sum_{j\in J_m(h)} \gamma_m(i+j) \right)_{i=0}^{p-1}. \end{aligned} \tag{18}$$

Hence, holding $\widehat{\sigma}_m$ and $\widehat{b}_{m,p,h}$ fixed,

$$\mathbb{E}[(x_{t+h} - \widehat{x}_{t+h})^2] = \widehat{\sigma}^2_m \left(a_{m,h} - 2\widehat{b}^\top_{m,p,h} d_{m,p,h} + \widehat{b}^\top_{m,p,h} \Gamma_{m,p} \widehat{b}_{m,p,h} \right). \tag{19}$$

Replacing unknown autocovariances by training estimates yields Equation (4).

A.3 Uncertainty and Level-variation Adjustment

Following the standard large-sample covariance form for autoregressive coefficient estimates (Brockwell & Davis, 1991), we use

$$\widehat{\Sigma}_{\phi,m,p} = \frac{\widehat{\sigma}^2_{\varepsilon,m,p}}{n_m} \left(\widehat{\Gamma}_{m,p} + \lambda_{m,p} I_p \right)^{-1}, \qquad \lambda_{m,p} = 10^{-10} \max\{\mathrm{tr}(\widehat{\Gamma}_{m,p})/p, 1\}. \tag{20}$$

For $r = 1,\dots,B$, draw $\delta^{(r)}_{m,p} \sim \mathcal{N}(0, \widehat{\Sigma}_{\phi,m,p})$ and set $\widehat{\phi}^{(r)}_{m,p} = \widehat{\phi}_{m,p} + \delta^{(r)}_{m,p}$. Replacing $\widehat{\phi}_{m,p}$ in the companion evolution and reconstruction gives $\widehat{b}^{(r)}_{m,p,h}$ used in Equation (5); common standard-normal directions are shared across candidates.

Let $n_b = \lfloor N/T \rfloor$, the number of complete non-overlapping blocks. With $T \le \lfloor N/10 \rfloor$, $n_b \ge 10$. Define

$$M_j = \frac{1}{T} \sum_{i=0}^{T-1} x^{(c)}_{jT+i}, \quad u_j = M_{j+1} - M_j, \quad \mathrm{VR}_c = \frac{\widehat{\mathrm{Var}}_j(u_j + u_{j+1})}{2\widehat{\mathrm{Var}}_j(u_j)}. \tag{21}$$

When increments have common variance, $\mathrm{VR}_c = 1 + \mathrm{Corr}(u_j, u_{j+1})$. Under the independent-increment reference (Lo & MacKinlay, 1988), $z_c = (\mathrm{VR}_c - 1)\sqrt{n_b}$ has an asymptotically standard-normal scale, so

$$G_c = \begin{cases} 1, & z_c > -1.96, \\ 0, & z_c \le -1.96. \end{cases} \tag{22}$$

suppresses the adjustment for pronounced alternation. Finally, with causal local level $L_t^{(c)} = T^{-1} \sum_{i=0}^{T-1} x_{t-i}^{(c)}$,

$$\Lambda_{c,H} = \frac{1}{H} \sum_{h=1}^{H} \widehat{\mathrm{Var}}_t(L_{t+h}^{(c)} - L_t^{(c)}), \qquad W_m = \begin{cases} 1, & m \in \{\mathrm{Id}, \mathcal{R}_T\}, \\ 0, & m \in \{\Delta_1, \Delta_T\}. \end{cases} \tag{23}$$

where each empirical variance uses all training indices for which both levels are defined. These definitions give D in Equation (6).

A.4 Shared-input Architectures

For architectures requiring one shared configuration across C channels, the shared structure family and extent are

$$\widehat{m}_H^{\mathrm{shared}} = \arg\min_m \frac{1}{C} \sum_{c=1}^{C} \frac{\min_{p \in \mathcal{P}} Q_{c,H}(m, p)}{\sigma_{x,c}^2}, \tag{24}$$

$$\widehat{p}_H^{\mathrm{shared}} = \mathrm{median}_c \left[\arg\min_{p \in \mathcal{P}} Q_{c,H}(\widehat{m}_H^{\mathrm{shared}}, p) \right]. \tag{25}$$

Here m ranges over the four temporal-structure families, seasonal formulations retain channel-specific lags T_c, and the median is mapped to the nearest value in $\mathcal{P}$ when necessary. Channel normalization prevents larger numerical scales from dominating the shared structure, while the median provides a robust shared extent.

B Additional Experimental Results

B.1 Full Mainstream-model Comparison

Table 4 expands the four-horizon averages in Table 1 into the corresponding per-horizon results.

Table 4: Full per-horizon comparison corresponding to Table 1.

Dataset	H	Ours		TimeFilter		DUET		iTransformer		Leddam		MSGNet		SOFTS		CrossGNN		PatchTST		Crossformer		TimesNet		DLinear		FEDformer	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	0.294	0.346	**0.313**	**0.354**	0.324	**0.354**	0.334	0.368	0.319	0.359	0.319	0.366	0.325	0.361	0.335	0.373	0.329	0.367	0.404	0.426	0.338	0.375	0.346	0.374	0.379	0.419
	192	0.330	0.368	**0.356**	0.380	0.369	**0.379**	0.377	0.391	0.369	0.383	0.376	0.397	0.375	0.389	0.372	0.390	0.367	0.385	0.450	0.451	0.374	0.387	0.382	0.391	0.426	0.441
	336	0.358	0.386	**0.386**	**0.402**	0.404	**0.402**	0.426	0.420	0.394	**0.402**	0.417	0.422	0.405	0.412	0.403	0.411	0.399	0.410	0.532	0.515	0.410	0.411	0.415	0.415	0.445	0.459
	720	0.395	0.409	**0.452**	**0.437**	0.463	**0.437**	0.491	0.459	0.460	0.442	0.481	0.458	0.466	0.447	0.461	0.442	0.454	0.439	0.666	0.589	0.478	0.450	0.473	0.451	0.543	0.490
	Avg.	0.344	0.377	**0.377**	**0.393**	0.390	**0.393**	0.407	0.410	0.386	0.397	0.398	0.411	0.393	0.403	0.393	0.404	0.387	0.400	0.513	0.496	0.400	0.406	0.404	0.408	0.448	0.452
ETTm2	96	0.163	0.255	**0.169**	0.255	0.174	0.255	0.180	0.264	0.176	**0.257**	0.177	0.262	0.180	0.261	0.176	0.266	0.175	0.259	0.287	0.366	0.187	0.267	0.193	0.293	0.203	0.287
	192	0.217	0.296	**0.235**	**0.299**	0.243	0.302	0.250	0.309	0.243	0.303	0.247	0.307	0.246	0.306	0.240	0.307	0.241	0.302	0.414	0.492	0.249	0.309	0.284	0.361	0.269	0.328
	336	0.266	0.331	**0.293**	**0.336**	0.304	0.341	0.311	0.348	0.303	0.341	0.312	0.346	0.319	0.352	0.304	0.345	0.305	0.343	0.597	0.542	0.321	0.351	0.382	0.429	0.325	0.366
	720	0.326	0.374	**0.390**	**0.393**	0.399	0.397	0.412	0.407	0.400	0.398	0.414	0.403	0.405	0.401	0.406	0.400	0.402	0.400	1.730	1.042	0.408	0.403	0.558	0.525	0.421	0.415
	Avg.	0.243	0.314	**0.272**	**0.321**	0.280	0.324	0.288	0.332	0.281	0.325	0.288	0.330	0.287	0.330	0.282	0.330	0.281	0.326	0.757	0.610	0.291	0.333	0.354	0.402	0.305	0.349
ETTh1	96	0.402	0.412	0.370	**0.394**	0.377	0.393	0.386	0.405	0.377	**0.394**	0.390	0.411	0.381	0.399	0.382	0.398	0.414	0.419	0.423	0.448	0.384	0.402	0.397	0.412	**0.376**	0.419
	192	0.429	0.430	0.413	0.420	0.429	0.425	0.441	0.436	0.424	**0.422**	0.442	0.442	0.435	0.431	0.427	0.425	0.460	0.445	0.471	0.474	0.436	0.429	0.446	0.441	**0.420**	0.448
	336	0.442	0.438	**0.450**	**0.440**	0.471	0.446	0.487	0.458	0.459	0.442	0.480	0.468	0.480	0.452	0.465	0.445	0.501	0.466	0.570	0.546	0.491	0.469	0.489	0.467	0.459	0.465
	720	0.437	0.449	**0.448**	**0.457**	0.496	0.480	0.503	0.491	0.463	0.459	0.494	0.488	0.499	0.488	0.472	0.468	0.500	0.488	0.653	0.621	0.521	0.500	0.513	0.510	0.506	0.507
	Avg.	**0.428**	0.432	0.420	0.428	0.443	0.436	0.454	0.447	0.431	**0.429**	0.452	0.452	0.449	0.442	0.437	0.434	0.469	0.454	0.529	0.522	0.458	0.450	0.461	0.457	0.440	0.460
ETTh2	96	**0.288**	0.352	0.283	0.337	0.296	0.345	0.297	0.349	0.292	**0.343**	0.328	0.371	0.297	0.347	0.309	0.359	0.302	0.348	0.745	0.584	0.340	0.374	0.340	0.394	0.358	0.397
	192	0.337	0.387	**0.362**	0.392	0.368	**0.389**	0.380	0.400	0.367	**0.389**	0.402	0.414	0.373	0.394	0.390	0.406	0.388	0.400	0.877	0.656	0.402	0.414	0.482	0.479	0.429	0.439
	336	0.365	0.411	**0.404**	0.424	0.411	**0.422**	0.428	0.432	0.412	0.424	0.435	0.443	0.410	0.426	0.426	0.444	0.426	0.433	1.043	0.731	0.452	0.452	0.591	0.541	0.496	0.487
	720	0.405	0.443	**0.407**	0.433	0.412	**0.434**	0.427	0.445	0.419	0.438	0.417	0.441	0.411	0.433	0.445	0.464	0.431	0.446	1.104	0.763	0.462	0.468	0.839	0.661	0.463	0.474
	Avg.	0.349	**0.398**	**0.364**	0.397	0.372	0.397	0.383	0.407	0.373	0.399	0.396	0.417	0.385	0.408	0.373	0.400	0.387	0.407	0.942	0.684	0.414	0.427	0.563	0.519	0.437	0.449
Weather	96	0.144	**0.201**	**0.153**	0.199	0.163	0.202	0.174	0.214	0.156	0.202	0.163	0.212	0.166	0.208	0.159	0.218	0.177	0.218	0.158	0.230	0.172	0.220	0.195	0.252	0.217	0.296
	192	0.186	0.244	**0.202**	**0.246**	0.218	0.252	0.221	0.254	0.207	0.250	0.212	0.254	0.217	0.253	0.211	0.266	0.225	0.259	0.206	0.277	0.219	0.261	0.237	0.295	0.276	0.336
	336	0.235	0.285	**0.260**	**0.289**	0.274	0.294	0.278	0.296	0.262	0.291	0.272	0.299	0.282	0.300	0.267	0.310	0.278	0.297	0.272	0.335	0.280	0.306	0.282	0.331	0.339	0.380
	720	0.304	0.335	**0.342**	**0.341**	0.349	0.343	0.358	0.349	0.343	0.343	0.350	0.348	0.356	0.351	0.352	0.362	0.354	0.348	0.398	0.418	0.365	0.359	0.345	0.382	0.403	0.428
	Avg.	0.217	0.266	**0.239**	**0.269**	0.251	0.273	0.258	0.279	0.242	0.272	0.249	0.278	0.255	0.278	0.247	0.289	0.259	0.281	0.259	0.315	0.259	0.287	0.265	0.315	0.309	0.360
Electricity	96	**0.135**	0.234	0.133	0.230	0.145	**0.233**	0.148	0.240	0.141	0.235	0.165	0.274	0.143	**0.233**	0.173	0.275	0.195	0.285	0.219	0.314	0.168	0.272	0.210	0.302	0.193	0.308
	192	0.154	**0.251**	0.154	0.248	0.163	0.248	0.162	0.253	0.159	0.252	0.184	0.292	**0.158**	0.248	0.195	0.288	0.199	0.289	0.231	0.322	0.184	0.289	0.210	0.305	0.201	0.315
	336	**0.169**	0.267	0.162	0.261	0.175	**0.262**	0.178	0.269	0.173	0.268	0.195	0.302	0.178	0.269	0.206	0.300	0.215	0.305	0.246	0.337	0.198	0.300	0.223	0.319	0.214	0.329
	720	0.202	0.295	0.184	0.284	0.204	**0.291**	0.225	0.317	**0.201**	0.295	0.231	0.332	0.218	0.305	0.231	0.335	0.256	0.337	0.280	0.363	0.220	0.320	0.258	0.350	0.246	0.355
	Avg.	**0.165**	0.262	0.158	0.256	0.172	**0.258**	0.178	0.270	0.169	0.263	0.194	0.300	0.174	0.264	0.201	0.300	0.216	0.304	0.244	0.334	0.193	0.295	0.225	0.319	0.214	0.327
Traffic	96	0.424	0.301	0.375	0.251	0.407	**0.252**	0.395	0.268	0.426	0.276	0.605	0.344	**0.376**	0.251	0.570	0.310	0.544	0.359	0.522	0.290	0.593	0.321	0.650	0.396	0.587	0.366
	192	0.440	0.304	0.395	**0.262**	0.431	**0.262**	0.417	0.276	0.458	0.289	0.613	0.359	**0.398**	0.261	0.577	0.321	0.540	0.354	0.530	0.293	0.617	0.336	0.598	0.370	0.604	0.373
	336	0.451	0.307	0.414	**0.271**	0.456	0.269	0.433	0.283	0.486	0.297	0.642	0.376	**0.415**	0.269	0.588	0.324	0.551	0.358	0.558	0.305	0.629	0.336	0.605	0.373	0.621	0.383
	720	0.480	0.315	0.445	**0.289**	0.509	0.292	0.467	0.302	0.498	0.313	0.702	0.401	**0.447**	0.287	0.597	0.337	0.586	0.375	0.589	0.328	0.640	0.350	0.645	0.394	0.626	0.382
	Avg.	0.449	0.307	0.407	**0.268**	0.451	0.269	0.428	0.282	0.467	0.294	0.641	0.370	**0.409**	0.267	0.583	0.323	0.555	0.362	0.550	0.304	0.620	0.336	0.625	0.383	0.610	0.376

B.2 Per-dataset and Per-horizon Comparisons

Table 5 compares global normalization, fixed $p = H$, channel-specific configuration, and dataset-shared configuration. The final two columns report the dataset-level choices from Exp17. A single dataset-level configuration achieves MSE 0.3132, compared with 0.3135 for channel-specific configuration; their mean task difference is -0.00028 with 95% bootstrap CI $[-0.00263, 0.00218]$. Thus, shared configuration preserves overall accuracy while reducing configuration granularity; on Traffic, it improves MSE from 0.4486 to 0.4369.

Table 5: Per-dataset and per-horizon MSE under channel-specific and dataset-shared historical-input configurations.

Dataset	H	Global norm.	Fixed $p = H$	Channel-specific	Dataset-shared	m_{ds}	p_{ds}
ETTh1	96	0.4046	0.3708	**0.4022**	0.4039	template	912
ETTh1	192	0.4545	0.4193	**0.4289**	0.4295	template	768
ETTh1	336	0.4562	0.4354	**0.4422**	0.4467	template	720
ETTh1	720	0.4671	**0.4379**	0.4369	0.4584	template	672
ETTh2	96	0.3864	0.2964	**0.2882**	0.2870	diff1	864
ETTh2	192	0.6372	0.3623	**0.3369**	0.3330	diff1	528
ETTh2	336	0.8372	0.3737	**0.3647**	0.3615	diff1	528
ETTh2	720	1.2129	0.4055	**0.4046**	0.4022	diff1	672
ETTm1	96	0.2938	0.3342	**0.2939**	0.2949	template	1728
ETTm1	192	**0.3306**	0.3399	0.3304	0.3316	template	1728
ETTm1	336	0.3598	0.3568	**0.3580**	0.3596	template	1728
ETTm1	720	0.4001	0.4117	0.3948	**0.3989**	template	1632
ETTm2	96	0.1815	0.1858	**0.1627**	0.1625	diff1	1872
ETTm2	192	0.2646	0.2316	**0.2174**	0.2162	diff1	1872
ETTm2	336	0.3614	0.2746	**0.2661**	0.2645	diff1	1872
ETTm2	720	0.5405	0.3584	0.3263	**0.3264**	diff1	1872
Weather	96	0.1416	0.1947	**0.1435**	0.1446	template	1920
Weather	192	0.1832	0.2087	**0.1857**	0.1898	seasdiff	1680
Weather	336	0.2299	0.2590	**0.2351**	0.2386	seasdiff	1680
Weather	720	0.2922	0.3236	**0.3039**	0.3060	seasdiff	1680
Electricity	96	0.1356	0.1478	**0.1347**	0.1336	template	1296
Electricity	192	**0.1519**	0.1504	0.1543	0.1555	seasdiff	1536
Electricity	336	0.1676	**0.1677**	0.1693	0.1706	seasdiff	1536
Electricity	720	**0.2026**	0.2060	0.2023	0.2064	seasdiff	1488
Traffic	96	0.4264	0.4985	**0.4237**	0.4169	template	1248
Traffic	192	**0.4382**	0.4392	0.4400	0.4279	template	1200
Traffic	336	0.4490	0.4354	0.4507	**0.4386**	template	1152
Traffic	720	0.4806	**0.4694**	0.4802	0.4643	template	1008

B.3 Channel-level Configuration Summary

Table 6 reports the number of channels assigned to each temporal structure and the distribution of configured extents. Complete channel-level records are released as supplementary CSV files.

Table 6: Channel-level temporal-structure counts and configured temporal extents.

Dataset	H	Id	Δ_1	$\mathcal{R}_T$	Δ_T	Median p	IQR p
ETTh1	96	0	2	5	0	816	264
ETTh1	192	0	2	5	0	768	360
ETTh1	336	0	2	5	0	720	600
ETTh1	720	0	1	6	0	288	528
ETTh2	96	0	7	0	0	864	768
ETTh2	192	0	7	0	0	528	600
ETTh2	336	0	6	0	1	528	600
ETTh2	720	0	5	0	2	672	600
ETTm1	96	0	2	5	0	1728	264
ETTm1	192	0	2	5	0	1728	240
ETTm1	336	0	2	5	0	1728	216
ETTm1	720	0	2	5	0	1488	696
ETTm2	96	1	6	0	0	1872	264
ETTm2	192	1	5	0	1	1872	312
ETTm2	336	1	5	0	1	1872	456
ETTm2	720	1	6	0	0	1872	552
Weather	96	3	5	7	6	1824	288
Weather	192	3	4	5	9	1632	240
Weather	336	3	4	4	10	1632	240
Weather	720	3	5	4	9	1632	240
Electricity	96	1	12	244	64	1392	624
Electricity	192	1	12	126	182	1536	480
Electricity	336	1	12	103	205	1536	480
Electricity	720	1	12	105	203	1536	576
Traffic	96	0	2	828	32	1296	768
Traffic	192	0	6	797	59	1200	768
Traffic	336	0	6	777	79	1152	720
Traffic	720	0	4	783	75	1056	720

B.4 Per-horizon Deep-forecaster Transfer

Table 7 expands the DLinear and PatchTST rows of Table 3 by forecast horizon. The configured-p arm uses the median channel-wise extent under the identity representation; the configured-(m, p) arm additionally uses the channel-majority temporal structure. For Hankel-OLS and projected SVD-Hankel, the original-structure $L = 96$, $L = H$, and configured-p arms use the Exp19 audit, while the full channel-specific (m, p) arms use the final Exp11 and Exp12 aggregates, respectively. Complete per-horizon records are provided with the supplementary results.

Table 7: Per-horizon MSE for transfer to PatchTST and DLinear.

Dataset	Forecaster	H	L=96	L=H	Configured p	Configured (m, p)
ETTh1	PatchTST	96	0.3959	0.3959	0.3787	**0.3815**
	DLinear	96	**0.3829**	**0.3829**	0.3868	0.3811
	PatchTST	192	0.4454	0.4281	0.4153	**0.4210**
	DLinear	192	0.4328	**0.4218**	0.4326	0.4160
	PatchTST	336	0.4880	0.4408	**0.4390**	0.4316
	DLinear	336	0.4914	0.4469	0.4549	**0.4485**
	PatchTST	720	0.4880	**0.4546**	0.4552	0.4438
	DLinear	720	0.5285	**0.5066**	**0.5066**	0.4751
ETTm2	PatchTST	96	**0.1774**	**0.1774**	0.1667	0.1800
	DLinear	96	0.1870	0.1870	0.1652	**0.1659**
	PatchTST	192	0.2479	0.2293	**0.2345**	0.2534
	DLinear	192	0.2720	**0.2426**	0.2223	0.2223
	PatchTST	336	0.3025	0.2769	**0.2893**	0.3300
	DLinear	336	0.3435	**0.2950**	0.2974	0.2712
	PatchTST	720	0.4002	0.3511	**0.3592**	0.4848
	DLinear	720	0.4386	0.4073	**0.3925**	0.3477
Weather	PatchTST	96	0.1766	0.1766	0.1503	**0.1523**
	DLinear	96	0.1969	0.1969	0.1669	**0.1725**
	PatchTST	192	0.2234	**0.2030**	0.1930	0.4922
	DLinear	192	0.2392	**0.2254**	0.2118	0.9154
	PatchTST	336	0.2779	**0.2474**	0.2400	0.8821
	DLinear	336	0.2835	**0.2626**	0.2527	0.8394
	PatchTST	720	0.3537	**0.3060**	0.2999	1.1649
	DLinear	720	0.3485	**0.3153**	0.3088	0.6978

B.5 Capacity Provenance

For our channel-wise scalar forecaster, the capacity entry in Table 1 is

$$\overline{p} = \frac{1}{28} \sum_{d=1}^{7} \sum_{H \in \{96,192,336,720\}} \frac{1}{C_d} \sum_{c=1}^{C_d} p_{d,H,c} = 1246.0,$$

the equal-task average number of fitted autoregressive coefficients per channel. Baseline entries are total trainable parameters in their ETTh1 $L = 96, H = 96$ reference configurations. Table 8 records both definitions explicitly; these values provide descriptive capacity context and are not an identical parameter unit. CrossGNN uses the parameter-equivalent official ETTm2 configuration with $L = 96$, $H = 96$, and seven channels because its official repository does not provide an ETTh1 script.

Table 8: Capacity values and provenance for the accuracy–capacity comparison.

Method	Capacity	Definition/source
Ours	1,246/channel	28-task mean fitted coefficients per channel
TimeFilter	764,448	Official configuration record
DUET	3,657,183	Official instantiation
iTransformer	841,568	Official configuration record
Leddam	1,129,206	Official instantiation
MSGNet	927,035	Official instantiation
SOFTS	510,816	Official instantiation
CrossGNN	10,163	Parameter-equivalent official configuration
PatchTST	35,171	Official configuration record
Crossformer	42,063,000	Published ETTh1 efficiency table
TimesNet	605,479	Published ETTh1 efficiency table
DLinear	18,624	Official instantiation
FEDformer	10,536,000	Published ETTh1 efficiency table

B.6 Controlled Temporal-Extent Selector Comparison

To isolate temporal-extent selection, we fix $m = \mathrm{Id}$ and use the same channel-wise Yule–Walker scalar AR forecaster, data splits, and candidate grid $p \in \{96, 144, \ldots, 1920\}$ for every selector. The comparison covers seven datasets, four horizons, 1,232 channels, 39 candidate orders, and 192,192 channel–horizon–order configurations. Since $m = \mathrm{Id}$, $D = 0$, and frozen Ours-Q is exactly $V + U$, not $V + 2U$. Ours-Q selections were frozen before this comparison and align with the Exp13/15 MSE, MAE, and origin audit. Validation selection reads validation outcomes; the test oracle reads test outcomes solely to quantify a nondeployable diagnostic upper bound. This is development-period evidence and was not used to modify the criterion or candidate grid.

Table 9: Controlled comparison of temporal-extent selectors under a fixed identity representation. All selectors use the same channel-wise scalar AR forecaster and candidate grid. The test oracle is a nondeployable diagnostic.

Selector	Task-equal MSE ↓	Paired reduction vs. Fixed-96 ↑	Oracle regret ↓	Ours-Q W/T/L
Fixed-96	0.553313	0.00%	72.09%	26/0/2
Fixed-H	0.433065	14.91%	27.22%	22/0/6
Random (20 seeds)	0.418741 ± 0.009926	18.57%	12.99%	23.6/0/4.4
AIC	0.418014	19.52%	11.35%	20/0/8
BIC	0.501910	10.14%	31.22%	26/0/2
Validation	**0.404198**	22.00%	6.64%	14/0/14
Ours-Q ($V + U$)	0.388830	**20.83%**	**6.78%**	–
Test oracle (diagnostic)	0.349955	25.99%	0.00%	0/0/28

Task-equal MSE first averages channels within each dataset–horizon task and then averages the 28 tasks. Paired reduction and oracle regret first compute paired channel-level ratios and then average them with equal task weight; their ordering therefore need not match task-equal MSE. W/T/L compares Ours-Q with each selector using task-level channel-mean MSE over the 28 dataset–horizon tasks. The comparison characterizes Ours-Q as a robust training-only adapter relative to short fixed histories, Fixed-H, random selection, AIC, and BIC, with performance comparable to validation selection without validation labels.

B.7 Joint versus Staged Structure–Extent Configuration

We compare the pre-specified Joint rule with deployable staged policies on the seven development datasets, covering 1,232 channels and $H \in \{96, 192, 336, 720\}$. This is development-period evidence, not frozen or blind confirmation: validation selection and the test oracle are diagnostic only and do not support the claims below. All reported values are finite, the eight required output hashes in the manifest are consistent, and the associated test suite passes 24/24 tests.

Table 10: Task-equal development-period results for structure–extent policies. The lowest numerical MSE does not replace Joint, which remains the pre-specified default.

Policy	MSE ↓	MAE ↓
Extent only	0.388830	0.374566
Sequential, $p = H$	0.312835	**0.336754**
Joint	**0.313481**	0.336653

For the two reported pre-registered primary comparisons in Table 11, positive gains favor Joint. We use 10,000 bootstrap draws with seed 0 over the 28 dataset–horizon tasks. W/T/L is computed across those tasks.

Table 11: Paired development-period comparisons. Positive gains favor Joint; intervals are 95% bootstrap confidence intervals over 28 dataset–horizon tasks.

Reference	Joint gain	W/T/L	95% CI
Extent only	+10.60%	19/0/9	[4.19, 18.10]%
Sequential, $p = H$	−0.26%	11/5/12	[−0.62, 0.06]%

We retain Joint as the pre-specified default. Although Sequential-H has a lower numerical MSE, its difference from Joint is not significant, so replacing the default post hoc would violate the development protocol. Taken together, these development-period results show that temporal structure is important relative to extent-only selection and do not establish the necessity or superiority of a simultaneous joint argmin. Complete selection distributions are provided in the supplementary CSV files.

B.8 Frozen Confirmation on PEMS08 and Solar

We evaluate the frozen method on PEMS08 and Solar, neither of which was used for method development. PEMS08 is sampled every 5 minutes and has train/validation/test lengths 10,713/3,571/3,572 with 170 eligible channels; Solar is sampled every 10 minutes and has lengths 36,792/5,256/10,512 with 137 eligible channels. We use $H \in \{96, 192, 336, 720\}$, corresponding to $\{8, 16, 28, 60\}$ hours on PEMS08 and $\{16, 32, 56, 120\}$ hours on Solar. The pre-specified defaults are $p \in \{96, 144, \ldots, 1920\}$, $B = 16$, seed 0, $T_{\max} = 400$, gate -1.96, and $\lambda_D = 1$. Training-only scores and choices were frozen before test evaluation. The score, choice, and test-result hash chain was verified; all 10,880 PEMS08 and 8,768 Solar test records are finite. We resample channels as clusters while retaining their four horizons for 10,000 bootstrap replicates. Results remain separate by dataset and frequency, with no cross-dataset headline mean. Expensive sensitivity to B/seed and $T_{\max}$ was deferred because of the submission deadline; we therefore do not claim that these quantities have been validated for robustness.

Table 12: Frozen-confirmation MSE, averaged within each dataset over four horizons. PEMS08 and Solar are reported separately.

Setting	PEMS08 MSE ↓	Solar MSE ↓
Id, $p = H$	**0.5623**	0.3118
Extent only	0.6484	**0.2062**
Joint	0.3858	0.2060

Table 12 confirms that historical-input configuration improves substantially over the fixed $\mathrm{Id}, p = H$ input on both held-out datasets. Relative to $\mathrm{Id}, p = H$, joint selection reduces MSE by 31.391% (95% CI [9.323, 59.568]%) on PEMS08 and 33.940% ([33.335, 34.533]%) on Solar. The extent-only selector provides further context: it is less favorable on PEMS08 and close to Joint on Solar.

Table 13: Frozen-confirmation MSE by horizon. Each panel is averaged within its dataset over channels; "Avg." is the four-horizon average in Table 12.

Dataset	Setting	$H = 96$	$H = 192$	$H = 336$	$H = 720$	Avg.
PEMS08	Id, $p = H$	**0.6292**	**0.6590**	**0.4325**	**0.5284**	**0.5623**
	Extent only	0.6495	0.6732	0.6236	0.6472	0.6484
	Joint	0.3436	0.3674	0.3957	0.4364	0.3858
Solar	Id, $p = H$	0.4548	0.2613	0.2870	0.2443	0.3118
	Extent only	0.1832	**0.2019**	**0.2166**	0.2230	**0.2062**
	Joint	**0.1842**	0.2011	0.2142	**0.2246**	0.2060

Table 14: Paired percentage MSE reduction of joint selection relative to each reference, with 95% channel-cluster bootstrap confidence intervals. Each resample retains all four horizons from a selected channel. Positive values favor joint selection.

Reference	PEMS08	Solar
Id, $p = H$	31.391 [9.323, 59.568]	33.940 [33.335, 34.533]
Extent only	40.502 [20.510, 45.754]	0.087 [−0.220, 0.381]

Table 15: Cheap post-freeze sensitivity diagnostics (four-horizon average MSE). The default is $\lambda_D = 1$, gate -1.96, $p_{\max} = 1920$, and step 48. These diagnostics do not alter the frozen default.

Setting	PEMS08	Solar
Default joint	0.3858	0.2060
$\lambda_D = 0$	0.4046	0.2060
$\lambda_D = 0.5$	0.4010	0.2060
$\lambda_D = 2$	0.3748	0.2060
Gate $= -1.64$	0.3860	0.2060
Gate $= -2.33$	0.3849	0.2060
$p_{\max} = 1440$	0.4518	0.2068
Step $= 96$	0.3860	0.2060
Step $= 192$	0.4079	0.2061

The diagnostics reveal complementary limitations. On PEMS08, removing D worsens MSE, whereas Solar is unchanged by D and the gate variations. The $p_{\max} = 1440$ setting is less favorable on PEMS08, while Solar changes little. The favorable post-freeze value at $\lambda_D = 2$ is not used to revise the default. The step-96 results are near the default on both datasets, while the coarser step-192 setting is less favorable on PEMS08.

B.9 Full-Candidate Alignment Between Training Criterion and Post-hoc Test Error

We separately audit ETTh1 and ETTm2 over all $4 \times 39 = 156$ (m, p) candidates for each channel–horizon cell. Each dataset contains 7 channels $\times 4$ horizons $\times 4$ representations $\times 39$ extents, yielding 4,368 training-derived Q records and 4,368 finite test-error records, with no missing or non-finite values. The training-derived Q values and configuration choices are frozen before test errors are computed; the test oracle is used solely as a post-hoc diagnostic. Results are reported separately for ETTh1 and ETTm2 and are not aggregated across datasets. This is a development-period diagnostic.

Table 16: Full-candidate alignment between the training criterion and post-hoc test error. Metrics are macro averaged within each dataset over its 28 channel–horizon cells.

Dataset	Spearman	Kendall tau-b	Weighted misorder	Oracle regret	Within 2%	Representation hit	Exact (m, p) hit
ETTh1	0.2353	0.1844	0.3865	9.52%	17.86%	53.57%	0.00%
ETTm2	0.3152	0.2821	0.3011	3.97%	39.29%	50.00%	0.00%

Table 17: Full-candidate alignment by forecast horizon. Results remain separate by dataset.

Dataset	H	Spearman	Weighted misorder	Oracle regret
ETTh1	96	-0.0539	0.5334	11.91%
ETTh1	192	0.1848	0.4152	10.31%
ETTh1	336	0.3656	0.3157	7.83%
ETTh1	720	0.4449	0.2819	8.02%
ETTm2	96	0.1835	0.3562	3.49%
ETTm2	192	0.2285	0.3548	4.73%
ETTm2	336	0.3326	0.2894	4.83%
ETTm2	720	0.5164	0.2040	2.85%

For a channel–horizon cell with candidates i, j and test errors e_i, e_j, weighted misorder is

$$\mathrm{WM}(Q, e) = \frac{\sum_{i<j} |e_i - e_j| \, \mathbf{1}[(Q_i - Q_j)(e_i - e_j) < 0]}{\sum_{i<j} |e_i - e_j|}.$$

Oracle regret is $e_{\mathrm{sel}}/e_{\mathrm{opt}} - 1$, where sel minimizes Q and opt minimizes post-hoc test error. "Within 2%" records whether $e_{\mathrm{sel}} \le 1.02 e_{\mathrm{opt}}$; representation hit and exact (m, p) hit record whether the two minimizers agree in m and in the complete configuration, respectively.

We compute every metric over the 156 candidates within each channel–horizon cell and then take a macro average within each dataset. Exact oracle recovery is zero for both datasets. The increasing rank association at longer horizons is a descriptive observation limited to these two ETT datasets; oracle regret is not monotone in the forecast horizon. These results indicate that Q captures useful information about the relative quality of candidate configurations, particularly at longer forecast horizons, although its minimizer does not generally coincide with the post-hoc test optimum.

Post-freeze explanation of limited alignment. The following diagnostics use only the already frozen choices and subsequently computed test panel; they do not feed back into Q, its candidate grid, or any primary experiment. Table 18 first distinguishes the strictness of exact recovery from residual ranking error. The test-error surface contains many near-optimal candidates, so selecting the unique test minimizer is sensitive to small post-hoc differences. Nevertheless, near ties alone cannot explain the modest rank correlations: on ETTh1, Q reaches the 2%-near-optimal set in only 17.86% of cells. Both rates exceed the random-candidate reference, but leave substantial room for improvement.

Table 18: Near-optimal test-error surfaces in the post-freeze full-candidate diagnostic. A random-candidate rate is the expected fraction of the 156 candidates within 2% of the cell oracle.

Dataset	Avg. # within 2%	Avg. # within 5%	Random within 2%	Q within 2%	Median best–second relative gap
ETTh1	9.93	26.93	6.36%	17.86%	0.181%
ETTm2	26.89	71.46	17.24%	39.29%	0.183%

To avoid comparing a four-way representation hit directly with a 39-way extent hit, we use a loss-based decomposition. Let $e^* = \min_{m,p} e(m,p)$, $(\widehat{m},\widehat{p}) = \arg\min Q$, and $e^*_{\widehat{m}} = \min_p e(\widehat{m},p)$. We define

$$r_{\text{structure}} = e^*_{\widehat{m}}/e^* - 1, \qquad r_{\text{extent}|\widehat{m}} = e(\widehat{m},\widehat{p})/e^*_{\widehat{m}} - 1.$$

Cellwise, $(1+r_{\text{total}}) = (1+r_{\text{structure}})(1+r_{\text{extent}|\widehat{m}})$; the macro means in Table 19 therefore do not add exactly. The residual loss is predominantly associated with extent ranking within the selected representation, while structure selection remains imperfect. In particular, ETTm2 representation accuracy is only modestly above its majority-class reference.

Table 19: Representation baselines and loss-based regret decomposition in the post-freeze diagnostic. Oracle representation counts are over 28 channel–horizon cells. Regrets are macro means.

Dataset	Majority-m baseline	Q m-hit	Total regret	Structure regret	Extent regret within $\widehat{m}$	Oracle m counts
ETTh1	32.14%	53.57%	9.52%	1.23%	8.17%	template 9; Δ_1 8; Id 6; Δ_T 5
ETTm2	42.86%	50.00%	3.97%	1.04%	2.90%	Δ_1 12; template 9; Id 5; Δ_T 2

Table 20 decomposes rank alignment by score components. On ETTh1, the negative association of original-scale V with test error is consistent with a working-model mismatch, but this diagnostic cannot distinguish contributions from non-stationarity, high-order AR estimation, or reconstruction error. On ETTm2, full Q has slightly lower Spearman correlation than V, so the corrections do not uniformly improve complete-candidate ranking. They nevertheless substantially reduce selection regret on both datasets, consistent with preventing costly selection errors rather than uniformly improving every pairwise ordering.

Table 20: Score-component alignment and selected-configuration regret in the post-freeze full-candidate diagnostic. V denotes the original-scale plug-in risk.

Dataset	$\rho(V,e)$	$\rho(V+U,e)$	$\rho(Q,e)$	V-only oracle regret	Q oracle regret
ETTh1	-0.4938	0.2003	0.2353	30.85%	9.52%
ETTm2	0.3260	0.2475	0.3152	20.58%	3.97%

The stronger ETTm2 alignment is consistent with a sampling-density hypothesis: its 15-minute sampling supplies more nominal observations and 10,801–11,425 test origins, compared with hourly ETTh1 and 2,161–2,785 origins; the same discrete p and H grid also corresponds to finer and shorter physical-time windows and horizons. This comparison does not identify a causal frequency effect, because dataset identity, physical windows and horizons, noise, and non-stationarity are confounded. Overall, near ties and genuine ranking error coexist in this audit; these post-freeze diagnostics do not modify Q.

B.10 Original-Scale Objective for Heterogeneous Historical Inputs

We isolate the scale on which heterogeneous historical-input formulations are compared. All three selectors use the same complete training segment, adjusted-period estimate, four formulations, order grid, channel-wise Yule–Walker AR forecaster, and test origins. Representation-space V treats every formulation as a direct standardized transformed target, so $J_{\text{rep}}(h) = \{h\}$. Original-scale V applies the formulation-specific reconstruction chain before scoring. Full $Q = V + U + D$ retains the original-scale construction and the frozen uncertainty and level-variation corrections. Each channel–horizon choice is frozen before test evaluation, and every selected forecast is reconstructed through the corresponding I_m before MSE and MAE are computed on the original series.

Table 21: Comparison of representation-space risk, original-scale risk, and the full criterion. W/T/L and paired gain use representation-space V as the reference.

Selector	MSE ↓	MAE ↓	W/T/L	Paired gain	95% CI
Representation-space V	**0.364548**	**0.372241**	–	–	–
Original-scale V	0.415068	0.385113	19/0/9	-8.95%	[-20.74, 0.77]%
Full $V + U + D$	0.313481	0.336653	25/0/3	10.68%	[6.19, 15.68]%

Original-scale V wins more tasks than it loses against representation-space V, while three large ETTh2 errors at $H = 192, 336, 720$ dominate its task-equal average. Full Q improves over representation-space V on 25 of 28 tasks, and its paired bootstrap interval remains strictly positive.

Table 22: Per-dataset and per-horizon MSE for the risk-scale comparison.

Dataset	H	Representation-space V	Original-scale V	Full Q
ETTh1	96	0.4555	**0.4530**	0.4022
ETTh1	192	0.4992	**0.4937**	0.4289
ETTh1	336	0.5240	**0.5129**	0.4422
ETTh1	720	0.5356	**0.4964**	0.4369
ETTh2	96	0.4747	**0.4739**	0.2882
ETTh2	192	**0.4164**	0.7185	0.3369
ETTh2	336	**0.5023**	0.9577	0.3647
ETTh2	720	**0.6871**	1.4424	0.4046
ETTm1	96	**0.2964**	0.2969	0.2939
ETTm1	192	0.3339	**0.3335**	0.3304
ETTm1	336	0.3626	**0.3612**	0.3580
ETTm1	720	0.4022	**0.3976**	0.3948
ETTm2	96	**0.1812**	0.1816	0.1627
ETTm2	192	**0.2623**	0.2628	0.2174
ETTm2	336	**0.3573**	0.3577	0.2661
ETTm2	720	0.5323	**0.5283**	0.3263
Weather	96	0.1469	**0.1450**	0.1435
Weather	192	0.1875	**0.1870**	0.1857
Weather	336	0.2339	**0.2345**	0.2351
Weather	720	0.2954	**0.2988**	0.3039
Electricity	96	0.1373	**0.1362**	0.1347
Electricity	192	**0.1539**	0.1524	0.1543
Electricity	336	0.1694	0.1672	**0.1693**
Electricity	720	**0.2044**	0.2023	0.2023
Traffic	96	0.4408	**0.4343**	0.4237
Traffic	192	0.4532	**0.4468**	0.4400
Traffic	336	0.4644	**0.4586**	0.4507
Traffic	720	0.4971	**0.4906**	0.4802

Table 23: Selected formulation counts and extent distributions over 4,928 channel–horizon cells.

Selector	Id	Δ_1	R_T	Δ_T	p med.	p IQR
Representation-space V	4,852	22	19	35	1,920	0
Original-scale V	37	52	4,580	259	1,920	0
Full Q	20	146	3,824	938	1,296	768

Representation-space and original-scale V agree on the complete (m, p) choice in 0.37% of cells and on m alone in 0.37%, while their p choices agree in 98.15%. Thus, reconstruction-aware scoring materially changes temporal-structure selection, whereas both uncorrected V criteria remain concentrated at the largest candidate order. The addition of U and D changes both the structure distribution and the selected extent.

Finally, we reuse the frozen 156-candidate test panels on ETTh1 and ETTm2 for a post-hoc ranking diagnostic. No diagnostic quantity feeds back into selection.

Taken together, the common original-scale construction supplies a coherent basis for comparing forecasts that obey different reconstruction rules. The full criterion then addresses the two failure modes shared by the uncorrected risks: concentration at the maximum order and sensitivity to a small number of costly configuration errors.

Table 24: Post-hoc ranking diagnostic for the three training criteria.

Dataset	Criterion	Spearman ↑	Kendall ↑	Weighted misorder ↓	Oracle regret ↓	Within 2% ↑
ETTh1	Representation-space V	**-0.3042**	**-0.2642**	**0.6622**	42.32%	**0.00%**
ETTh1	Original-scale V	-0.4938	-0.3873	0.7514	**30.85%**	**0.00%**
ETTh1	Full Q	0.2353	0.1844	0.3865	9.52%	17.86%
ETTm2	Representation-space V	0.0463	0.0623	0.4439	20.93%	25.00%
ETTm2	Original-scale V	0.3260	**0.2727**	0.2699	**20.58%**	**32.14%**
ETTm2	Full Q	**0.3152**	0.2821	**0.3011**	3.97%	39.29%