MORA: CORRECTING TIME SERIES FORECASTS WITH COARSE FUTURE SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent architectural advances have improved time-series forecasting. However, our analysis of forecasters with diverse architectures reveals a common pattern: a substantial share of residual energy is concentrated in a small number of low-frequency components. Simple linear probes further show that the forecasters' historical inputs still contain predictive information about errors in the forecast mean. Building on these findings, we propose **M**odel-**O**utput **R**esidual **A**daptation (**MORA**), a model-agnostic residual adapter that refines the outputs of frozen forecasters without requiring additional inputs. MORA estimates segment-wise means of the target sequence through structural extrapolation and learned estimation, rather than directly predicting backbone-specific residuals. It then converts discrepancies between these estimates and the corresponding statistics of the backbone forecast into stagewise corrections, scaling each update according to estimation reliability. Extensive experiments on twelve benchmarks with six forecasting backbones show average MSE reductions of **7.2%** for long-term forecasting and **28.1%** for short-term forecasting, with lower MSE in **96.2%** of the 288 evaluated cases. The benefits also extend to three time-series foundation models evaluated at different context lengths. Together, these results demonstrate that predictable coarse-scale residuals can be effectively exploited to improve existing forecasts. Code repository: `https://anonymous.4open.science/r/MORA-B0B0/`.

1 INTRODUCTION

Time-series forecasting aims to predict how a system will evolve over time based on historical observations and plays an important role in applications such as energy management, weather forecasting, and traffic flow prediction (Zhou et al., 2021; Song et al., 2020). Advances in deep learning have driven rapid progress in this area. Early approaches relied mainly on recurrent neural networks and temporal convolutional networks (Lai et al., 2018; Bai et al., 2018), followed by a growing variety of Transformer- and MLP-based models (Nie et al., 2023; Liu et al., 2024; Yi et al., 2023). These models capture temporal patterns and dependencies in different ways and have continued to improve forecasting performance.

However, recent work often pursues further gains through new architectural designs or more complex modules (Hu et al., 2025; Liu et al., 2026), while the improvements are typically incremental (Nematirad et al., 2025; Feng et al., 2026). This leads us to ask a different question: rather than continuing to refine the model architecture, do the residuals of existing forecasters exhibit systematic patterns? We analyze the residuals of forecasters with diverse architectures and find that residual energy is concentrated in a small number of low-frequency components. As shown in Figure 1(a), on Weather, the first eight DCT components account for roughly 80% of the residual energy for each of the three forecasters, despite comprising only $1/12$ of the frequency dimensions. Since these components capture overall level and slow variation, this result indicates that coarse-scale deviations constitute a substantial part of the remaining forecasting error.

We next examine whether the same observed history contains clues to these coarse-scale errors. Keeping each forecaster fixed, we fit a simple linear probe using the first eight DCT coefficients of its input history to predict the difference between the true and predicted means over the forecast horizon. Figure 1(b) further shows that, after grouping test samples by the probe output, all three

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.

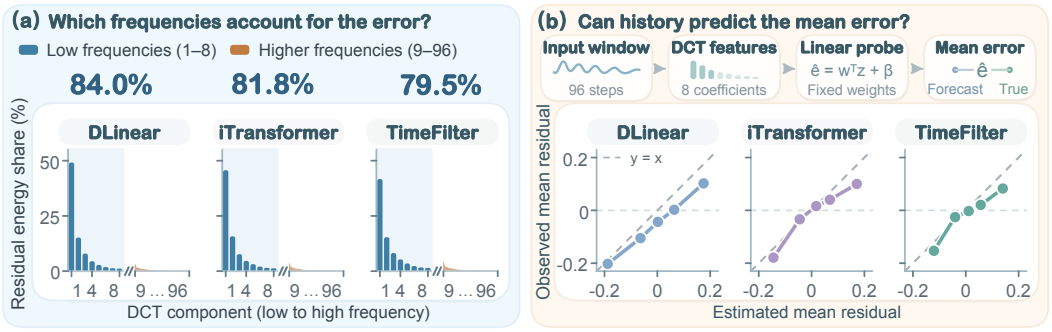

Figure 1: **Residual analysis of three forecasters on Weather. (a)** Residual energy in the first eight DCT components and the remaining components. **(b)** Test samples are grouped by the output of a history-based linear probe. Probe estimates and observed future-mean residuals are shown as group averages. Here, the future-mean residual refers to the true future mean minus the predicted future mean.

forecasters exhibit the same overall pattern: larger group-average estimates generally correspond to larger observed group-mean residuals. Thus, a simple linear probe using only the history already seen by the forecaster can still distinguish groups that are over- or underpredicted on average. Similar patterns are observed on other datasets, with additional results provided in Appendix G.

This association suggests scope for history-dependent correction beyond a constant offset. Assuming finite second moments, a nonzero population covariance between the residual along a fixed coarse-scale direction and a scalar function of history implies the existence of an affine correction with strictly lower expected MSE than the best constant correction along that direction (Proposition 1). In practice, however, estimation error can offset this potential gain, motivating corrections that account for the reliability of the estimates.

Building on this, we propose **M**odel-**O**utput **R**esidual **A**daptation (**MORA**), a model-agnostic residual adapter that operates directly on the output of a frozen backbone, without modifying the backbone or requiring additional inputs. MORA consists of two components: ***Coarse Future Estimation*** and ***Stagewise Residual Adaptation***. The former represents the future at a coarse temporal resolution using segment-wise means over contiguous intervals, and estimates these statistics from the same input history through structural extrapolation and learned estimation. Because these estimates target coarse-scale statistics of the future itself rather than backbone-specific residuals, they decouple future-state estimation from backbone-specific correction and become residual signals only when compared with the backbone forecast. The latter compares these estimates with the corresponding coarse statistics of the backbone forecast and, guided by estimation reliability, progressively converts their discrepancies into structured trajectory corrections. Rather than relearning the entire future trajectory, MORA uses the backbone forecast as its starting point and corrects only those coarse-scale deviations for which the observed history still contains predictive information.

We evaluate MORA on twelve benchmarks spanning long- and short-term forecasting, using six representative forecasting backbones. Across 288 settings, MORA reduces prediction error in **96.2%** of cases, as measured by MSE, with average MSE reductions of **7.2%** on the long-term benchmarks and **28.1%** on the short-term benchmarks. We further evaluate MORA on three time-series foundation models with frozen parameters, yielding average MSE reductions of **24.0%** and **6.9%** at context lengths of 96 and 512, respectively. These gains are also accompanied by lower MAE on average. Taken together, these results show that MORA is broadly effective across model architectures, forecast horizons, and context lengths.

In summary, our main contributions are as follows:

- By analyzing residuals from forecasters with diverse architectures, we show that the remaining errors exhibit pronounced coarse-scale structure and that the same input history used for forecasting still contains predictive signals about these errors.

- We propose **MORA**, a model-agnostic residual adapter that separates coarse future estimation from backbone-specific correction. MORA refines the outputs of frozen forecasters in stages, scaling each correction according to the reliability of the underlying estimate.
- Extensive experiments on twelve forecasting benchmarks show that MORA reduces MSE in **96.2%** of 288 long- and short-term forecasting settings across six representative backbones. Results on three time-series foundation models further demonstrate its broad effectiveness.

2 METHODOLOGY

2.1 PROBLEM SETUP

Given a historical window $\mathbf{X} \in \mathbb{R}^{L\times C}$ and its corresponding future sequence $\mathbf{Y} \in \mathbb{R}^{H\times C}$, where L, H, and C denote the lookback length, forecast horizon, and number of variables, respectively, a pretrained forecaster f_θ produces an H-step prediction

$$\widehat{\mathbf{Y}}^{(0)} = f_\theta(\mathbf{X}) \in \mathbb{R}^{H\times C}. \tag{1}$$

We freeze all parameters of f_θ and learn an output-space residual adapter that refines the backbone prediction through a structured correction,

$$\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}^{(0)} + \Delta_\phi\big(\mathbf{X}, \widehat{\mathbf{Y}}^{(0)}\big), \tag{2}$$

where $\Delta_\phi(\mathbf{X}, \widehat{\mathbf{Y}}^{(0)}) \in \mathbb{R}^{H\times C}$ estimates the predictable component of the residual $\mathbf{Y} - \widehat{\mathbf{Y}}^{(0)}$. The adapter is fitted offline with the backbone frozen. At inference, its parameters remain fixed, and it operates on the standard forecasting query and the backbone prediction without modifying the backbone or extending its input window.

2.2 MODEL-OUTPUT RESIDUAL ADAPTATION

We introduce **Model-Output Residual Adaptation (MORA)**, illustrated in Figure 2. MORA comprises two components: ***Coarse Future Estimation*** and ***Stagewise Residual Adaptation***. The first estimates a coarse representation of the future rather than backbone-specific residuals. The second compares these estimates with the backbone forecast at the same coarse scale and translates their discrepancies into structured residual updates. Estimates based on the current level and recurring patterns guide an initial update; a learned estimate is then compared with the updated forecast to refine the remaining residual. The resulting correction is calibrated, reconstructed over the forecast horizon, and added to the backbone prediction.

2.2.1 COARSE FUTURE ESTIMATION

MORA first estimates a coarse representation of the future rather than a residual tied to a particular backbone. For channel c, let $\mathbf{x}_c = \mathbf{X}_{:,c} \in \mathbb{R}^L$ and $\mathbf{y}_c = \mathbf{Y}_{:,c} \in \mathbb{R}^H$ denote its historical and future trajectories, respectively. We partition the forecast horizon into Q contiguous, non-overlapping intervals $\{\mathcal{B}_j\}_{j=1}^{Q}$ of approximately equal length and define a fixed linear operator $\mathbf{A} \in \mathbb{R}^{Q\times H}$ such that

$$\mathbf{q}_c = \mathbf{A}\mathbf{y}_c, \qquad q_{j,c} = \frac{1}{|\mathcal{B}_j|}\sum_{t\in\mathcal{B}_j} y_{t,c}. \tag{3}$$

Here $\mathbf{q}_c \in \mathbb{R}^Q$ captures the evolution of the future at a coarser temporal scale. We choose $Q = \min\{Q_{\max}, \lfloor H/b_{\min}\rfloor\}$, where $Q_{\max}$ caps the representation dimension and $b_{\min}$ sets the minimum temporal span of each interval. In our experiments, $Q_{\max} = 64$ and $b_{\min} = 4$.

We estimate the same coarse representation in two ways. First, we extrapolate simple temporal structure from the observed history. Let $\boldsymbol{\tau}$ denote the temporal coordinates of the forecast timestamps, and let $\mathcal{C}_j(\mathbf{x}_c, \boldsymbol{\tau}) \in \mathbb{R}^H$ denote the j-th extrapolated trajectory. These extrapolations use the current level, recent-pattern continuation, and, when enabled, a periodic template fitted on the training data. Each extrapolation yields

$$\mathbf{m}^e_{j,c} = \mathbf{A}\mathcal{C}_j(\mathbf{x}_c, \boldsymbol{\tau}) \in \mathbb{R}^Q, \qquad j = 1, \ldots, J. \tag{4}$$

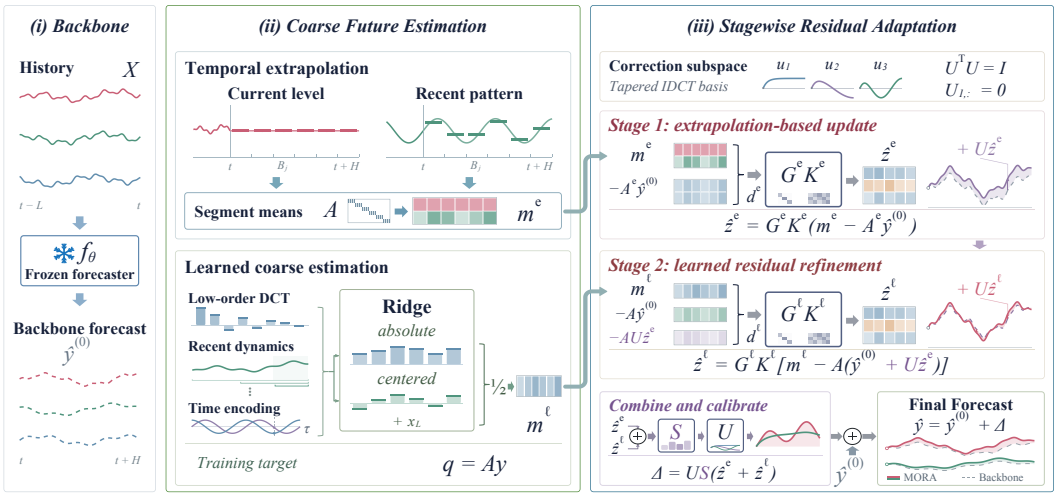

Figure 2: **Overview of MORA.** (i) A frozen backbone produces the initial forecast. (ii) Temporal extrapolation and learned coarse estimation provide complementary estimates of future segment means. (iii) These estimates guide two sequential residual updates, with the second stage addressing the discrepancy remaining after the first-stage correction. The two updates are combined, calibrated, and reconstructed into a correction that is added to the initial forecast. Curves and matrix entries are schematic; channel indices are omitted for clarity.

We retain the J estimates separately and stack them as $\mathbf{m}_c^e \in \mathbb{R}^{JQ}$. The exact extrapolation rules and periodic-template construction are given in Appendix C.1.

These temporal extrapolations provide low-capacity coarse future estimates under explicit temporal assumptions. We further introduce a learned predictor for the same coarse future target $\mathbf{q}_c$. For target channel c, we construct a deterministic feature vector $\boldsymbol{\phi}_c(\mathbf{X}, \boldsymbol{\tau}) \in \mathbb{R}^D$ from the multivariate history $\mathbf{X}$ and known temporal context $\boldsymbol{\tau}$, where D denotes the feature dimension. The feature vector summarizes low-order DCT components (Ahmed et al., 1974) of the history, whole-window and block-level statistics, multi-scale recent dynamics, and periodic time encodings. For $L = 96$, the compact representation has $D = 78$ dimensions. We fit ridge regressors (Hoerl & Kennard, 1970) separately for each target channel to map these historical features to the Q-dimensional coarse future quantity. To capture both the absolute level and changes relative to the current state, we use two coordinate systems: one retains the original coordinates, while the other is centered at the latest observation. With $x_{L,c}$ denoting the latest observation of channel c, the centered history and target are

$$\widetilde{\mathbf{x}}_c = \mathbf{x}_c - x_{L,c}\mathbf{1}_L, \qquad \widetilde{\mathbf{q}}_c = \mathbf{q}_c - x_{L,c}\mathbf{1}_Q, \tag{5}$$

where $\mathbf{1}_n$ denotes the n-dimensional all-ones vector. Let $\mathbf{m}_c^{\mathrm{abs}} \in \mathbb{R}^Q$ denote the prediction in the original coordinates and $\widetilde{\mathbf{m}}_c^{\mathrm{rel}} \in \mathbb{R}^Q$ the prediction in the centered coordinates. After restoring the latter to the original level, the learned coarse estimate is

$$\mathbf{m}_c^{\ell} = \frac{1}{2}\mathbf{m}_c^{\mathrm{abs}} + \frac{1}{2}\left(\widetilde{\mathbf{m}}_c^{\mathrm{rel}} + x_{L,c}\mathbf{1}_Q\right). \tag{6}$$

Thus, the learned predictor estimates the future coarse representation $\mathbf{q}_c$, rather than a backbone-specific residual. The exact feature construction, feature and target standardization, and ridge-fitting details are provided in Appendix C.1.

For series with many variables, fitting a separate learned estimator for every channel makes the parameter count and fitting cost scale linearly with C. We therefore replace the channel-wise estimator with a factorized coarse estimator based on a shared low-dimensional channel-factor model. Specifically, we first estimate a channel-wise periodic baseline from the training data by averaging each channel at every phase of a configured period. After subtracting this baseline and standardizing the remaining multivariate deviations, we project them into the shared factor space. A shared ridge model predicts the coarse future states of these factors, which are then mapped back to the original

channels with the channel-wise mean, scale, and periodic baseline restored. The resulting channel-level estimates retain the form $\mathbf{m}_c^\ell \in \mathbb{R}^Q$, leaving the subsequent residual adaptation unchanged. Details are provided in Appendix C.2.

Backbone-relative residual signal. The coarse future estimates become backbone-relative by comparing them with the backbone forecast at the same temporal scale. For any individual coarse estimate $\mathbf{m}_c \in \mathbb{R}^Q$ of $\mathbf{q}_c$,

$$\mathbf{d}_c = \mathbf{m}_c - \mathbf{A}\widehat{\mathbf{y}}_c^{(0)} = \mathbf{A}\left(\mathbf{y}_c - \widehat{\mathbf{y}}_c^{(0)}\right) + (\mathbf{m}_c - \mathbf{A}\mathbf{y}_c). \tag{7}$$

The first term is the backbone residual at the same coarse scale, while the second is the error of the coarse future estimate. Thus, a coarse future estimate can be converted into a backbone-relative residual signal by subtracting the corresponding coarse prediction of the backbone.

2.2.2 Stagewise Residual Adaptation

Correction subspace. The coarse representation does not uniquely determine a pointwise correction. We therefore restrict the correction to a predefined low-dimensional subspace over the forecast horizon, reducing its degrees of freedom while retaining the backbone forecast as the reference prediction. Let $\mathbf{U} \in \mathbb{R}^{H\times k}$ contain an orthonormal basis for this correction subspace, where k is the correction rank and $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$. For the backbone residual $\mathbf{r}_c = \mathbf{y}_c - \widehat{\mathbf{y}}_c^{(0)}$, we write

$$\mathbf{r}_c = \mathbf{U}\mathbf{z}_c + \mathbf{r}_{\perp,c}, \qquad \mathbf{z}_c = \mathbf{U}^\top\mathbf{r}_c, \tag{8}$$

where $\mathbf{z}_c \in \mathbb{R}^k$ contains the coordinates of the residual within the correction subspace, and $\mathbf{r}_{\perp,c} \in \mathbb{R}^H$ is the orthogonal remainder.

MORA estimates these residual coordinates as $\widehat{\mathbf{z}}_c$ and reconstructs the correction within the span of $\mathbf{U}$. We construct $\mathbf{U}$ by tapering low-order inverse discrete cosine transform (IDCT) modes near the forecast boundary and orthonormalizing the resulting columns. The first row of $\mathbf{U}$ is zero, so the correction leaves the backbone's first predicted value unchanged. For long-term forecasting, we use $k = 64$; construction details are provided in Appendix C.3.

For any individual coarse estimate $\mathbf{m}_c \in \mathbb{R}^Q$, the backbone-relative discrepancy satisfies

$$\mathbf{d}_c = \mathbf{A}\mathbf{U}\mathbf{z}_c + \boldsymbol{\eta}_c, \qquad \boldsymbol{\eta}_c = (\mathbf{m}_c - \mathbf{q}_c) + \mathbf{A}\mathbf{r}_{\perp,c}. \tag{9}$$

Here $\boldsymbol{\eta}_c \in \mathbb{R}^Q$ contains the coarse-estimation error and the contribution of residuals outside the correction space. Thus, the discrepancy $\mathbf{d}_c$ provides a noisy signal of the residual coordinates $\mathbf{z}_c$.

Residual refinement. Rather than combining the extrapolation-based and learned estimates in a single update, MORA performs residual adaptation in two stages. The extrapolation-based estimates first produce an initial correction, after which the learned estimate is evaluated against the updated forecast to refine the residual discrepancy.

Let $\mathbf{A}^e \in \mathbb{R}^{JQ\times H}$ denote J vertically stacked copies of $\mathbf{A}$. The channel-specific matrices $\mathbf{K}_c^e \in \mathbb{R}^{k\times JQ}$ and $\mathbf{K}_c^\ell \in \mathbb{R}^{k\times Q}$ map coarse discrepancies to residual coordinates. The diagonal matrices $\mathbf{G}^e, \mathbf{G}^\ell \in \mathbb{R}^{k\times k}$ provide coordinate-wise reliability scaling shared across channels. The ordered updates are

$$\begin{aligned} \mathbf{d}_c^e &= \mathbf{m}_c^e - \mathbf{A}^e\widehat{\mathbf{y}}_c^{(0)}, & \widehat{\mathbf{z}}_c^e &= \mathbf{G}^e\mathbf{K}_c^e\mathbf{d}_c^e, \\ \mathbf{d}_c^\ell &= \mathbf{m}_c^\ell - \mathbf{A}\left(\widehat{\mathbf{y}}_c^{(0)} + \mathbf{U}\widehat{\mathbf{z}}_c^e\right), & \widehat{\mathbf{z}}_c^\ell &= \mathbf{G}^\ell\mathbf{K}_c^\ell\mathbf{d}_c^\ell. \end{aligned} \tag{10}$$

Here, $\widehat{\mathbf{z}}_c^e$ and $\widehat{\mathbf{z}}_c^\ell$ denote the coefficient-space residual updates from the first and second corrections, respectively. The learned discrepancy is then computed against the updated forecast $\widehat{\mathbf{y}}_c^{(0)} + \mathbf{U}\widehat{\mathbf{z}}_c^e$, so the second update targets the residual remaining after the extrapolation-based correction.

Final calibration and prediction. The two ordered updates produce a combined correction $\widehat{\mathbf{z}}_c^e + \widehat{\mathbf{z}}_c^\ell$ in the k-dimensional coefficient space. Before reconstructing this correction over the forecast horizon, we calibrate its magnitude to account for differences across channels and basis-coordinate ranges. For channel c, the k basis coordinates are partitioned into four consecutive groups, with

one validation-fitted scale shared within each group. We collect these scales in a diagonal matrix $\mathbf{S}_c \in \mathbb{R}^{k\times k}$. Thus, although $\mathbf{S}_c$ is written as a $k \times k$ matrix, it contains only four free scalar parameters.

The final forecast is

$$\boxed{\widehat{\mathbf{y}}_c = \widehat{\mathbf{y}}_c^{(0)} + \mathbf{U}\mathbf{S}_c\left(\widehat{\mathbf{z}}_c^e + \widehat{\mathbf{z}}_c^\ell\right)}. \tag{11}$$

The matrix $\mathbf{S}_c$ calibrates the combined residual coefficients before $\mathbf{U}$ maps them back to the forecast horizon. This calibration is applied only after both residual updates and does not affect the intermediate forecast used by the learned stage. Stacking the channel-wise forecasts in Eq. equation 11 yields $\widehat{\mathbf{Y}}$, while the corresponding corrections form the residual adapter $\Delta_\phi(\mathbf{X}, \widehat{\mathbf{Y}}^{(0)})$ introduced in Eq. (2).

2.2.3 Fitting the Residual Adapter

For the coarse estimates from temporal extrapolation, the residual mapping is estimated using pairs $(\mathbf{d}_c^e, \mathbf{z}_c)$. For the learned coarse estimate, we define the remaining target as $\mathbf{z}_c^\ell = \mathbf{z}_c - \widehat{\mathbf{z}}_c^e$. The corresponding discrepancy $\mathbf{d}_c^\ell$ is computed against the forecast corrected by $\widehat{\mathbf{z}}_c^e$, following Equation equation 10. The learned residual mapping is then estimated using pairs $(\mathbf{d}_c^\ell, \mathbf{z}_c^\ell)$.

The channel-specific matrices $\mathbf{K}_c^e$ and $\mathbf{K}_c^\ell$ are computed from regularized second-order statistics of the corresponding discrepancy–target pairs. The gains $\mathbf{G}^e$ and $\mathbf{G}^\ell$ are estimated separately using a coordinate-wise reliability rule shared across channels. The complete estimators, regularization, and reliability-scaling rules are provided in Appendix C.4.

The learned coarse predictor is fitted on the training split. For estimating the residual mappings, we use blocked out-of-fold (OOF) predictions on the same split to construct the required statistics and the corresponding $\widehat{\mathbf{z}}_c^e$ used to form $\mathbf{z}_c^\ell$. Once $\mathbf{K}_c^e$, $\mathbf{K}_c^\ell$, $\mathbf{G}^e$, and $\mathbf{G}^\ell$ are fixed, the final calibration $\mathbf{S}_c$ is fitted on validation data using the calibration objective defined above. The complete fitting protocol is deferred to Appendix C.5.

3 Experiments

3.1 Experimental Setup

Datasets and backbones. We evaluate MORA on eight long-horizon forecasting datasets, ETTh1, ETTh2, ETTm1, ETTm2, Weather, ECL, Traffic, and Solar, and four short-horizon traffic datasets, PEMS03, PEMS04, PEMS07, and PEMS08. All datasets are split chronologically into training, validation, and test sets. ETT follows the standard 12/4/4-month split; the other long-horizon datasets and PEMS use ratios of 7:1:2 and 6:2:2, respectively. We select six representative backbones spanning classical architectures and recent state-of-the-art methods: DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024), TimeFilter (Hu et al., 2025), CGTFra (Yu et al., 2026), and VPNet (Liu et al., 2026). For long-horizon forecasting, we additionally evaluate three pretrained time-series foundation models: Chronos-2 (Ansari et al., 2025), Sundial (Liu et al., 2025), and TimesFM 2.5 (Das et al., 2024; Google Research, 2025). More details on the datasets and backbones are provided in Appendices D.1 and D.2.

Implementation details. Following common forecasting protocols (Liu et al., 2024), we use a lookback length of $L = 96$ for task-specific models, with $H \in \{96, 192, 336, 720\}$ for long-horizon forecasting and $H \in \{12, 24, 48, 96\}$ for short-horizon forecasting. To examine adaptation with longer histories, foundation models are evaluated at both $L = 96$ and $L = 512$, with an adapter fitted separately for each context length. All backbone parameters remain frozen during adaptation. MORA's coarse estimators and residual maps are fitted on training data, followed by final calibration on validation forecasts; neither the backbone nor the adapter is updated at test time. For ECL, Traffic, and PEMS, we use the factorized coarse estimator, while the remaining datasets use the channel-wise estimator. We report MSE and MAE, and use Imp. to denote the percentage error reduction relative to the corresponding baseline. More details on hyperparameters, fitting, and computing resources are given in Appendix D.3; results for individual horizons and seeds are provided in Appendix F.

Table 1: Baseline and MORA results for six backbones on eight long-term forecasting datasets. MSE and MAE are averaged over $H \in \{96, 192, 336, 720\}$ at $L = 96$. In Tables 1–3, bold marks the lower displayed error within each baseline–MORA comparison.

Dataset		TimeFilter		iTransformer		CGTFra		PatchTST		DLinear		VPNet	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	Base	0.418	0.427	0.455	0.447	**0.435**	**0.428**	0.448	0.449	0.462	0.459	0.436	0.429
	+MORA	**0.410**	**0.423**	**0.448**	**0.442**	0.438	0.430	**0.428**	**0.436**	**0.439**	**0.445**	**0.433**	**0.427**
	Imp. (%)	1.985	0.978	1.510	1.196	-0.546	-0.334	4.673	2.761	5.048	2.926	0.737	0.314
ETTh2	Base	0.372	0.403	0.383	0.407	0.376	0.397	0.390	0.413	0.562	0.519	**0.359**	**0.385**
	+MORA	**0.370**	**0.399**	**0.372**	**0.398**	**0.369**	**0.393**	**0.367**	**0.397**	**0.487**	**0.466**	0.363	0.389
	Imp. (%)	0.508	0.925	2.895	2.078	1.734	1.069	5.752	3.997	13.410	10.098	-1.044	-1.106
ETTm1	Base	0.377	0.394	0.407	0.411	0.393	0.390	0.387	0.402	0.403	0.408	0.376	0.382
	+MORA	**0.357**	**0.385**	**0.369**	**0.394**	**0.360**	**0.380**	**0.362**	**0.390**	**0.362**	**0.386**	**0.355**	**0.378**
	Imp. (%)	5.184	2.088	9.317	4.279	8.295	2.544	6.616	2.936	10.261	5.209	5.695	1.111
ETTm2	Base	0.271	0.321	0.290	0.335	0.278	0.318	0.293	0.336	0.352	0.400	0.274	0.315
	+MORA	**0.264**	**0.317**	**0.270**	**0.322**	**0.265**	**0.313**	**0.275**	**0.327**	**0.303**	**0.354**	**0.265**	**0.311**
	Imp. (%)	2.637	1.059	7.073	3.853	4.724	1.703	6.265	2.642	13.825	11.501	3.423	1.205
Weather	Base	0.240	0.270	0.259	0.279	0.240	0.262	0.258	0.280	0.267	0.319	0.238	0.262
	+MORA	**0.227**	**0.265**	**0.231**	**0.275**	**0.222**	**0.257**	**0.231**	**0.277**	**0.236**	**0.294**	**0.228**	**0.259**
	Imp. (%)	5.522	1.683	10.970	1.611	7.356	1.758	10.624	1.176	11.648	7.568	4.021	1.188
ECL	Base	0.164	0.261	0.175	0.267	0.167	0.255	0.204	0.294	0.209	0.295	0.164	0.252
	+MORA	**0.156**	**0.256**	**0.160**	**0.259**	**0.159**	**0.252**	**0.174**	**0.275**	**0.177**	**0.277**	**0.156**	**0.250**
	Imp. (%)	4.827	1.943	8.571	2.991	4.639	1.026	14.626	6.630	15.337	6.132	4.784	1.158
Traffic	Base	0.408	0.269	0.423	0.282	0.430	0.257	0.481	0.308	0.625	0.385	0.421	0.274
	+MORA	**0.400**	**0.266**	**0.411**	**0.277**	**0.424**	**0.256**	**0.457**	**0.297**	**0.537**	**0.343**	**0.413**	**0.271**
	Imp. (%)	1.931	1.051	2.767	1.837	1.599	0.085	5.146	3.604	14.123	10.696	1.886	0.883
Solar	Base	0.227	0.258	0.235	0.262	0.226	0.231	0.259	0.300	0.330	0.401	0.204	**0.211**
	+MORA	**0.198**	**0.250**	**0.198**	**0.251**	**0.210**	**0.230**	**0.205**	**0.270**	**0.212**	**0.278**	**0.199**	0.212
	Imp. (%)	12.910	3.043	15.829	4.282	6.917	0.253	20.638	10.044	35.827	30.677	2.557	-0.369

Table 2: Baseline and MORA results for six backbones on the four PEMS datasets. MSE and MAE are averaged over $H \in \{12, 24, 48, 96\}$ at $L = 96$.

Dataset		TimeFilter		iTransformer		CGTFra		PatchTST		DLinear		VPNet	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS03	Base	0.101	0.210	0.115	0.224	0.109	0.213	0.206	0.295	0.245	0.341	0.117	0.227
	+MORA	**0.098**	**0.206**	**0.114**	**0.216**	**0.099**	**0.203**	**0.127**	**0.232**	**0.137**	**0.241**	**0.107**	**0.215**
	Imp. (%)	3.019	2.008	1.178	3.494	9.246	4.928	38.291	21.592	44.103	29.397	8.201	5.479
PEMS04	Base	0.095	0.201	0.113	0.223	0.107	0.210	0.256	0.333	0.246	0.350	0.111	0.224
	+MORA	**0.079**	**0.180**	**0.090**	**0.193**	**0.082**	**0.183**	**0.097**	**0.196**	**0.105**	**0.201**	**0.084**	**0.187**
	Imp. (%)	16.706	10.467	20.277	13.700	23.066	12.876	62.321	41.041	57.482	42.593	24.279	16.433
PEMS07	Base	0.084	0.184	0.101	0.204	0.096	0.189	0.220	0.301	0.288	0.381	0.101	0.211
	+MORA	**0.071**	**0.167**	**0.091**	**0.178**	**0.078**	**0.166**	**0.086**	**0.185**	**0.098**	**0.197**	**0.076**	**0.176**
	Imp. (%)	15.459	9.520	10.030	12.593	18.250	12.044	60.714	38.640	65.851	48.317	24.449	16.480
PEMS08	Base	0.098	0.203	0.152	0.227	0.145	0.226	0.210	0.296	0.239	0.346	0.106	0.215
	+MORA	**0.085**	**0.187**	**0.129**	**0.210**	**0.110**	**0.201**	**0.105**	**0.209**	**0.114**	**0.221**	**0.088**	**0.193**
	Imp. (%)	13.234	7.754	14.797	7.281	23.796	11.092	50.224	29.300	52.499	36.130	16.811	10.205

3.2 Forecasting Results

Main Results. Tables 1 and 2 report horizon-averaged results on long- and short-term forecasting, respectively. Across the 192 dataset–backbone–horizon settings on the long-term benchmarks, MORA reduces MSE in 187 cases (97.4%). Averaged across the eight datasets, MORA lowers the MSE of DLinear and PatchTST by 14.9% and 9.3%, respectively. The gains remain evident for more competitive recent forecasters, with average MSE reductions of 4.4% for TimeFilter and 4.3% for CGTFra. Overall, across all six backbones, MORA reduces MSE by 7.2% on average. Similar improvements are observed on the four short-term PEMS benchmarks, where MORA improves both MSE and MAE for all 24 dataset–backbone combinations in the horizon-averaged results. Across the six backbones, average MSE reductions range from 11.6% to 55.0%. Notably, the benefit persists for TimeFilter, a strong recent backbone: MORA further reduces its MSE by 3.0%–16.7% across the four datasets, averaging 12.1% relative to the same frozen checkpoints. Together, these results show that MORA complements diverse forecasting architectures in both long- and short-term settings, including strong backbones, without updating their weights.

Table 3: Baseline and MORA results for three frozen time-series foundation models at $L = 96$ and $L = 512$. MSE and MAE are averaged over $H \in \{96, 192, 336, 720\}$ separately for each context length.

Dataset		Chronos-2				Sundial				TimesFM 2.5			
		$L = 96$		$L = 512$		$L = 96$		$L = 512$		$L = 96$		$L = 512$	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	Base	0.492	0.439	0.418	**0.412**	0.473	0.435	0.397	0.412	0.484	0.435	0.421	**0.409**
	+MORA	**0.456**	**0.431**	**0.402**	0.415	**0.436**	**0.428**	**0.387**	**0.409**	**0.435**	**0.429**	**0.403**	0.412
	Imp. (%)	7.370	1.880	3.900	-0.753	7.739	1.610	2.463	0.682	10.157	1.439	4.437	-0.866
ETTh2	Base	0.406	0.405	0.352	0.377	0.399	0.410	0.344	0.383	0.413	0.406	0.372	0.388
	+MORA	**0.384**	**0.393**	**0.345**	0.377	**0.389**	**0.405**	**0.340**	**0.378**	**0.397**	**0.404**	**0.365**	**0.384**
	Imp. (%)	5.413	3.038	2.031	0.005	2.598	1.068	1.001	1.170	3.895	0.512	1.898	1.184
ETTm1	Base	0.616	0.477	0.387	0.372	0.537	0.457	0.366	0.370	0.875	0.584	0.381	0.372
	+MORA	**0.366**	**0.391**	**0.338**	**0.360**	**0.369**	**0.390**	**0.338**	**0.365**	**0.370**	**0.390**	**0.341**	**0.366**
	Imp. (%)	40.496	18.127	12.787	3.163	31.329	14.632	7.521	1.389	57.682	33.247	10.354	1.629
ETTm2	Base	0.318	0.338	0.280	0.311	0.304	0.340	0.275	0.320	0.372	0.378	0.291	0.318
	+MORA	**0.275**	**0.315**	**0.257**	**0.299**	**0.271**	**0.320**	**0.252**	**0.308**	**0.277**	**0.321**	**0.273**	**0.309**
	Imp. (%)	13.376	6.946	8.182	3.574	10.916	5.858	8.162	3.754	25.500	14.896	6.079	2.624
Weather	Base	0.307	0.291	0.258	0.252	0.285	0.302	0.244	0.275	0.314	0.299	0.263	0.268
	+MORA	**0.243**	**0.290**	**0.213**	**0.250**	**0.233**	**0.291**	**0.213**	**0.262**	**0.230**	**0.283**	**0.215**	**0.249**
	Imp. (%)	20.807	0.330	17.467	0.761	18.407	3.803	12.890	4.853	26.704	5.584	18.340	6.779
ECL	Base	0.234	0.294	0.163	0.244	0.204	0.284	0.148	0.243	0.239	0.308	0.164	0.249
	+MORA	**0.184**	**0.274**	**0.154**	**0.241**	**0.173**	**0.270**	**0.143**	**0.240**	**0.187**	**0.283**	**0.156**	**0.246**
	Imp. (%)	21.589	6.881	5.026	1.022	14.910	4.971	2.899	0.979	22.001	8.226	4.987	1.161
Traffic	Base	0.574	0.313	0.392	**0.238**	0.604	0.355	0.458	0.290	0.599	0.334	0.382	0.244
	+MORA	**0.518**	**0.301**	**0.389**	0.239	**0.531**	**0.327**	**0.441**	**0.283**	**0.544**	**0.316**	**0.379**	0.244
	Imp. (%)	9.796	3.687	0.846	-0.171	12.143	7.929	3.734	2.131	9.119	5.503	0.748	0.082
Solar	Base	0.528	0.327	0.187	**0.193**	0.666	0.530	0.229	0.257	0.913	0.520	0.281	**0.250**
	+MORA	**0.215**	**0.286**	0.187	0.194	**0.210**	**0.278**	**0.208**	**0.256**	**0.222**	**0.279**	**0.222**	0.261
	Imp. (%)	59.281	12.521	0.174	-0.196	68.421	47.491	9.212	0.170	75.707	46.260	21.222	-4.503

Time-Series Foundation Models. Table 3 reports the results of MORA on three time-series foundation models using their officially released pretrained checkpoints. Since all three models support substantially longer context windows at inference, we evaluate $L = 512$ in addition to $L = 96$ to examine whether MORA remains effective when the backbone is provided with more historical information. Each adapted forecast is compared against its unadapted counterpart using the same context length. At $L = 96$, MORA reduces MSE by 22.3%, 20.8%, and 28.8% for Chronos-2, Sundial, and TimesFM 2.5, respectively, averaged equally across the eight datasets. Even at $L = 512$, where the foundation models can draw on substantially more historical context, MORA yields further average MSE reductions of 6.3%, 6.0%, and 8.5% for the same three models. MORA also reduces MAE on average at both context lengths. Together with the results on task-specific forecasters, these findings show that MORA provides a broadly applicable output-space refinement mechanism whose benefits persist across model families and context lengths.

3.3 Ablation Study

We first conduct component ablations to examine the roles of MORA's coarse future estimates and stagewise residual updates. We then compare alternative adaptation strategies to assess the benefit of deriving residual corrections from coarse future estimates.

Table 4: Component ablations on TimeFilter.

	ETTm1		ETTm2		ECL	
Variant	MSE	MAE	MSE	MAE	MSE	MAE
TimeFilter	0.452	0.437	0.390	0.393	0.206	0.302
w/o learned	0.427	0.431	0.389	0.390	0.194	0.291
w/o extrapolation	0.436	0.427	0.380	0.391	0.195	0.294
Independent	0.427	0.429	0.379	0.389	0.191	0.290
MORA	**0.422**	**0.426**	**0.376**	**0.388**	**0.190**	**0.288**

Table 5: Alternative strategies on TimeFilter.

	ETTm1		Weather		Solar	
Strategy	MSE	MAE	MSE	MAE	MSE	MAE
TimeFilter	0.313	0.354	0.153	0.199	0.199	0.239
Affine calibration	0.314	0.357	0.153	0.228	0.200	0.239
Direct residual	**0.300**	0.357	0.150	0.211	0.181	0.239
δ-Adapter (O)	0.370	0.415	0.156	0.224	0.192	0.243
δ-Adapter (I+O)	0.312	0.356	0.153	0.201	**0.171**	0.234
MORA	0.304	**0.350**	**0.148**	**0.198**	0.173	**0.231**

O: output-only; I+O: input–output.

Component Ablations. Table 4 examines the contributions of the two sources of coarse future estimates and the design of the second-stage correction. Removing either *structural extrapolation* or the *learned estimator* increases both MSE and MAE on all three datasets. Although both sources target the same coarse future quantities, neither alone matches the performance of the full adapter, indicating that they provide complementary correction signals. With both sources retained, we compare the stagewise design against an *independent-update variant* in which both corrections are fitted to the original backbone residual. The stagewise design achieves slightly lower MSE and MAE on all three datasets. This result suggests that the learned correction benefits from accounting for the first-stage correction, allowing it to target the remaining residual rather than fitting another correction to the original backbone residual.

Comparison with Alternative Strategies. Table 5 compares MORA with *affine calibration*, *direct residual regression*, and two variants of δ-Adapter (Liang et al., 2026). The direct regressor uses the same historical features and correction basis as MORA, with a comparable parameter count, but predicts the backbone residual directly. Affine calibration improves neither MSE nor MAE. The learned alternatives reduce MSE in several settings, but their gains do not consistently extend to MAE. Direct residual regression achieves the lowest MSE on ETTm1, while δ-Adapter (I+O) does so on Solar. MORA achieves the lowest MAE on all three datasets while reducing MSE relative to the unadapted backbone in every case. These results suggest that deriving corrections from coarse future estimates provides an effective adaptation strategy, yielding consistent improvements in both metrics across the evaluated settings. Additional ablations and experimental details are provided in Appendix E.

3.4 Low-Frequency Residual Analysis

To examine whether MORA effectively corrects low-frequency forecasting errors, we visualize the low-frequency residuals of DLinear and TimeFilter on Solar before and after adaptation. As shown in Figure 3, MORA reduces low-frequency reconstruction MSE by 41.0% and 13.9%, respectively. The corresponding overall forecasting MSE decreases by 38.9% and 12.8%. The reduction in low-frequency error accounts for nearly all of the overall MSE improvement in these settings, supporting the intended role of MORA's coarse future estimates in correcting coarse-scale deviations.

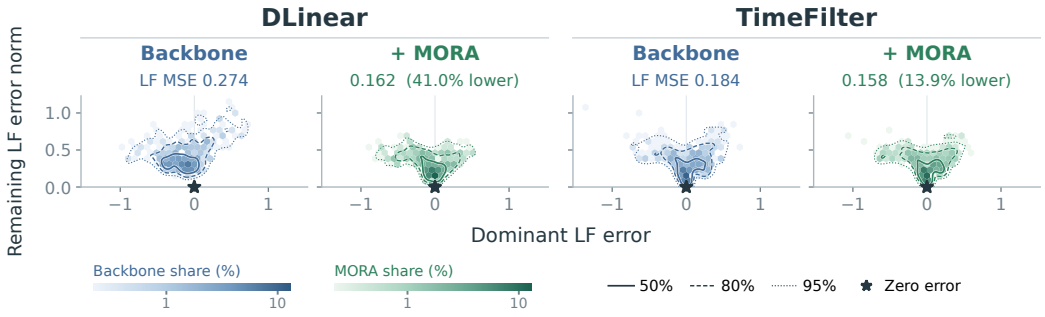

Figure 3: **Low-frequency residuals before and after applying MORA on Solar.** DCT-8 residuals are represented by their signed projection onto a shared principal direction fitted on the validation set, together with the norm of the orthogonal residual. The squared distance from the origin corresponds to the low-frequency MSE.

4 Conclusion

We presented MORA, a model-agnostic residual adapter for frozen time-series forecasters. Motivated by the coarse-scale structure of forecasting errors, MORA estimates segment-wise future means and uses their discrepancies with the corresponding means of backbone forecasts to construct stagewise corrections, scaled by estimation reliability. Experiments on long- and short-term forecasting benchmarks demonstrate broad improvements across diverse architectures, including time-series foundation models at different context lengths. These findings highlight coarse future estimation as an effective complement to existing forecasters.

AI USE STATEMENT

We used generative AI tools to assist with dataset preparation and reformatting, manuscript organization, drafting, translation, language editing, and LaTeX typesetting. We carefully reviewed all AI-assisted content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We describe MORA's coarse future estimation and stagewise residual adaptation in the main paper. The estimator definitions, correction subspace, and fitting and calibration procedures are detailed in Appendix C. Dataset splits, preprocessing, backbone configurations, hyperparameters, and computing resources are documented in Appendix D. Additional ablations and horizon-wise results are provided in Appendices E and F, respectively. The anonymous code repository is linked in the abstract.

REFERENCES

N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. doi: 10.1109/T-C.1974.223784. URL https://doi.org/10.1109/T-C.1974.223784.

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

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. URL https://arxiv.org/abs/1803.01271.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Wanjin Feng, Yuan Yuan, Jingtao Ding, and Yong Li. Beyond model ranking: Predictability-aligned evaluation for time series forecasting. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.23074.

Google Research. TimesFM 2.5: 200m pytorch checkpoint. Official model release and model card, 2025. URL https://huggingface.co/google/timesfm-2.5-200m-pytorch. Checkpoint identifier: google/timesfm-2.5-200m-pytorch.

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. URL https://doi.org/10.1080/00401706.1970.10488634.

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 of *Proceedings of Machine Learning Research*, pp. 24893–24911, 2025. URL https://proceedings.mlr.press/v267/hu25ac.html.

Jeonghwan Im and Hyuk-Yoon Kwon. COSA: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2a8ce71baac4c89bf9ff479d8240c7d9-Abstract-Conference.html.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025. doi: 10.1609/aaai.v39i17.33965. URL https://ojs.aaai.org/index.php/AAAI/article/view/33965.

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.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018. URL https://arxiv.org/abs/1703.07015.

Daojun Liang, Qi Li, Yinglong Wang, Jing Chen, Hu Zhang, Xiaoxiao Cui, Qizheng Wang, and Shuo Li. The forecast after the forecast: A post-processing shift in time series. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=syfWdclGE1.

Kun Liu, Renjun Jia, Ruifeng Yang, Xirui Zeng, Yuqi Liang, and Cen Chen. Are global dependencies necessary? scalable time series forecasting via local cross-variate modeling. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a832b65da26c4867a8581cb13d46ad6d-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*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html.

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 *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06625.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.00816.

Max Planck Institute for Biogeochemistry. Weather data from the jena meteorological station. Meteorological data archive, n.d. URL https://www.bgc-jena.mpg.de/wetter/. Accessed September 25, 2026.

Reza Nematirad, Anil Pahwa, and Balasubramaniam Natarajan. Are data embeddings effective in time series forecasting? *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=yeu44ZRvZZ.

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://arxiv.org/abs/2211.14730.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 914–921, 2020. doi: 10.1609/aaai.v34i01.5438. URL https://ojs.aaai.org/index.php/AAAI/article/view/5438.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

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*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/701251e1db4a2e4dd2ef23f5265d5936-Abstract-Conference.html.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper/2023/hash/f1d16af76939f476b5f040fd1398c0a3-Abstract.html.

Wuqing Yu, Weichen Guo, Jian Zhou, Shuyu Luo, and Jiacai Zhang. Robust inter-series dependency modeling for time series forecasting via information-theoretic alignment. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=YQPQuHVLLQ.

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. URL https://arxiv.org/abs/2205.13504.

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. URL https://arxiv.org/abs/2012.07436.

Tian Zhou, Ziqing Ma, Xue Wang, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, and Rong Jin. FiLM: Frequency improved Legendre memory model for long-term time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/524ef58c2bd075775861234266e5e020-Abstract.html.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286, 2022b. URL https://proceedings.mlr.press/v162/zhou22g.html.

A Related Work

Time-Series Forecasting. Early neural forecasting models used recurrent and convolutional networks to capture temporal patterns, as exemplified by LSTNet (Lai et al., 2018). Subsequent Transformer-based approaches extended this line of research to long-horizon prediction: Informer (Zhou et al., 2021) improved attention efficiency, while Autoformer (Wu et al., 2021) incorporated progressive trend–seasonal decomposition. Spectral structure became another useful modeling tool, with FEDformer combining decomposition and frequency-domain representations and FiLM using Fourier filtering alongside Legendre memory (Zhou et al., 2022a;b). Later work explored simpler predictors and alternative representations. DLinear (Zeng et al., 2023) revisited decomposition-based linear forecasting, and PatchTST (Nie et al., 2023) combined temporal patching with channel-independent modeling. FreTS (Yi et al., 2023) and FITS (Xu et al., 2024) further developed spectral forecasting through frequency-domain MLPs and complex-valued interpolation, respectively, while iTransformer (Liu et al., 2024) used variate tokens to model cross-variate dependencies. More recent approaches have refined multivariate dependency modeling through patch-specific graph filtration in TimeFilter (Hu et al., 2025), local cross-variate interactions in VPNet (Liu et al., 2026), and information-theoretic alignment in CGTFra (Yu et al., 2026). Across these designs, temporal and cross-variate structure is incorporated into the predictor through its architecture or representation.

Time-Series Foundation Models. Large-scale pretraining has broadened forecasting beyond dataset-specific model development, allowing a single model to learn from heterogeneous time-series collections and transfer to new tasks. Chronos (Ansari et al., 2024) formulates forecasting as language modeling over quantized time-series values, whereas TimesFM (Das et al., 2024) adopts a patched decoder-only architecture. Moirai (Woo et al., 2024) further supports heterogeneous sampling frequencies and variable dimensions within a unified forecasting model. Subsequent developments have expanded both the modeling approach and the range of supported tasks: Sundial (Liu et al., 2025) uses flow matching for continuous-valued generative forecasting, while Chronos-2 (Ansari et al., 2025) supports univariate, multivariate, and covariate-informed forecasting through group attention. These models provide reusable forecasting backbones that can be applied without task-specific fine-tuning, creating a natural setting for studying how their predictions can be refined without modifying pretrained parameters.

Forecast Adaptation and Post-Processing. Beyond backbone design and pretraining, another line of research improves forecasts by addressing non-stationarity or adapting existing predictions. RevIN (Kim et al., 2022) removes and restores instance-level statistics, while Non-stationary Transformers (Liu et al., 2022) combine series stationarization with de-stationary attention to preserve informative non-stationary structure. These mechanisms are incorporated into the forecasting pipeline rather than fitted as standalone residual corrections. Test-time adaptation instead adjusts a deployed forecasting system as new observations become available. TAFAS (Kim et al., 2025) uses partially observed targets for online calibration, and COSA (Im & Kwon, 2026) updates a gated output-space residual adapter using recently observed ground truth. Closer to our setting, δ-Adapter (Liang et al., 2026) learns lightweight input perturbations and output residual corrections around a frozen forecaster. MORA also keeps the backbone frozen, but separates coarse future estimation from backbone-specific residual adaptation. Its coarse estimators predict segment-wise means of the future from the same historical window. Backbone-specific residual mappings then translate discrepancies between these estimates and the corresponding coarse statistics of the current forecast into reliability-scaled, stagewise corrections. Unlike TAFAS and COSA, MORA is fitted offline, and both the backbone and adapter remain fixed throughout testing.

B HISTORY-DEPENDENT RESIDUAL CORRECTION

This appendix proves Proposition 1 from the introduction. The result connects a history-dependent residual signal to a reduction in expected MSE beyond the best constant correction along the same coarse-scale direction. We first give the population least-squares argument, then relate it to the forecast-mean probe and the residual coordinates used by MORA.

B.1 IMPROVEMENT BEYOND A CONSTANT CORRECTION

Following the problem setup, keep the pretrained forecaster f_θ frozen and consider channel c, with backbone residual $\mathbf{r}_c = \mathbf{y}_c - \widehat{\mathbf{y}}_c^{(0)} \in \mathbb{R}^H$. Let $\mathbf{u} \in \mathbb{R}^H$ be a fixed unit vector along a coarse-scale direction, and let $p = g(\mathbf{X})$ be a scalar signal from the same historical window available to the backbone. The function g need not be linear; one example is the fixed linear probe used in the introduction. All expectations below refer to the joint distribution of $(\mathbf{X}, \mathbf{Y})$, with f_θ and g held fixed. Define

$$z = \mathbf{u}^\top \mathbf{r}_c, \qquad \mu_z = \mathbb{E}[z], \qquad \mu_p = \mathbb{E}[p], \qquad \sigma_p^2 = \mathrm{Var}(p), \qquad \gamma = \mathrm{Cov}(z, p). \tag{12}$$

Here z is a scalar projection of the backbone residual; its relation to the vector $\mathbf{z}_c = \mathbf{U}^\top \mathbf{r}_c$ is given in Appendix B.3. Assume $\mathbb{E}\|\mathbf{r}_c\|_2^2 < \infty$ and $\mathbb{E}[p^2] < \infty$. For coefficients $a, b \in \mathbb{R}$, the corrected forecast is $\widehat{\mathbf{y}}_c^{(0)} + \mathbf{u}(a + bp)$, with expected MSE

$$\mathcal{R}_c(a, b) = \frac{1}{H} \mathbb{E}\left[\|\mathbf{r}_c - \mathbf{u}(a + bp)\|_2^2 \right]. \tag{13}$$

The factor $1/H$ averages the error over the forecast horizon. A constant correction has $b = 0$: its coefficient is the same for every query, while its temporal shape is fixed by $\mathbf{u}$.

Proposition 1 (History-dependent correction beyond a constant offset) *Under the assumptions above, suppose $\sigma_p^2 > 0$. The best constant correction along $\mathbf{u}$ has coefficient μ_z. The optimal affine correction has*

$$b^\star = \frac{\gamma}{\sigma_p^2}, \qquad a^\star = \mu_z - b^\star \mu_p, \qquad a^\star + b^\star p = \mu_z + \frac{\gamma}{\sigma_p^2}(p - \mu_p). \tag{14}$$

Its improvement over the best constant correction is exactly

$$\underbrace{\min_a \mathcal{R}_c(a, 0)}_{\mathcal{R}_{\mathrm{const}}} - \underbrace{\min_{a,b} \mathcal{R}_c(a, b)}_{\mathcal{R}_{\mathrm{aff}}} = \frac{\gamma^2}{H \sigma_p^2}. \tag{15}$$

Consequently, $\mathrm{Cov}(z, p) \neq 0$ *implies a strict reduction in expected MSE beyond the best constant correction along that direction.*

Proof. Since $\|\mathbf{u}\|_2 = 1$ and $\mathbf{u}^\top(\mathbf{r}_c - \mathbf{u}z) = 0$, the squared error decomposes pointwise as

$$\|\mathbf{r}_c - \mathbf{u}(a + bp)\|_2^2 = \|\mathbf{r}_c - \mathbf{u}z\|_2^2 + (z - a - bp)^2. \tag{16}$$

The first term is unaffected by either coefficient. Centering z and p gives

$$\mathbb{E}[(z - a - bp)^2] = \mathrm{Var}(z) - 2b\gamma + b^2\sigma_p^2 + (a + b\mu_p - \mu_z)^2. \tag{17}$$

For any fixed b, the last term is minimized at $a = \mu_z - b\mu_p$. Completing the square in the remaining terms yields

$$\mathrm{Var}(z) - 2b\gamma + b^2\sigma_p^2 = \mathrm{Var}(z) - \frac{\gamma^2}{\sigma_p^2} + \sigma_p^2 \left(b - \frac{\gamma}{\sigma_p^2} \right)^2. \tag{18}$$

Because $\sigma_p^2 > 0$, the minimizing slope is $b^\star = \gamma/\sigma_p^2$. Restricting instead to $b = 0$ gives the best constant coefficient $a = \mu_z$. Their minimum scalar squared errors differ by γ^2/σ_p^2. The common orthogonal term cancels, and division by H proves Eq. (15). ∎

If $\sigma_p^2 = 0$, the affine and constant correction classes coincide; if $\sigma_p^2 > 0$ but $\gamma = 0$, the chosen signal yields no additional improvement over the best constant correction within the affine class.

B.2 Connection to the Forecast-Mean Probe

For the forecast-mean probe, take $\mathbf{u} = \mathbf{1}_H/\sqrt{H}$ and define the future-mean residual $\bar{r}_c = H^{-1}\mathbf{1}_H^\top \mathbf{r}_c$, namely the true future mean minus the predicted future mean. Then $z = \sqrt{H}\bar{r}_c$, and Proposition 1 gives

$$\mathcal{R}_{\text{const}} - \mathcal{R}_{\text{aff}} = \frac{\text{Cov}(\bar{r}_c, p)^2}{\text{Var}(p)}. \tag{19}$$

The corresponding optimal correction is

$$\mathbf{u}(a^\star + b^\star p) = \mathbf{1}_H \left[\mathbb{E}[\bar{r}_c] + \frac{\text{Cov}(\bar{r}_c, p)}{\text{Var}(p)}(p - \mathbb{E}[p])\right]. \tag{20}$$

This history-dependent level correction is added to every forecast step. It can therefore reduce the MSE of the full future sequence without changing the deviations around the predicted future mean.

Because $p = g(\mathbf{X})$, the law of total expectation also gives

$$\text{Cov}(z, p) = \text{Cov}(\mathbb{E}[z \mid \mathbf{X}], p)\,. \tag{21}$$

Nonzero covariance rules out a conditional mean $\mathbb{E}[z \mid \mathbf{X}]$ that is constant almost surely. Hence, the historical window contains predictive information about the residual beyond a constant offset. The diagnostic probe differs from MORA's learned coarse estimator, which predicts the future quantity $\mathbf{q}_c = \mathbf{A}\mathbf{y}_c$ rather than a backbone-specific residual.

B.3 Connection to Stagewise Residual Adaptation

The correction subspace in Section 2.2.2 makes the connection to the residual coordinates explicit. For any fixed $\mathbf{v} \in \mathbb{R}^k$ with $\|\mathbf{v}\|_2 = 1$, orthonormality of $\mathbf{U}$ gives

$$\mathbf{u} = \mathbf{U}\mathbf{v}, \qquad \|\mathbf{u}\|_2 = 1, \qquad z = \mathbf{v}^\top \mathbf{z}_c. \tag{22}$$

Thus Proposition 1 applies to a direction within MORA's correction subspace whenever its covariance condition holds. The forecast-mean direction illustrates the motivation, whereas the tapered basis used by MORA leaves the first predicted value unchanged. A mean-probe association alone does not establish nonzero covariance for every direction in that subspace.

The remaining issue is estimation reliability. For $\sigma_p^2 > 0$ and any fixed fitted coefficients $(\widehat{a}, \widehat{b})$ in the affine family above, the difference from the best constant correction is

$$\begin{aligned}\mathcal{R}_c(\widehat{a}, \widehat{b}) - \mathcal{R}_{\text{const}} = \frac{1}{H}\Big[&(\widehat{a} + \widehat{b}\mu_p - \mu_z)^2 \\ &+ \sigma_p^2(\widehat{b} - b^\star)^2 - \frac{\gamma^2}{\sigma_p^2}\Big].\end{aligned} \tag{23}$$

The two nonnegative terms quantify coefficient-estimation error, which can offset the population improvement. This motivates accounting for estimation reliability, as in MORA's coordinate-wise reliability scaling $\mathbf{G}^e$ and $\mathbf{G}^\ell$. These gains are estimated from training data; the final calibration $\mathbf{S}_c$ is fitted separately on validation data, following Section 2.2.3.

Proposition 1 establishes the existence of an improving correction under the evaluation distribution. It does not guarantee that finite-sample estimates, or coefficients fitted on a different distribution, achieve that improvement. The grouped probe results provide empirical evidence relevant to the covariance condition, rather than a proof of it. Window overlap leaves the risk identity unchanged but must be considered when estimating uncertainty.

C Additional Method Details

C.1 Coarse Future Estimation

Temporal extrapolation. The current-level estimate extends the latest observation over the forecast horizon. Recent-pattern continuation instead repeats the last $P_{\text{eff}} = \min\{P, L\}$ observations, where P is the configured base period in sampling steps. For $h = 1, \ldots, H$,

$$\begin{aligned}\big[\mathcal{C}_1(\mathbf{x}_c, \boldsymbol{\tau})\big]_h &= x_{L,c},\\ \big[\mathcal{C}_2(\mathbf{x}_c, \boldsymbol{\tau})\big]_h &= x_{L-P_{\text{eff}}+1+((h-1) \bmod P_{\text{eff}}),c}.\end{aligned} \tag{24}$$

Both estimates use only the supplied historical window. When $P > L$, recent-pattern continuation repeats the available window rather than accessing observations outside it.

Training-fitted periodic template. When enabled, the periodic estimate uses a channel-wise harmonic template fitted on the training data,

$$T_c(s) = a_{0,c} + \sum_{p \in \mathcal{P}} \left[a_{p,c} \cos(2\pi s/p) + b_{p,c} \sin(2\pi s/p)\right]. \tag{25}$$

Here s denotes the absolute sampling index. The period dictionary $\mathcal{P}$ combines configured daily and weekly periods with separated spectral peaks identified from detrended training series. The harmonic coefficients are estimated by regularized least squares. For a query ending at index s_0, we estimate a local level offset $b_c^{\text{loc}} = \sum_{\ell=1}^{L} w_\ell [x_{\ell,c} - T_c(s_0 - L + \ell)]$, where the normalized exponential weights w_ℓ give greater weight to recent observations. The extrapolated future is $[\mathcal{C}_3(\mathbf{x}_c, \boldsymbol{\tau})]_h = T_c(s_0 + h) + b_c^{\text{loc}}$. Each trajectory is mapped through $\mathbf{A}$; the resulting estimates are retained separately in $\mathbf{m}_c^e$.

The dictionary contains at most eight periods, including P and $7P$, with the remaining periods selected from training-only spectral peaks. Harmonic fitting uses ridge regularization of 10^{-3} under an average squared-loss convention. The local level offset uses normalized exponential weights with a half-life of 16 sampling steps.

Historical features. The compact feature vector contains 78 entries at $L = 96$: 16 orthonormal DCT coefficients, eight whole-window statistics, 12 block means, 36 recent-dynamics features, and six time encodings. The whole-window statistics comprise the mean, standard deviation, first and last values, endpoint difference, linear slope, and the mean and standard deviation of first differences. Recent dynamics comprise 12 block means of the last 24 observations, 12 block means of their first differences, and the mean, standard deviation, slope, and endpoint difference over the last 8, 16, and 24 observations. Time encodings are sine–cosine pairs evaluated at the forecast origin for periods P, $P/2$, and $7P$.

The channel-wise estimator also uses extended feature representations. At $L = 96$, the full representation has 252 dimensions, additionally including raw history, block standard deviations, recent increments, and cross-channel summaries. Adding interactions between historical state features and temporal encodings gives 348 dimensions. These representations use the same historical window and coarse future target, without changing the stagewise residual updates.

Channel-wise ridge estimation. We fit the absolute and last-centered views separately. For view $v \in \{\text{abs}, \text{rel}\}$, let $\boldsymbol{\Phi}_c^v$ and $\mathbf{Q}_c^v$ collect the standardized historical features and coarse targets over the fitting samples. The channel-wise coefficient matrix is estimated as

$$\mathbf{W}_c^v = \arg\min_{\mathbf{W}} \|\mathbf{Q}_c^v - \boldsymbol{\Phi}_c^v \mathbf{W}\|_F^2 + \lambda_v \|\mathbf{W}\|_F^2. \tag{26}$$

Feature and target statistics are computed on the fitting samples, with standard-deviation floors of 10^{-5} and 10^{-4}, respectively. Predictions are restored to the target scale, and the centered prediction is additionally shifted by $x_{L,c}\mathbf{1}_Q$. The two estimates are then averaged as in Equation (6). The fitting and regularization-selection protocol is specified in Appendix C.5.

C.2 Factorized Coarse Estimation

Periodic baseline. Let $\mathbf{x}_t \in \mathbb{R}^C$ denote the multivariate observation at absolute sampling index t, and let P_{f} be the configured period. For training indices $\mathcal{T}$, define $\mathcal{T}_t = \{s \in \mathcal{T} : s \bmod P_{\mathrm{f}} = t \bmod P_{\mathrm{f}}\}$. The channel-wise periodic baseline is

$$\mathbf{b}(t) = \frac{1}{|\mathcal{T}_t|} \sum_{s \in \mathcal{T}_t} \mathbf{x}_s. \tag{27}$$

Unlike the harmonic template used for temporal extrapolation, this baseline directly averages observations sharing the same phase.

Shared factor space. We subtract the periodic baseline and standardize the remaining deviations using their training means $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma}$:

$$\boldsymbol{\xi}_t = (\mathbf{x}_t - \mathbf{b}(t) - \boldsymbol{\mu}) \oslash \boldsymbol{\sigma}, \qquad \mathbf{s}_t = \mathbf{V}^\top \boldsymbol{\xi}_t. \tag{28}$$

Here $\oslash$ denotes elementwise division, with standard deviations floored at 10^{-5}. The columns of $\mathbf{V} \in \mathbb{R}^{C \times d_{\mathrm{f}}}$ are the leading d_{f} eigenvectors of the training covariance of $\boldsymbol{\xi}_t$, satisfying $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$. The resulting factor states summarize variation shared across channels.

Coarse factor prediction. For a query ending at s_0, the factor-space target for interval $\mathcal{B}_j$ is

$$\mathbf{q}_j^{\mathrm{f}} = \frac{1}{|\mathcal{B}_j|} \sum_{h \in \mathcal{B}_j} \mathbf{s}_{s_0+h}. \tag{29}$$

The predictor uses the latest factor state; means over the last 6, 12, 24, 48, and 96 steps; differences from the states 6, 12, and 24 steps earlier; slopes over the last 24 steps; and the six time encodings described in Appendix C.1. At $L = 96$, these features contain $10d_{\mathrm{f}}+6$ entries. A single multi-output ridge regressor is fitted to the concatenated absolute targets $\mathbf{q}_j^{\mathrm{f}}$ and last-relative targets $\mathbf{q}_j^{\mathrm{f}} - \mathbf{s}_{s_0}$ for all factors and intervals. The relative predictions are restored to absolute factor coordinates by adding $\mathbf{s}_{s_0}$. Training-internal rolling validation selects the ridge strength and one fixed prediction rule: the absolute prediction, the restored relative prediction, or their equal-weight average. The selected rule remains fixed at inference. Feature and target standardization follows the channel-wise ridge estimator. Denote the selected prediction by $\widehat{\mathbf{q}}_j^{\mathrm{f}}$.

Reconstruction in channel space. The factor predictions are mapped back to the original channels, with the channel-wise mean, scale, and periodic baseline restored:

$$[\mathbf{m}_c^\ell]_j = \frac{1}{|\mathcal{B}_j|} \sum_{h \in \mathcal{B}_j} b_c(s_0 + h) + \mu_c + \sigma_c \mathbf{V}_{c,:} \widehat{\mathbf{q}}_j^{\mathrm{f}}. \tag{30}$$

We use $d_{\mathrm{f}} = 32$ and $P_{\mathrm{f}} = 7P$, where $P = 24$ for ECL/Traffic and $P = 288$ for PEMS. All train-dependent quantities follow the fitting protocol in Appendix C.5. The resulting $\mathbf{m}_c^\ell \in \mathbb{R}^Q$ is used in the stagewise updates without changing their formulation.

C.3 Correction Subspace

Cosine modes and boundary taper. For a horizon H and correction rank $k < H$, let $\mathbf{B} \in \mathbb{R}^{H \times k}$ contain the first k orthonormal IDCT modes. For $h = 1, \ldots, H$ and $j = 0, \ldots, k-1$,

$$B_{h,j+1} = a_j \cos\left(\frac{\pi j (h - \frac{1}{2})}{H}\right), \qquad a_0 = H^{-1/2}, \quad a_j = \sqrt{2/H} \text{ for } j \geq 1. \tag{31}$$

We apply a taper over the first n_{b} forecast steps, where $2 \leq n_{\mathrm{b}} \leq H$:

$$w_h^{\mathrm{tap}} = \begin{cases} \sin^2\left(\dfrac{\pi}{2} \dfrac{h-1}{n_{\mathrm{b}}-1}\right), & 1 \leq h \leq n_{\mathrm{b}}, \\ 1, & n_{\mathrm{b}} < h \leq H. \end{cases} \tag{32}$$

The taper starts at zero and reaches one at n_{b}. The tapered modes and their reduced QR factorization are

$$\mathbf{B}_{\mathrm{tap}} = \mathrm{diag}(w_1^{\mathrm{tap}}, \ldots, w_H^{\mathrm{tap}}) \mathbf{B} = \mathbf{U}\mathbf{R}, \tag{33}$$

where $\mathbf{U} \in \mathbb{R}^{H \times k}$ has orthonormal columns. This orthonormalization preserves the span of the tapered modes.

Boundary preservation. The first row of $\mathbf{B}_{\mathrm{tap}}$ is zero. Since $\mathbf{R}$ is nonsingular, the first row of $\mathbf{U}$ is also zero. Consequently, the final correction satisfies

$$\left[\mathbf{U}\mathbf{S}_c\left(\widehat{\mathbf{z}}_c^e + \widehat{\mathbf{z}}_c^\ell\right)\right]_1 = 0, \qquad \widehat{y}_{1,c} = \widehat{y}_{1,c}^{(0)}. \tag{34}$$

The constraint therefore preserves the backbone's first predicted value; it does not force that value to equal the last observation. The basis is computed once for each configuration and shared across channels. Because tapering modifies the cosine modes, its span is not identical to an untapered DCT truncation.

Dimensions. For long-term forecasting, we use $k = 64$ and $n_{\mathrm{b}} = 13$. The fixed short-horizon settings for PEMS are listed in Table 9. These settings retain the same construction with $k < H$.

C.4 Estimating the Residual Updates

Stage-specific signals. For the extrapolation-based stage, we use $(\mathbf{d}_c^e, \mathbf{z}_c)$ and $\mathbf{M}^e = \mathbf{A}^e\mathbf{U}$. For the learned stage, we use $(\mathbf{d}_c^\ell, \mathbf{z}_c^\ell)$ and $\mathbf{M}^\ell = \mathbf{A}\mathbf{U}$. We describe the estimator for either stage below, omitting the stage superscript. Let p be the dimension of $\mathbf{d}_c$ and define $\boldsymbol{\eta}_{i,c} = \mathbf{d}_{i,c} - \mathbf{M}\mathbf{z}_{i,c}$.

Residual mapping. From N training pairs, we form the uncentered second-order moments

$$\mathbf{C}_{ab,c} = \frac{1}{N}\sum_{i=1}^{N} \omega_{i,c}\mathbf{a}_{i,c}\mathbf{b}_{i,c}^\top, \qquad a, b \in \{z, d, \eta\}. \tag{35}$$

No sample mean is subtracted. The weights are $\omega_{i,c} = s_{i,c}^{-2\gamma}$, where $s_{i,c} = \max\{\mathrm{RMS}(\mathbf{d}_{i,c}), \underline{s}_c\}$. The floor $\underline{s}_c$ is the larger of 10^{-5} and the training tenth percentile of the discrepancy RMS. Setting $\gamma = 0$ gives uniform weights. These weights affect moment estimation; the deployed mapping remains linear in $\mathbf{d}_c$.

We apply diagonal shrinkage $\mathcal{D}_\beta(\mathbf{C}) = (1-\beta)\mathbf{C} + \beta\,\mathrm{Diag}(\mathbf{C})$, where $\mathrm{Diag}(\mathbf{C})$ retains only the diagonal entries. Writing $\overline{\mathbf{C}}_{zz,c} = \mathcal{D}_\beta(\mathbf{C}_{zz,c})$ and $\overline{\mathbf{C}}_{\eta\eta,c} = \mathcal{D}_\beta(\mathbf{C}_{\eta\eta,c})$, the regularized mapping is

$$\begin{aligned} \boldsymbol{\Sigma}_c^d &= (1-\rho)\left(\mathbf{M}\overline{\mathbf{C}}_{zz,c}\mathbf{M}^\top + \nu\overline{\mathbf{C}}_{\eta\eta,c}\right) + \rho\mathbf{C}_{dd,c}, \\ \mathbf{K}_c &= \left[(1-\rho)\overline{\mathbf{C}}_{zz,c}\mathbf{M}^\top + \rho\mathbf{C}_{zd,c}\right](\boldsymbol{\Sigma}_c^d + \epsilon_c\mathbf{I}_p)^{-1}. \end{aligned} \tag{36}$$

Here $\beta, \rho \in [0,1]$, $\nu > 0$, and $\epsilon_c = 10^{-6}\max\{\mathrm{tr}(\boldsymbol{\Sigma}_c^d)/p, 10^{-12}\}$. The parameter ρ interpolates between structured and empirical second-order estimates. The mapping is computed by solving the corresponding linear system, rather than explicitly forming its inverse.

Reliability scaling. Let $\widetilde{z}_{i,c,r}$ be the unscaled prediction for coordinate r from a mapping fitted without the temporal block containing sample i. For a set of training samples $\mathcal{I}$, define

$$\alpha_r(\mathcal{I}) = \frac{\sum_{i\in\mathcal{I}}\sum_c \widetilde{z}_{i,c,r}z_{i,c,r}}{\sum_{i\in\mathcal{I}}\sum_c \widetilde{z}_{i,c,r}^2}. \tag{37}$$

The ratio is set to zero when its denominator does not exceed 10^{-12}. We use the pooled rule for $\mathbf{G}^e$ and the blockwise rule for $\mathbf{G}^\ell$:

$$\begin{aligned} g_r^{\mathrm{pool}} &= \mathrm{clip}_{[0,g_{\max}]}\big(\alpha_r(\mathcal{I})\big), \\ g_r^{\mathrm{block}} &= \max\left\{0, \min\left[g_{\max}, \alpha_r(\mathcal{I}), \min_b \alpha_r(\mathcal{I}_b)\right]\right\}, \end{aligned} \tag{38}$$

where $\mathcal{I}_b$ is temporal block b and $g_{\max} = 1.5$. Each stage uses its own targets and cross-fitted predictions. Both rules pool channels, producing one gain per residual coordinate. The gains form the diagonal matrices $\mathbf{G}^e$ and $\mathbf{G}^\ell$ and remain fixed at inference. Parameter selection and block construction are described in Appendix C.5.

C.5 Fitting and Calibration

Training signals. We construct blocked out-of-fold (OOF) forecasts on the training split. The latter half of the training-window origins is partitioned into four contiguous blocks, and held queries are sampled at stride six. For each block, task-specific backbone replicas exclude its raw-time span and any training windows whose histories or targets overlap that span. Train-dependent coarse estimators follow the same exclusion. The replicas are used only to construct training signals; deployment uses the original frozen backbone. For foundation models, pretrained weights remain fixed throughout, while the local coarse estimators are fitted using the corresponding training partitions. Coarse-predictor ridge strengths are selected from $\{1, 10, 100, 1000, 10000\}$ using training-internal rolling validation with an $L + H - 1$ separation between fitting and held window origins.

Stagewise fitting. The first-stage mapping is cross-fitted over the training blocks. Its held-block predictions are used to estimate $\mathbf{G}^e$ according to Appendix C.4. After applying this gain, the resulting first-stage coefficients define the remaining target $\mathbf{z}_c^\ell$ and the adjusted discrepancy $\mathbf{d}_c^\ell$ used to fit the second stage. The second-stage mapping is cross-fitted over the same blocks, and its held-block predictions determine $\mathbf{G}^\ell$. Each mapping is then refitted on all of its corresponding training pairs. The learned coarse predictor continues to target the future coarse quantities, not the remaining backbone residual. Train-dependent coarse estimators are fitted on the full training split for validation and test inference.

Regularization settings. The residual mappings use parameter values drawn from $\beta \in \{0.5, 1\}$, $\nu \in \{0.5, 1, 2\}$, $\rho \in \{0, 0.25, 0.5, 1\}$, and $\gamma \in \{0, 0.5\}$. Candidates are ranked by their minimum relative reduction in residual-coordinate squared error across training blocks, using the blockwise gain rule. Mean block improvement breaks ties. The factorized realization uses the fixed setting $(\beta, \nu, \rho, \gamma) = (1, 1, 0, 0)$ in both stages.

Validation calibration. Once both update mappings and their gains are fixed, we fit $\mathbf{S}_c$ using validation forecasts. The k coordinates are partitioned into four consecutive groups, with one scale $s_{c,g} \in [0, 1.5]$ per channel and group. We divide the validation queries into four contiguous blocks and minimize the equally weighted normalized objective

$$\mathcal{L}_{\mathrm{cal}} = \frac{1}{4} \sum_{b=1}^{4} \left[0.1 \frac{\mathrm{MSE}_b}{\mathrm{MSE}_b^{(0)}} + 0.9 \frac{\mathrm{MAE}_{b,\epsilon}}{\mathrm{MAE}_b^{(0)}} \right]. \tag{39}$$

Each numerator is evaluated from the calibrated forecasts and averaged over queries, forecast steps, and channels within its block. Superscript (0) denotes the corresponding unadapted backbone error; both denominators are floored at 10^{-12}. The smoothed absolute error replaces $|e|$ by $\sqrt{e^2 + \epsilon^2}$, with $\epsilon = \max\{10^{-3} \frac{1}{4} \sum_b \mathrm{MAE}_b^{(0)}, 10^{-7}\}$. Starting from unit scales, we perform at most eight iteratively reweighted least-squares updates, each solving a four-dimensional box-constrained quadratic problem per channel. Iteration stops when the largest scale change is below 10^{-5}.

The resulting calibration is applied only to the sum of the two updates; it does not alter the intermediate forecast used by the learned stage. At inference, the backbone, coarse estimators, residual mappings, reliability gains, and calibration remain fixed.

D Additional Experimental Details

D.1 Datasets and Forecasting Windows

Datasets. Table 6 summarizes the twelve benchmarks, including their dimensions, sampling intervals, forecast horizons, and split sizes. **ETT** (Zhou et al., 2021) records oil temperature and six load measurements from two electricity transformers, with hourly and 15-minute versions. **Weather** (Max Planck Institute for Biogeochemistry, n.d.) contains meteorological measurements from the Max Planck Institute for Biogeochemistry in Jena. **ECL**, **Traffic**, and **Solar** use the LSTNet releases (Lai et al., 2018) and record electricity consumption, road occupancy, and photovoltaic generation, respectively. **PEMS03/04/07/08** use the STSGCN traffic-flow releases (Song et al., 2020); we forecast the first feature of each sensor in the supplied arrays. The data sources are the ETT repository, MPI-BGC weather archive, LSTNet data repository, and STSGCN repository.

Chronological splits. ETT uses the first 20 benchmark months, divided into 12/4/4 months for training/validation/test; each month contains 30 days in the released loader. The remaining long-horizon datasets use 7:1:2 splits, and PEMS uses 6:2:2. Boundaries follow the corresponding loaders. Table 6 reports raw observations, not overlapping forecasting windows.

Table 6: Forecasting benchmarks. C is the number of variables and H is the forecast horizon in time steps. Split sizes count raw observations before windowing.

Dataset	C	Horizons H	Split sizes Train / Val / Test	Sampling	Domain
ETTh1, ETTh2	7	$\{96, 192, 336, 720\}$	(8640, 2880, 2880)	Hourly	Electricity
ETTm1, ETTm2	7	$\{96, 192, 336, 720\}$	(34560, 11520, 11520)	15 min	Electricity
Weather	21	$\{96, 192, 336, 720\}$	(36887, 5270, 10539)	10 min	Weather
ECL	321	$\{96, 192, 336, 720\}$	(18412, 2632, 5260)	Hourly	Electricity
Traffic	862	$\{96, 192, 336, 720\}$	(12280, 1756, 3508)	Hourly	Transportation
Solar	137	$\{96, 192, 336, 720\}$	(36792, 5256, 10512)	10 min	Energy
PEMS03	358	$\{12, 24, 48, 96\}$	(15724, 5242, 5242)	5 min	Transportation
PEMS04	307	$\{12, 24, 48, 96\}$	(10195, 3398, 3399)	5 min	Transportation
PEMS07	883	$\{12, 24, 48, 96\}$	(16934, 5645, 5645)	5 min	Transportation
PEMS08	170	$\{12, 24, 48, 96\}$	(10713, 3571, 3572)	5 min	Transportation

Long-horizon preprocessing. Each channel is standardized using its training mean and population standard deviation, with unit scale for constant channels. These statistics are reused on validation and test data. Forecast origins advance by one step; the future stays within its evaluation split, while the initial history may extend into the preceding split. Each baseline–MORA pair uses identical forecast origins, as do the foundation-model comparisons between $L = 96$ and $L = 512$. Incomplete evaluation batches are retained.

PEMS preprocessing. For a record of length T, split boundaries are $\lfloor 0.6T \rfloor$ and $\lfloor 0.8T \rfloor$; windows stay within their split without borrowing preceding-split history. iTransformer and CGTFra use training-fitted standardization and test stride one. TimeFilter uses split-specific standardization and test stride 12, also adopted by the PatchTST, DLinear, and VPNet transfer experiments. The latter convention uses test-split normalization statistics, but does not fit MORA parameters on test labels. Every local baseline/MORA pair shares preprocessing and window enumeration, including the endpoint convention. Absolute normalized errors are not directly comparable across the two preprocessing groups.

D.2 Models and Inference

Task-specific backbones. We use the released architectures and training objectives, with model- and dataset-specific configurations at $L = 96$. Optimizers, batch sizes, and training budgets follow these configurations. Checkpoints are selected on validation data before adaptation and then

frozen. Each local baseline–MORA pair uses the same checkpoint; backbone hyperparameters are not retuned during adaptation.

PEMS baselines. TimeFilter, iTransformer, and CGTFra reproductions are checked against the corresponding paper versions and horizon-specific references. TimeFilter at $H = 96$ is an extension without an original-paper score. PatchTST, DLinear, and VPNet have no matching PEMS results in the selected source papers; their transfer recipes retain the original training principles while adapting to $L = 96$, the PEMS channel count, and TimeFilter's preprocessing. VPNet uses its Traffic recipe with patch length 12 and stride six.

Table 15 marks six quoted baseline entries with †: five from Table 9 of iTransformer (Liu et al., 2024) and one from the archived CGTFra public submission, rather than a verified final-ICML result table. The remaining 90 baseline entries are local reproductions, and all +MORA forecasts use local checkpoints. These six comparisons use literature references rather than paired local checkpoints. Local paired scores remain in the result records.

Foundation models. Table 7 lists the frozen checkpoints. Chronos-2 uses its multivariate interface and released median forecast; TimesFM 2.5 and Sundial use their univariate interfaces, returning the released point forecast and the mean of 20 sampled trajectories, respectively. These choices are fixed within each baseline/MORA pair. The evaluation does not establish whether the benchmark datasets were excluded from the external models' pretraining corpora.

Table 7: Pretrained checkpoints used at both context lengths. All foundation-model weights remain frozen.

Model	Checkpoint identifier
Chronos-2	`amazon/chronos-2`
Sundial	`thuml/sundial-base-128m`
TimesFM 2.5	`google/timesfm-2.5-200m-pytorch`

D.3 Adapter Configuration and Fitting

Tables 8 and 9 collect the shared configuration and fixed short-horizon dimensions. The correction basis, residual estimators, and fitting procedure are specified in Appendices C.3, C.4, and C.5. The feature and periodic-template configurations are kept fixed within each dataset–backbone group across horizons, without horizon-wise test selection or per-query fallback.

Table 8: Shared MORA configuration. PEMS-specific dimensions are given in Table 9.

Setting	Value or selection rule
Long-horizon correction rank	$k = 64$
Coarse intervals	$Q = \min\{64, \lfloor H/4 \rfloor\}$
Coarse-predictor ridge candidates	$\{1, 10, 100, 1000, 10000\}$
Channel-factor dimension	32, for the factorized estimator
Factor-predictor training stride	One time step
Residual-fitting OOF blocks	Four; held origins in the later half of training
Residual-fitting origin stride	Six time steps
Final calibration	Four coefficient groups per channel
Calibration scale range	$[0, 1.5]$; fitted on validation data
Test-time parameter updates	None

Dataset-specific fitting. ECL, Traffic, and PEMS use the factorized estimator described in Appendix C.2; other datasets use the channel-wise estimator. This dataset-level choice was fixed during development on the inspected benchmarks. For PEMS OOF replicas, checkpoint selection uses a 6% training-internal validation interval; the official validation and test splits are not used for replica fitting or checkpoint selection.

Table 9: PEMS dimensions, fixed across datasets and backbones. Boundary taper is measured in forecast steps.

Forecast horizon H	12	24	48	96
Coarse intervals Q	3	6	12	24
Correction rank k	8	16	32	64
Boundary taper	2	3	6	13

Repeated runs reuse the residual mappings fitted for the initial run. For Sundial, runs with the channel-wise estimator refit validation calibration; runs with the factorized estimator reuse the complete adapter. These repetitions vary inference or adaptation, not the pretrained foundation-model weights.

Computing environment. Independent experiment jobs run across eight NVIDIA GeForce RTX 4090 GPUs with 24 GB each. Model-specific environments follow the released implementations: VPNet uses Python 3.8.20, PyTorch 1.12.1, and CUDA 11.3; CGTFra uses Python 3.9.24, PyTorch 2.4.1, and CUDA 12.1. The foundation-model environments use PyTorch 2.7.1. Chronos-2 and TimesFM 2.5 use Transformers 4.57.6, and Sundial uses Transformers 4.40.1. The corresponding packages are `chronos-forecasting` 2.3.2 and `timesfm` 2.0.2.

D.4 Repeated Runs, Metrics, and Aggregation

The long-horizon task-specific experiments use one run per horizon on ETTh1, ETTm1, ETTm2, and Weather, and three backbone seeds on ETTh2, ECL, Traffic, and Solar. DLinear, PatchTST, TimeFilter, and VPNet use seed 2021 for the single-run datasets and seeds 2021–2023 for repeated datasets. CGTFra uses seed 2023 and seeds 2023–2025, respectively. iTransformer uses seed 2021 for single-run settings except ETTm2 at $H = 96$, which uses seed 2023; repeated datasets use seeds 2023–2025. PEMS retains one local seed per backbone: 2021 for TimeFilter/PatchTST/DLinear/VPNet and 2023 for iTransformer/CGTFra.

Chronos-2 and TimesFM 2.5 each use one fixed pretrained checkpoint. Sundial uses sampling seed 2023 on ETTh1, ETTm1, ETTm2, and Weather, and seeds 2023–2025 on ETTh2, ECL, Traffic, and Solar. The result tables report mean errors over the recorded runs for each setting; a single-run entry retains the error from that run. Available standard deviations are retained in the result records and measure training or sampling variability, not uncertainty over independent test episodes.

For N queries, MSE and MAE average over queries, forecast steps, and variables:

$$\begin{aligned} \mathrm{MSE} &= \frac{1}{NHC}\sum_{i,t,c}(y_{i,t,c}-\widehat{y}_{i,t,c})^2, \\ \mathrm{MAE} &= \frac{1}{NHC}\sum_{i,t,c}|y_{i,t,c}-\widehat{y}_{i,t,c}|. \end{aligned} \tag{40}$$

Local baseline–MORA pairs use identical queries, targets, and normalization. Let $E^{(0)}_{h,s}$ and $E_{h,s}$ denote the baseline and adapted errors at horizon h for run s. We first average the S_h runs within a horizon, then give each horizon equal weight:

$$\overline{E}^{(0)} = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}}\frac{1}{S_h}\sum_{s=1}^{S_h}E^{(0)}_{h,s}, \qquad \overline{E} = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}}\frac{1}{S_h}\sum_{s=1}^{S_h}E_{h,s}. \tag{41}$$

For either metric, the reported improvement is

$$\mathrm{Imp.}(E) = 100\,\frac{\overline{E}^{(0)}-\overline{E}}{\overline{E}^{(0)}}. \tag{42}$$

Imp. is computed from unrounded mean errors, not by averaging seed- or horizon-specific percentages. Dataset-level Imp. values are then averaged equally for model-level summaries, so additional runs do not increase a dataset's weight.

Long-horizon and foundation-model tables use local paired baselines. For PEMS, Imp. uses the displayed reference score for the six quoted entries and the local baseline otherwise. Thus, the literature-reference comparison is distinct from the local paired comparison. Win counts are descriptive because overlapping windows and shared checkpoints make the evaluation settings dependent.

D.5 Variation Across Datasets and Metrics

Dataset-dependent gains. The benefit of residual adaptation depends on the dataset–backbone pair. For TimeFilter, the horizon-averaged MSE reduction is 12.910% on Solar, 1.931% on Traffic, and 0.508% on ETTh2. CGTFra on ETTh1 and VPNet on ETTh2 increase both errors, while VPNet on Solar improves MSE but not MAE. The full results retain these cases, so average gains should not be read as improvements in every setting.

Metric-dependent gains. At $L = 512$, all 24 foundation-model–dataset pairs improve in MSE, but only 19 improve in MAE. For example, TimesFM 2.5 on Solar reduces MSE by 21.222% while increasing MAE by 4.503%. Squared loss gives larger errors greater weight than absolute loss. The two metrics therefore capture different aspects of a correction, and a reduction in MSE need not produce a reduction in MAE.

Reference-dependent gains. On PEMS, the gains for DLinear and PatchTST are measured against their local transfer baselines. Across 96 settings, local paired comparisons improve MSE and MAE in 92 and 94 cases; using the displayed reference baselines gives 90 and 94, respectively. For PEMS08/iTransformer at $H = 96$, the adapted MSE matches the quoted 0.221 at three decimals despite a larger gain over the local baseline. These comparisons distinguish adaptation gains from margins over published reference values.

E ADDITIONAL ABLATIONS

Tables 4 and 5 use TimeFilter (seed 2021) with $L = 96$, at $H = 720$ and $H = 96$, respectively. The component examples are a retrospectively selected subset of the full study; ECL uses the channel-wise estimator in that comparison. Methods share the same frozen checkpoint and evaluation windows within each setting.

Individual extrapolation sources. We remove each extrapolation source in turn, retain the learned estimate, and refit the residual adapter (Table 10). The study covers eight datasets, TimeFilter, iTransformer, PatchTST, and DLinear, and $H \in \{96, 720\}$ at $L = 96$, using the channel-wise estimator. For either error metric E, entries average the paired change $100(E_{\text{removed}}/E_{\text{full}} - 1)$ over matched settings; positive values indicate higher error after removal. Removing recent-pattern continuation or the static periodic template increases both errors on average over their matched settings.

Table 10: Contribution of individual extrapolation sources. Entries are mean relative error changes (%) after removal and refitting. The last two rows use the same 40 template-enabled settings.

Estimate removed	**Settings**	**MSE change**	**MAE change**
Current level	64	+0.038	-0.173
Recent-pattern continuation	40	+0.906	+0.796
Static periodic template	40	+1.603	+0.380

Additional comparisons. Table 11 extends the input–output δ-Adapter comparison to five further settings at $L = H = 96$. MORA achieves lower MSE in all five and lower MAE in three, including both metrics on ECL/TimeFilter. These results complement the three-dataset comparison in the main text.

Table 11: Additional comparisons with δ-Adapter (input + output) at $L = H = 96$. MORA uses the factorized estimator on ECL and Traffic and the channel-wise estimator elsewhere.

Dataset	**Backbone**	**MSE**			**MAE**		
		Base	δ (I+O)	MORA	Base	δ (I+O)	MORA
ECL	PatchTST	0.180	0.161	**0.150**	0.273	**0.252**	0.253
Traffic	PatchTST	0.459	0.432	**0.431**	0.298	**0.278**	0.285
ETTm2	PatchTST	0.183	0.182	**0.171**	0.268	0.269	**0.258**
ETTh2	PatchTST	0.299	0.298	**0.285**	0.349	0.348	**0.337**
ECL	TimeFilter	0.133	0.132	**0.129**	0.230	0.230	**0.228**

Comparison details. Affine calibration fits a channel-wise scale and offset on validation data. Direct residual regression uses the original OOF training pairs and reduced-rank ridge, matching MORA's predictive coefficient count to within approximately 2%. The δ-Adapter comparisons use the authors' released implementations (Liang et al., 2026) with the backbone frozen and online test-time updates disabled. Neural adapters use seed 2025; input–output candidates (additive budgets 0.01 and 0.1, or a multiplicative budget of 0.2) are selected by validation MSE.

F Complete Forecasting Results

Tables 12–15 report the horizon-level results for the main comparisons, with runs aggregated as described in Appendix D.4. Bold marks the lower displayed error within each baseline–MORA pair; ties at the displayed precision are unbolded. Shaded rows give Imp. (%).

F.1 Task-Specific Forecasters

Table 12: Complete long-horizon results for six task-specific backbones at $L = 96$.

Dataset	H	Forecast	TimeFilter		iTransformer		CGTFra		PatchTST		DLinear		VPNet	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	Base	0.369	0.394	0.386	0.404	0.373	**0.388**	0.379	0.399	0.396	0.411	0.375	0.387
		+MORA	**0.358**	**0.388**	**0.382**	**0.402**	**0.371**	0.389	**0.365**	**0.392**	**0.378**	**0.401**	**0.368**	**0.385**
		Imp. (%)	2.890	1.499	1.046	0.529	0.442	-0.234	3.652	1.700	4.486	2.376	1.945	0.647
	192	Base	0.412	0.421	0.442	0.436	0.429	0.422	0.425	0.431	0.447	0.442	0.428	0.419
		+MORA	**0.403**	**0.415**	**0.431**	**0.428**	**0.424**	**0.418**	**0.411**	**0.422**	**0.424**	**0.430**	**0.420**	**0.415**
		Imp. (%)	2.304	1.442	2.565	1.966	1.012	0.815	3.418	2.178	5.117	2.840	1.843	0.940
	336	Base	0.447	0.439	0.491	0.461	0.470	0.442	0.471	0.458	0.491	0.470	0.466	0.441
		+MORA	**0.438**	**0.435**	**0.484**	**0.455**	0.470	0.442	**0.452**	**0.448**	**0.459**	**0.452**	**0.462**	**0.438**
		Imp. (%)	2.022	0.791	1.462	1.178	0.098	0.078	4.167	2.282	6.483	3.860	0.960	0.535
	720	Base	0.445	0.456	0.499	0.487	**0.470**	**0.463**	0.518	0.505	0.515	0.513	**0.476**	**0.468**
		+MORA	**0.441**	**0.454**	**0.494**	**0.482**	0.485	0.471	**0.482**	**0.482**	**0.494**	**0.499**	0.483	0.471
		Imp. (%)	0.900	0.280	0.984	1.078	-3.398	-1.857	6.912	4.532	4.052	2.584	-1.427	-0.731
ETTh2	96	Base	0.288	0.340	0.301	0.350	0.296	0.339	0.302	0.353	0.349	0.401	0.290	0.335
		+MORA	**0.283**	**0.336**	**0.292**	**0.344**	**0.288**	**0.335**	**0.285**	**0.338**	**0.300**	**0.355**	**0.288**	**0.334**
		Imp. (%)	1.768	1.024	3.091	1.843	2.538	1.269	5.748	4.233	13.991	11.526	0.716	0.187
	192	Base	0.381	0.406	0.380	0.399	0.368	0.386	0.383	0.403	0.475	0.475	0.358	0.377
		+MORA	**0.369**	**0.393**	**0.369**	**0.391**	**0.362**	**0.381**	**0.364**	**0.388**	**0.404**	**0.421**	**0.356**	**0.376**
		Imp. (%)	3.060	3.108	2.847	1.950	1.755	1.193	4.942	3.803	15.018	11.344	0.592	0.175
	336	Base	0.408	0.431	0.423	0.432	0.418	0.425	0.429	0.438	0.599	0.544	**0.397**	**0.409**
		+MORA	**0.403**	**0.421**	**0.413**	**0.424**	**0.413**	**0.421**	**0.410**	**0.424**	**0.541**	**0.507**	0.409	0.417
		Imp. (%)	1.323	2.224	2.438	1.891	1.130	0.777	4.487	3.206	9.598	6.871	-3.032	-1.933
	720	Base	**0.410**	**0.436**	0.429	0.447	0.422	0.439	0.444	0.460	0.825	0.655	**0.393**	**0.419**
		+MORA	0.424	0.446	**0.415**	**0.435**	**0.415**	**0.434**	**0.410**	**0.438**	**0.701**	**0.583**	0.400	0.430
		Imp. (%)	-3.559	-2.469	3.249	2.558	1.751	1.088	7.678	4.741	15.004	10.999	-1.824	-2.483
ETTm1	96	Base	0.313	0.354	0.341	0.375	0.321	0.349	0.323	0.364	0.344	0.373	0.312	0.341
		+MORA	**0.304**	**0.350**	**0.313**	**0.357**	**0.296**	**0.340**	**0.307**	**0.353**	**0.310**	**0.351**	**0.297**	**0.340**
		Imp. (%)	3.056	1.275	8.057	4.800	7.705	2.537	5.183	3.112	10.051	5.962	4.663	0.352
	192	Base	0.355	0.380	0.380	0.394	0.369	0.373	0.369	0.388	0.380	0.391	0.362	0.372
		+MORA	**0.338**	**0.373**	**0.350**	**0.380**	**0.346**	**0.367**	**0.346**	**0.379**	**0.343**	**0.376**	**0.340**	**0.367**
		Imp. (%)	4.919	1.684	7.791	3.549	6.327	1.601	6.190	2.504	9.807	3.929	6.013	1.229
	336	Base	0.386	0.403	0.420	0.419	0.397	0.394	0.399	0.409	0.414	0.415	0.386	0.390
		+MORA	**0.365**	**0.393**	**0.379**	**0.403**	**0.372**	**0.390**	**0.369**	**0.399**	**0.368**	**0.394**	**0.363**	**0.385**
		Imp. (%)	5.417	2.597	9.639	3.855	6.318	1.128	7.586	2.505	11.071	4.986	5.916	1.117
	720	Base	0.452	0.437	0.486	0.456	0.485	0.442	0.458	0.445	0.475	0.453	0.445	0.425
		+MORA	**0.422**	**0.426**	**0.432**	**0.434**	**0.428**	**0.422**	**0.426**	**0.429**	**0.427**	**0.426**	**0.418**	**0.418**
		Imp. (%)	6.667	2.628	11.114	4.873	11.800	4.609	7.127	3.567	10.072	5.898	5.968	1.611
ETTm2	96	Base	0.169	0.255	0.185	0.272	0.172	0.249	0.183	0.268	0.195	0.295	0.173	0.247
		+MORA	**0.165**	**0.252**	**0.169**	**0.254**	**0.165**	**0.247**	**0.171**	**0.258**	**0.175**	**0.265**	**0.167**	**0.244**
		Imp. (%)	1.887	1.126	8.865	6.487	4.080	1.056	6.539	3.623	10.288	10.081	3.457	0.988
	192	Base	0.235	0.298	0.253	0.311	0.239	0.294	0.247	0.309	0.285	0.363	0.236	0.291
		+MORA	**0.229**	**0.296**	**0.233**	**0.300**	**0.229**	**0.290**	**0.233**	**0.302**	**0.251**	**0.323**	**0.229**	**0.288**
		Imp. (%)	2.355	0.971	7.956	3.390	4.505	1.269	5.720	2.290	11.955	11.012	3.033	1.022
	336	Base	0.293	0.336	0.316	0.353	0.304	0.337	0.318	0.352	0.379	0.424	0.301	0.333
		+MORA	**0.286**	**0.333**	**0.293**	**0.341**	**0.289**	**0.330**	**0.299**	**0.345**	**0.336**	**0.380**	**0.290**	**0.329**
		Imp. (%)	2.205	0.783	7.219	3.481	5.011	2.020	5.806	1.904	11.226	10.345	3.688	1.153
	720	Base	0.390	0.393	0.406	0.403	0.397	0.393	0.426	0.414	0.548	0.520	0.387	0.388
		+MORA	**0.376**	**0.388**	**0.384**	**0.392**	**0.377**	**0.385**	**0.397**	**0.402**	**0.451**	**0.449**	**0.373**	**0.382**
		Imp. (%)	3.454	1.318	5.592	2.757	4.915	2.166	6.805	2.896	17.850	13.590	3.440	1.525

Table 12 (continued, part 2 of 2). Complete long-horizon results for six task-specific backbones at $L = 96$.

Dataset	H	Forecast	TimeFilter		iTransformer		CGTFra		PatchTST		DLinear		VPNet	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	Base	0.153	0.199	0.174	0.213	0.153	**0.190**	0.176	0.217	0.195	0.255	0.154	0.192
		+MORA	**0.148**	**0.198**	**0.151**	**0.210**	**0.144**	0.191	**0.151**	**0.212**	**0.151**	**0.219**	**0.149**	**0.191**
		Imp. (%)	3.505	0.523	13.153	1.671	5.384	-0.091	13.943	2.322	22.545	13.879	3.417	0.872
	192	Base	0.202	0.246	0.223	0.255	0.205	0.240	0.222	0.257	0.240	0.299	0.202	0.240
		+MORA	**0.194**	**0.243**	**0.194**	**0.251**	**0.190**	**0.238**	**0.194**	**0.253**	**0.204**	**0.272**	**0.195**	**0.236**
		Imp. (%)	4.144	1.450	12.943	1.735	7.612	1.003	12.354	1.580	14.817	9.169	3.391	1.568
	336	Base	0.259	0.289	0.281	0.298	0.260	0.282	0.280	0.299	0.284	0.335	0.259	0.281
		+MORA	**0.247**	**0.285**	**0.247**	**0.293**	**0.241**	**0.275**	**0.247**	**0.295**	**0.258**	**0.313**	**0.247**	**0.278**
		Imp. (%)	4.839	1.462	12.270	1.779	7.612	2.394	11.930	1.201	9.358	6.615	4.557	0.999
	720	Base	0.345	0.344	0.358	0.350	0.342	0.336	0.354	0.348	0.348	0.385	0.337	0.336
		+MORA	**0.319**	**0.335**	**0.331**	**0.345**	**0.315**	**0.326**	**0.330**	**0.347**	**0.330**	**0.374**	**0.322**	**0.332**
		Imp. (%)	7.736	2.708	7.663	1.341	7.888	2.815	6.860	0.141	5.220	2.983	4.263	1.257
ECL	96	Base	0.135	0.233	0.148	0.240	0.139	0.228	0.180	0.273	0.195	0.277	0.135	0.224
		+MORA	**0.130**	**0.229**	**0.136**	**0.234**	**0.132**	**0.226**	**0.150**	**0.253**	**0.156**	**0.258**	**0.130**	**0.223**
		Imp. (%)	3.618	1.409	8.454	2.368	4.623	1.083	16.788	7.136	19.679	6.946	3.873	0.783
	192	Base	0.158	0.253	0.166	0.257	0.155	0.243	0.188	0.280	0.194	0.280	0.152	0.240
		+MORA	**0.150**	**0.247**	**0.152**	**0.249**	**0.148**	**0.240**	**0.160**	**0.261**	**0.162**	**0.262**	**0.146**	**0.238**
		Imp. (%)	5.580	2.411	8.171	2.842	4.441	1.107	14.528	6.697	16.359	6.497	4.164	1.025
	336	Base	0.166	0.264	0.179	0.271	0.170	0.260	0.204	0.296	0.207	0.296	0.169	0.258
		+MORA	**0.158**	**0.259**	**0.164**	**0.263**	**0.162**	**0.256**	**0.175**	**0.276**	**0.177**	**0.278**	**0.160**	**0.254**
		Imp. (%)	4.588	1.933	8.115	2.967	4.841	1.394	14.147	6.584	14.562	6.061	5.388	1.466
	720	Base	0.196	0.294	0.209	0.299	0.204	0.288	0.245	0.328	0.242	0.329	0.200	0.287
		+MORA	**0.186**	**0.288**	**0.190**	**0.288**	**0.194**	**0.286**	**0.212**	**0.307**	**0.214**	**0.312**	**0.189**	**0.283**
		Imp. (%)	5.255	1.973	9.358	3.639	4.633	0.579	13.510	6.194	11.696	5.201	5.362	1.285
Traffic	96	Base	0.375	0.251	0.392	0.267	0.390	0.240	0.460	0.297	0.650	0.397	0.384	0.257
		+MORA	**0.368**	**0.249**	**0.380**	**0.262**	**0.384**	**0.239**	**0.432**	**0.285**	**0.530**	**0.340**	**0.376**	**0.255**
		Imp. (%)	1.890	1.050	2.931	2.075	1.496	0.255	5.969	4.234	18.453	14.303	2.078	0.980
	192	Base	0.396	0.262	0.415	0.278	0.422	0.249	0.467	0.301	0.599	0.371	0.408	0.267
		+MORA	**0.387**	**0.259**	**0.401**	**0.271**	**0.415**	**0.248**	**0.439**	**0.288**	**0.501**	**0.324**	**0.401**	**0.264**
		Imp. (%)	2.224	1.346	3.356	2.517	1.666	0.207	6.001	4.410	16.338	12.698	1.879	1.107
	336	Base	0.415	0.271	0.427	0.284	0.438	0.261	0.482	0.307	0.606	0.374	0.427	0.275
		+MORA	**0.406**	**0.268**	**0.415**	**0.278**	**0.430**	0.261	**0.457**	**0.296**	**0.522**	**0.336**	**0.419**	**0.273**
		Imp. (%)	2.080	1.194	2.897	1.949	1.841	0.104	5.302	3.463	13.813	10.232	2.046	0.879
	720	Base	0.447	0.291	0.457	0.300	0.471	**0.277**	0.516	0.325	0.646	0.396	0.465	0.296
		+MORA	**0.440**	**0.289**	**0.448**	**0.297**	**0.465**	0.278	**0.498**	**0.317**	**0.594**	**0.373**	**0.458**	**0.294**
		Imp. (%)	1.569	0.653	1.969	0.890	1.398	-0.190	3.494	2.416	8.006	5.631	1.587	0.599
Solar	96	Base	0.197	0.232	0.203	0.237	0.192	**0.207**	0.223	0.278	0.290	0.377	0.176	**0.190**
		+MORA	**0.171**	**0.225**	**0.171**	**0.228**	**0.175**	0.209	**0.174**	**0.248**	**0.177**	**0.253**	**0.170**	0.192
		Imp. (%)	13.130	2.871	15.631	3.670	8.900	-0.762	22.035	10.796	38.906	32.980	3.163	-0.814
	192	Base	0.228	0.257	0.238	0.263	0.221	**0.227**	0.260	0.303	0.320	0.398	0.195	**0.206**
		+MORA	**0.201**	**0.251**	**0.194**	**0.249**	**0.204**	0.228	**0.198**	**0.266**	**0.200**	**0.271**	**0.189**	0.207
		Imp. (%)	12.056	2.465	18.301	5.446	7.695	-0.314	23.851	12.277	37.287	31.772	3.342	-0.476
	336	Base	0.239	0.269	0.250	0.274	0.241	0.242	0.277	0.308	0.352	0.415	0.217	0.221
		+MORA	**0.204**	**0.258**	**0.207**	**0.259**	**0.221**	**0.239**	**0.217**	**0.279**	**0.221**	**0.281**	**0.211**	0.221
		Imp. (%)	14.671	4.224	17.291	5.522	8.019	1.464	21.616	9.454	37.231	32.337	3.018	0.071
	720	Base	0.244	0.274	0.250	0.276	0.248	0.245	0.274	0.309	0.357	0.413	0.229	**0.227**
		+MORA	**0.215**	**0.267**	**0.220**	**0.269**	**0.239**	**0.244**	**0.232**	**0.285**	**0.247**	**0.306**	**0.227**	0.228
		Imp. (%)	11.808	2.572	12.184	2.467	3.622	0.441	15.471	7.764	30.627	25.852	0.986	-0.326

F.2 FOUNDATION MODELS AT $L = 96$

Table 13: Complete long-horizon results for frozen foundation models at $L = 96$.

Dataset	H	Forecast	Chronos-2		Sundial		TimesFM 2.5	
			MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	Base	0.424	0.397	0.406	0.396	0.429	0.402
		+MORA	**0.394**	**0.393**	**0.376**	**0.392**	**0.379**	**0.394**
		Imp. (%)	6.967	1.111	7.240	1.176	11.662	2.190
	192	Base	0.484	0.431	0.469	0.430	0.485	0.432
		+MORA	**0.441**	**0.419**	**0.423**	**0.419**	**0.424**	**0.420**
		Imp. (%)	8.880	2.795	9.854	2.636	12.682	2.680
	336	Base	0.536	0.456	0.516	0.451	0.526	0.450
		+MORA	**0.492**	**0.444**	**0.479**	**0.442**	**0.466**	**0.441**
		Imp. (%)	8.227	2.607	7.068	1.884	11.358	2.106
	720	Base	0.526	0.471	0.500	0.461	0.498	**0.457**
		+MORA	**0.497**	**0.467**	**0.466**	**0.457**	**0.472**	0.462
		Imp. (%)	5.431	0.988	6.852	0.758	5.130	-1.050
ETTh2	96	Base	0.317	0.344	0.317	0.353	0.330	0.345
		+MORA	**0.295**	**0.331**	**0.288**	**0.337**	**0.298**	**0.335**
		Imp. (%)	7.021	3.977	9.078	4.622	9.758	2.901
	192	Base	0.413	0.400	0.405	0.405	0.420	0.402
		+MORA	**0.379**	**0.382**	**0.375**	**0.388**	**0.396**	**0.391**
		Imp. (%)	8.174	4.417	7.521	4.031	5.806	2.620
	336	Base	0.449	0.433	0.438	0.436	0.453	0.434
		+MORA	**0.419**	**0.417**	**0.418**	**0.424**	**0.434**	**0.425**
		Imp. (%)	6.601	3.812	4.479	2.620	4.285	2.192
	720	Base	0.443	0.442	**0.437**	**0.445**	**0.450**	**0.443**
		+MORA	**0.441**	**0.441**	0.474	0.472	0.462	0.465
		Imp. (%)	0.487	0.303	-8.556	-5.959	-2.575	-4.912
ETTm1	96	Base	0.592	0.452	0.478	0.420	0.841	0.560
		+MORA	**0.312**	**0.356**	**0.317**	**0.356**	**0.311**	**0.354**
		Imp. (%)	47.221	21.329	33.785	15.219	63.051	36.912
	192	Base	0.607	0.467	0.527	0.449	0.881	0.582
		+MORA	**0.352**	**0.379**	**0.353**	**0.379**	**0.358**	**0.380**
		Imp. (%)	42.092	18.864	32.871	15.731	59.363	34.684
	336	Base	0.618	0.482	0.548	0.467	0.872	0.588
		+MORA	**0.376**	**0.399**	**0.378**	**0.398**	**0.381**	**0.397**
		Imp. (%)	39.086	17.133	31.103	14.681	56.328	32.427
	720	Base	0.646	0.508	0.596	0.494	0.905	0.607
		+MORA	**0.425**	**0.429**	**0.428**	**0.429**	**0.431**	**0.429**
		Imp. (%)	34.188	15.544	28.205	13.086	52.357	29.275
ETTm2	96	Base	0.210	0.273	0.201	0.276	0.257	0.312
		+MORA	**0.174**	**0.250**	**0.169**	**0.254**	**0.174**	**0.255**
		Imp. (%)	16.965	8.321	15.637	8.134	32.146	18.140
	192	Base	0.274	0.314	0.266	0.318	0.333	0.357
		+MORA	**0.236**	**0.291**	**0.233**	**0.299**	**0.238**	**0.299**
		Imp. (%)	13.757	7.339	12.434	6.183	28.510	16.374
	336	Base	0.340	0.355	0.326	0.355	0.395	0.393
		+MORA	**0.298**	**0.332**	**0.293**	**0.337**	**0.298**	**0.338**
		Imp. (%)	12.324	6.508	10.226	5.103	24.502	14.135
	720	Base	0.448	0.412	0.424	0.409	0.501	0.448
		+MORA	**0.393**	**0.386**	**0.389**	**0.390**	**0.397**	**0.394**
		Imp. (%)	12.262	6.111	8.263	4.723	20.878	12.128

Table 13 (continued, part 2 of 2). Complete long-horizon results for frozen foundation models at $L = 96$.

Dataset	H	Forecast	Chronos-2		Sundial		TimesFM 2.5	
			MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	Base	0.244	0.230	0.195	0.235	0.207	0.229
		+MORA	**0.171**	**0.224**	**0.151**	**0.218**	**0.152**	**0.210**
		Imp. (%)	29.872	2.529	22.542	7.029	26.717	8.175
	192	Base	0.273	0.270	0.254	0.283	0.264	0.277
		+MORA	**0.206**	**0.267**	**0.196**	**0.266**	**0.194**	**0.258**
		Imp. (%)	24.501	1.258	22.674	6.140	26.376	6.830
	336	Base	0.319	**0.305**	0.309	0.322	0.336	0.318
		+MORA	**0.256**	0.306	**0.251**	**0.309**	**0.248**	**0.302**
		Imp. (%)	19.551	-0.134	18.685	3.811	26.109	4.913
	720	Base	0.392	**0.357**	0.382	**0.369**	0.450	0.374
		+MORA	**0.339**	0.362	**0.331**	0.370	**0.327**	**0.360**
		Imp. (%)	13.635	-1.391	13.228	-0.046	27.335	3.647
ECL	96	Base	0.214	0.274	0.180	0.261	0.223	0.293
		+MORA	**0.159**	**0.253**	**0.149**	**0.247**	**0.164**	**0.264**
		Imp. (%)	25.886	7.622	17.197	5.210	26.459	9.869
	192	Base	0.213	0.279	0.187	0.269	0.220	0.292
		+MORA	**0.167**	**0.260**	**0.159**	**0.256**	**0.171**	**0.268**
		Imp. (%)	21.610	6.879	14.650	4.846	22.592	8.431
	336	Base	0.230	0.295	0.204	0.286	0.236	0.307
		+MORA	**0.184**	**0.275**	**0.174**	**0.272**	**0.186**	**0.283**
		Imp. (%)	20.155	6.719	14.334	4.920	20.969	7.820
	720	Base	0.279	0.329	0.245	0.319	0.278	0.341
		+MORA	**0.224**	**0.307**	**0.211**	**0.304**	**0.226**	**0.317**
		Imp. (%)	19.458	6.412	13.908	4.926	18.830	7.000
Traffic	96	Base	0.574	0.310	0.605	0.353	0.602	0.338
		+MORA	**0.497**	**0.295**	**0.513**	**0.317**	**0.530**	**0.311**
		Imp. (%)	13.280	5.018	15.189	10.148	11.970	8.010
	192	Base	0.550	0.301	0.582	0.344	0.575	0.324
		+MORA	**0.491**	**0.289**	**0.502**	**0.313**	**0.515**	**0.303**
		Imp. (%)	10.781	4.186	13.652	9.223	10.328	6.483
	336	Base	0.566	0.309	0.596	0.351	0.592	0.329
		+MORA	**0.513**	**0.298**	**0.524**	**0.324**	**0.538**	**0.313**
		Imp. (%)	9.313	3.519	12.173	7.749	9.023	5.011
	720	Base	0.606	0.331	0.635	0.372	0.628	0.346
		+MORA	**0.569**	**0.324**	**0.585**	**0.354**	**0.594**	**0.337**
		Imp. (%)	6.057	2.142	7.832	4.801	5.372	2.600
Solar	96	Base	0.640	0.343	0.625	0.502	0.875	0.503
		+MORA	**0.186**	**0.265**	**0.177**	**0.254**	**0.184**	**0.254**
		Imp. (%)	70.987	22.878	71.722	49.450	78.941	49.516
	192	Base	0.527	0.324	0.680	0.534	0.937	0.528
		+MORA	**0.208**	**0.285**	**0.201**	**0.272**	**0.211**	**0.274**
		Imp. (%)	60.569	12.318	70.495	49.014	77.473	48.213
	336	Base	0.494	0.323	0.684	0.537	0.909	0.520
		+MORA	**0.221**	**0.289**	**0.222**	**0.285**	**0.234**	**0.287**
		Imp. (%)	55.344	10.342	67.609	46.916	74.258	44.733
	720	Base	0.451	0.319	0.676	0.545	0.930	0.528
		+MORA	**0.246**	**0.307**	**0.243**	**0.301**	**0.258**	**0.303**
		Imp. (%)	45.493	3.786	64.109	44.763	72.301	42.712

F.3 FOUNDATION MODELS AT $L = 512$

Table 14: Complete long-horizon results for frozen foundation models at $L = 512$.

Dataset	H	Forecast	Chronos-2		Sundial		TimesFM 2.5	
			MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	Base	0.371	**0.375**	0.356	0.383	0.384	0.382
		+MORA	**0.350**	0.376	**0.343**	**0.378**	**0.354**	**0.380**
		Imp. (%)	5.652	-0.287	3.593	1.318	7.781	0.545
	192	Base	0.414	0.406	0.390	0.406	0.428	**0.408**
		+MORA	**0.392**	0.406	**0.378**	**0.401**	**0.401**	0.409
		Imp. (%)	5.260	-0.093	3.025	1.106	6.117	-0.190
	336	Base	0.439	**0.423**	0.403	0.415	0.446	**0.419**
		+MORA	**0.421**	0.427	**0.396**	**0.414**	**0.427**	0.424
		Imp. (%)	4.179	-0.906	1.857	0.266	4.340	-1.193
	720	Base	0.449	**0.445**	0.438	0.444	0.428	**0.427**
		+MORA	**0.445**	0.452	**0.431**	**0.443**	0.428	0.437
		Imp. (%)	0.925	-1.602	1.603	0.134	-0.148	-2.455
ETTh2	96	Base	0.280	0.320	0.279	0.334	0.306	0.338
		+MORA	**0.273**	**0.316**	**0.271**	**0.325**	**0.295**	**0.332**
		Imp. (%)	2.535	1.068	2.938	2.550	3.730	1.796
	192	Base	0.351	0.369	0.342	0.376	0.376	0.384
		+MORA	**0.342**	**0.364**	**0.338**	**0.371**	**0.365**	**0.379**
		Imp. (%)	2.717	1.316	1.211	1.468	2.858	1.444
	336	Base	0.377	0.397	**0.361**	**0.395**	0.401	0.409
		+MORA	**0.371**	**0.395**	0.367	0.397	**0.394**	**0.405**
		Imp. (%)	1.726	0.402	-1.778	-0.482	1.713	0.875
	720	Base	0.398	**0.424**	0.392	0.426	**0.405**	0.422
		+MORA	**0.393**	0.434	**0.384**	**0.420**	0.406	**0.418**
		Imp. (%)	1.362	-2.310	1.997	1.358	-0.195	0.756
ETTm1	96	Base	0.308	0.323	0.292	0.325	0.313	0.329
		+MORA	**0.277**	**0.315**	**0.275**	**0.323**	**0.284**	**0.326**
		Imp. (%)	10.024	2.463	5.920	0.676	9.169	1.008
	192	Base	0.362	0.355	0.342	0.357	0.361	0.359
		+MORA	**0.318**	**0.345**	**0.318**	**0.353**	**0.321**	**0.353**
		Imp. (%)	12.170	2.873	7.018	1.176	10.897	1.813
	336	Base	0.395	0.381	0.378	0.380	0.394	0.382
		+MORA	**0.345**	**0.370**	**0.350**	**0.375**	**0.351**	**0.374**
		Imp. (%)	12.803	2.734	7.557	1.434	10.820	2.128
	720	Base	0.484	0.430	0.451	0.418	0.456	0.418
		+MORA	**0.411**	**0.412**	**0.411**	**0.409**	**0.409**	**0.411**
		Imp. (%)	14.997	4.306	8.910	2.086	10.337	1.501
ETTm2	96	Base	0.170	0.237	0.176	0.255	0.179	0.246
		+MORA	**0.161**	**0.233**	**0.158**	**0.244**	**0.168**	**0.241**
		Imp. (%)	5.504	1.841	10.200	4.244	6.448	2.062
	192	Base	0.241	0.286	0.239	0.298	0.249	0.293
		+MORA	**0.224**	**0.277**	**0.216**	**0.285**	**0.231**	**0.285**
		Imp. (%)	6.835	2.823	9.776	4.402	7.006	2.958
	336	Base	0.306	0.329	0.297	0.336	0.314	0.336
		+MORA	**0.277**	**0.315**	**0.272**	**0.323**	**0.294**	**0.326**
		Imp. (%)	9.327	4.403	8.294	3.623	6.455	2.956
	720	Base	0.404	0.390	0.387	0.393	0.420	0.396
		+MORA	**0.366**	**0.373**	**0.364**	**0.381**	**0.399**	**0.387**
		Imp. (%)	9.249	4.478	6.138	3.056	5.089	2.445

Table 14 (continued, part 2 of 2). Complete long-horizon results for frozen foundation models at $L = 512$.

Dataset	H	Forecast	Chronos-2		Sundial		TimesFM 2.5	
			MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	Base	0.156	**0.176**	0.160	0.207	0.163	0.194
		+MORA	**0.139**	0.178	**0.139**	**0.195**	**0.142**	**0.186**
		Imp. (%)	10.563	-1.281	13.187	5.588	12.716	4.399
	192	Base	0.211	0.226	0.213	0.255	0.216	0.242
		+MORA	**0.181**	**0.225**	**0.182**	**0.240**	**0.184**	**0.227**
		Imp. (%)	14.208	0.505	14.873	6.002	14.831	6.405
	336	Base	0.281	0.275	0.267	0.295	0.289	0.289
		+MORA	**0.229**	**0.269**	**0.231**	**0.280**	**0.232**	**0.267**
		Imp. (%)	18.699	2.069	13.580	5.082	19.744	7.461
	720	Base	0.383	0.332	0.336	0.344	0.385	0.345
		+MORA	**0.302**	**0.329**	**0.300**	**0.332**	**0.302**	**0.318**
		Imp. (%)	21.168	0.937	10.945	3.359	21.631	7.811
ECL	96	Base	0.128	0.211	0.118	0.213	0.133	0.220
		+MORA	**0.123**	**0.210**	**0.116**	**0.212**	**0.127**	**0.218**
		Imp. (%)	3.483	0.394	1.657	0.588	4.320	0.810
	192	Base	0.146	0.229	0.135	0.231	0.150	0.236
		+MORA	**0.141**	**0.228**	**0.132**	**0.229**	**0.143**	**0.234**
		Imp. (%)	3.732	0.538	2.126	0.754	4.480	1.014
	336	Base	0.167	0.249	0.150	0.247	0.167	0.253
		+MORA	**0.158**	**0.247**	**0.146**	**0.245**	**0.159**	**0.251**
		Imp. (%)	5.013	0.973	2.865	0.979	5.096	1.169
	720	Base	0.210	0.287	0.188	0.280	0.206	0.287
		+MORA	**0.196**	**0.281**	**0.180**	**0.276**	**0.195**	**0.283**
		Imp. (%)	6.869	1.912	4.257	1.461	5.696	1.545
Traffic	96	Base	0.353	0.219	0.425	0.270	0.355	0.231
		+MORA	**0.349**	0.219	**0.408**	**0.265**	**0.352**	**0.230**
		Imp. (%)	0.965	-0.212	3.920	1.923	0.873	0.149
	192	Base	0.379	0.231	0.448	0.283	0.371	0.238
		+MORA	**0.375**	0.231	**0.428**	**0.276**	**0.368**	0.238
		Imp. (%)	1.106	-0.106	4.341	2.471	0.903	0.155
	336	Base	0.396	0.240	0.462	0.292	0.382	0.244
		+MORA	**0.392**	0.240	**0.444**	**0.285**	**0.379**	0.244
		Imp. (%)	1.006	-0.158	3.944	2.350	0.730	-0.022
	720	Base	0.442	**0.264**	0.499	0.314	0.418	0.264
		+MORA	**0.440**	0.265	**0.485**	**0.308**	**0.416**	0.264
		Imp. (%)	0.384	-0.205	2.837	1.799	0.519	0.052
Solar	96	Base	0.157	0.172	0.212	0.234	0.236	**0.221**
		+MORA	0.157	0.172	**0.190**	0.234	**0.186**	0.231
		Imp. (%)	0.011	-0.032	10.075	-0.266	20.996	-4.608
	192	Base	0.180	0.190	0.229	0.253	0.270	**0.242**
		+MORA	0.180	0.190	**0.209**	0.253	**0.217**	0.253
		Imp. (%)	0.069	-0.119	8.832	-0.024	19.555	-4.666
	336	Base	0.197	0.202	0.234	0.264	0.303	**0.262**
		+MORA	0.197	0.202	**0.214**	0.264	**0.241**	0.274
		Imp. (%)	0.148	-0.214	8.562	0.161	20.378	-4.620
	720	Base	0.214	**0.210**	0.243	0.277	0.317	**0.274**
		+MORA	**0.213**	0.211	**0.220**	**0.275**	**0.242**	0.285
		Imp. (%)	0.405	-0.384	9.443	0.722	23.618	-4.162

F.4 SHORT-HORIZON FORECASTING ON PEMS

Table 15: Complete short-horizon results on PEMS at $L = 96$.

Dataset	H	Forecast	TimeFilter		iTransformer		CGTFra		PatchTST		DLinear		VPNet	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS03	12	Base	0.063	0.164	**0.068**	0.174	0.060	0.160	0.084	0.191	0.107	0.219	0.066	0.171
		+MORA	**0.062**	0.164	0.069	**0.172**	**0.059**	**0.159**	**0.075**	**0.182**	**0.083**	**0.193**	0.066	0.171
		Imp. (%)	1.704	-0.059	-0.168	1.204	1.745	0.576	10.531	4.782	22.063	11.975	-0.578	-0.060
	24	Base	0.079	0.186	0.095	0.204	0.079	0.184	0.130	0.239	0.184	0.290	0.090	0.202
		+MORA	**0.077**	**0.184**	**0.088**	**0.192**	**0.075**	**0.180**	**0.098**	**0.208**	**0.109**	**0.219**	**0.084**	**0.194**
		Imp. (%)	3.062	1.023	7.299	5.700	4.641	2.126	24.083	12.862	40.671	24.421	6.581	4.129
	48	Base	0.109	0.222	**0.125**†	0.236†	0.119†	0.228†	0.231	0.324	0.303	0.389	0.125	0.239
		+MORA	**0.106**	**0.217**	0.126	**0.230**	**0.110**	**0.217**	**0.142**	**0.248**	**0.152**	**0.256**	**0.115**	**0.225**
		Imp. (%)	3.046	1.989	-1.121	2.505	7.322	4.706	38.688	23.650	49.885	34.103	8.367	5.605
	96	Base	0.154	0.269	0.171	0.281	0.178	0.281	0.379	0.427	0.387	0.468	0.185	0.297
		+MORA	**0.148**	**0.258**	0.171	**0.269**	**0.151**	**0.255**	**0.193**	**0.289**	**0.204**	**0.296**	**0.163**	**0.269**
		Imp. (%)	3.519	3.970	-0.013	4.148	15.085	9.410	49.023	32.428	47.286	36.723	12.004	9.496
PEMS04	12	Base	0.068	0.167	0.081	0.189	0.070	0.169	0.098	0.207	0.111	0.226	0.073	0.177
		+MORA	**0.064**	**0.161**	**0.072**	**0.173**	**0.066**	**0.163**	**0.073**	**0.174**	**0.076**	**0.179**	**0.066**	**0.165**
		Imp. (%)	5.406	3.684	11.134	8.489	5.853	3.903	25.557	15.758	31.532	20.921	9.127	6.719
	24	Base	0.080	0.183	0.099	0.210	0.084	0.187	0.156	0.264	0.182	0.298	0.095	0.210
		+MORA	**0.072**	**0.171**	**0.082**	**0.185**	**0.074**	**0.174**	**0.086**	**0.189**	**0.093**	**0.195**	**0.078**	**0.181**
		Imp. (%)	9.494	6.310	16.886	12.036	11.082	6.860	45.180	28.572	48.635	34.410	17.956	13.426
	48	Base	0.100	0.209	0.120†	0.233†	0.113	0.220	0.281	0.364	0.305	0.404	0.123	0.241
		+MORA	**0.084**	**0.187**	**0.096**	**0.200**	**0.087**	**0.190**	**0.105**	**0.205**	**0.118**	**0.210**	**0.090**	**0.195**
		Imp. (%)	16.173	10.641	20.076	13.950	22.971	13.611	62.411	43.517	61.151	48.033	27.454	19.035
	96	Base	0.134	0.247	0.150†	0.262†	0.160	0.266	0.490	0.497	0.387	0.471	0.151	0.268
		+MORA	**0.097**	**0.202**	**0.109**	**0.213**	**0.101**	**0.207**	**0.122**	**0.217**	**0.131**	**0.219**	**0.101**	**0.207**
		Imp. (%)	27.137	17.987	27.632	18.567	36.958	22.196	75.102	56.359	66.221	53.504	32.947	22.874
PEMS07	12	Base	0.055	0.150	**0.067**	0.165	0.056	0.147	0.079	0.184	0.100	0.220	0.064	0.167
		+MORA	**0.053**	**0.145**	0.068	**0.156**	**0.055**	**0.143**	**0.063**	**0.161**	**0.071**	**0.174**	**0.059**	**0.155**
		Imp. (%)	3.918	2.748	-2.294	5.382	1.271	2.575	19.735	12.430	29.351	21.181	8.192	7.121
	24	Base	0.068	0.166	0.087	0.190	0.075	0.167	0.133	0.240	0.186	0.305	0.092	0.204
		+MORA	**0.063**	**0.159**	**0.084**	**0.171**	**0.067**	**0.158**	**0.080**	**0.181**	**0.089**	**0.192**	**0.071**	**0.176**
		Imp. (%)	7.450	4.720	3.459	9.633	10.056	5.688	40.006	24.766	52.014	36.994	22.476	14.099
	48	Base	0.090	0.193	0.110	0.214	0.104	0.198	0.252	0.336	0.350	0.436	0.107	0.220
		+MORA	**0.076**	**0.175**	**0.098**	**0.187**	**0.085**	**0.175**	**0.096**	**0.195**	**0.107**	**0.203**	**0.080**	**0.182**
		Imp. (%)	15.306	9.405	10.549	13.014	18.848	11.952	62.024	41.922	69.534	53.554	24.571	16.933
	96	Base	0.122	0.229	0.141	0.247	0.148	0.243	0.416	0.444	0.517	0.564	0.139	0.252
		+MORA	**0.091**	**0.188**	**0.113**	**0.199**	**0.106**	**0.189**	**0.107**	**0.202**	**0.127**	**0.220**	**0.093**	**0.191**
		Imp. (%)	25.250	17.545	19.493	19.324	28.425	22.191	74.292	54.527	75.436	60.985	33.118	24.199
PEMS08	12	Base	0.064	0.162	0.081	0.184	0.071	0.168	0.084	0.191	0.102	0.218	0.067	0.174
		+MORA	**0.060**	**0.157**	**0.070**	**0.168**	**0.067**	**0.161**	**0.069**	**0.170**	**0.074**	**0.181**	**0.062**	**0.163**
		Imp. (%)	5.294	3.570	13.608	8.618	6.007	4.139	18.501	10.922	27.135	17.211	7.335	6.417
	24	Base	0.079	0.182	0.118	0.221	0.096	0.194	0.131	0.239	0.169	0.287	0.086	0.193
		+MORA	**0.073**	**0.173**	**0.091**	**0.191**	**0.086**	**0.182**	**0.088**	**0.194**	**0.096**	**0.207**	**0.076**	**0.179**
		Imp. (%)	7.611	4.865	22.635	13.575	9.965	6.118	32.501	18.715	43.043	27.909	10.998	7.132
	48	Base	0.104	0.212	0.186†	0.235†	0.153	0.244	0.235	0.326	0.295	0.398	0.113	0.224
		+MORA	**0.091**	**0.196**	**0.134**	**0.227**	**0.121**	**0.215**	**0.118**	**0.226**	**0.129**	**0.239**	**0.096**	**0.203**
		Imp. (%)	12.002	7.332	28.131	3.239	20.984	11.769	49.941	30.529	56.183	40.027	15.225	9.331
	96	Base	0.145	0.257	0.221†	0.267†	0.259	0.296	0.392	0.429	0.390	0.479	0.159	0.270
		+MORA	**0.115**	**0.224**	0.221	**0.254**	**0.167**	**0.244**	**0.144**	**0.247**	**0.154**	**0.257**	**0.119**	**0.228**
		Imp. (%)	20.679	12.789	-0.180	4.708	35.493	17.739	63.130	42.419	60.455	46.441	25.044	15.562

G Additional Residual Analysis

We extend Figure 1 to the eight long-term datasets using frozen DLinear, iTransformer, and TimeFilter forecasts at $L = H = 96$. The analysis retains the original chronological splits and training-set normalization. DLinear and TimeFilter use seed 2021. iTransformer uses seed 2021 on ETTh1, ETTm1, and Weather, and seed 2023 on the other five datasets. The linear probes are used only for this analysis; MORA is fitted as described in Appendix C.5.

G.1 Frequency Decomposition of Forecast Residuals

For test window i and channel c, let $\mathbf{r}_{i,c} = \mathbf{y}_{i,c} - \widehat{\mathbf{y}}^{(0)}_{i,c}$ be the backbone residual. Let $\mathbf{D}_H$ denote the orthonormal DCT-II matrix over the forecast horizon. With N test windows, the contribution of frequency index k to the overall forecasting MSE is

$$E_k = \frac{1}{NHC} \sum_{i=1}^{N} \sum_{c=1}^{C} \left[(\mathbf{D}_H \mathbf{r}_{i,c})_k \right]^2, \qquad k = 0, \ldots, H-1. \tag{43}$$

Orthonormality gives

$$\mathrm{MSE} = \sum_{k=0}^{H-1} E_k, \qquad F_8 = \frac{\sum_{k=0}^{7} E_k}{\sum_{k=0}^{H-1} E_k}. \tag{44}$$

Thus, F_8 measures the share of residual energy in the first eight modes, including the constant component. Display ranks 1–8 in Figure 1(a) correspond to DCT indices 0–7. The reported fractions use the full residual spectrum and all test windows and channels; they are ratios of summed energies, not averages of per-window energy fractions.

Table 16 reports F_8 alongside the history-based probe statistic defined below. On Weather, the energy shares are 84.0%, 81.8%, and 79.5% for DLinear, iTransformer, and TimeFilter, respectively, matching Figure 1(a). The first eight modes also account for substantial error on the other datasets, although the fraction varies: it exceeds 92% on Solar but is around 29%–33% on Traffic. The diagnostic band is fixed and is distinct from MORA's tapered correction subspace.

G.2 History-Based Forecast-Mean Probes

Following Appendix B.2, the probe targets the signed future-mean residual:

$$\bar{r}_{i,c} = \frac{1}{H} \mathbf{1}_H^\top \mathbf{r}_{i,c}, \qquad \mathbf{h}_{i,c} = \left[(\mathbf{D}_L \mathbf{x}_{i,c})_0, \ldots, (\mathbf{D}_L \mathbf{x}_{i,c})_7 \right]^\top. \tag{45}$$

Here $\mathbf{h}_{i,c}$ contains eight coefficients of the complete same-channel historical window, not eight selected time points. A positive $\bar{r}_{i,c}$ indicates that the backbone underpredicts the future mean; a negative value indicates overprediction.

For each backbone and channel, we fit a linear ridge probe with an intercept on that backbone's validation windows. Let $\widetilde{\mathbf{h}}_{i,c}$ denote the features standardized by their validation means and standard deviations, with a scale floor of 10^{-8}. The probe predicts $p_{i,c} = a_c + \mathbf{w}_c^\top \widetilde{\mathbf{h}}_{i,c}$ by minimizing

$$\frac{1}{N_{\mathrm{val}}} \sum_{i \in \mathcal{I}_{\mathrm{val}}} (\bar{r}_{i,c} - p_{i,c})^2 + 0.01 \|\mathbf{w}_c\|_2^2, \tag{46}$$

where $\mathcal{I}_{\mathrm{val}}$ contains the N_{val} validation windows. Predictions remain in the benchmark's standardized units. The probe uses neither future observations nor calendar covariates; its weights, normalization statistics, and group thresholds remain fixed on test data. Test targets are used only to evaluate the associations.

For Figure 1(b), each channel's validation probe scores define five groups through their 20th, 40th, 60th, and 80th percentiles. We apply these same thresholds to test scores computed from history. Within each group, we first average the estimated and observed mean errors over test windows separately for each channel, then average channels equally. To keep the channel composition the same across the five groups, this visualization uses channels with at least one test window in every

group. The resulting coverage on Weather is 19 of 21 channels for DLinear and iTransformer and 18 of 21 for TimeFilter. Coverage depends on the history-based group assignments, not on the observed errors. The plotted points are group means, not individual forecasts or a temporal trajectory.

For the cross-dataset summary, we compute the Pearson correlation between the frozen probe output and $\bar{r}_{i,c}$ over test windows within each channel, and report its median over channels with a defined correlation. This uses all such channels, independently of the five-group visualization's coverage criterion. Correlations are defined for 320 of 321 channels on ECL and for all channels in the other settings reported here.

Table 16: Residual diagnostics at $L = H = 96$. F_8 is the DCT-8 share of squared residual energy; r_{med} is the median channel-wise test correlation of the history-based mean-residual probe.

Dataset	**DLinear**		**iTransformer**		**TimeFilter**	
	F_8 (%)	r_{med}	F_8 (%)	r_{med}	F_8 (%)	r_{med}
ETTh1	48.195	0.115	50.060	0.264	50.558	0.219
ETTh2	81.734	0.161	77.132	0.068	78.749	0.277
ETTm1	88.041	0.008	86.825	0.128	87.196	0.122
ETTm2	84.858	0.206	82.751	0.318	82.226	0.210
Weather	83.984	0.471	81.847	0.302	79.537	0.206
ECL	61.541	0.176	57.438	0.228	55.567	0.225
Traffic	32.993	0.393	29.319	0.285	29.598	0.240
Solar	94.530	0.057	92.786	0.158	92.490	0.101

Interpretation. Spectral concentration and predictability answer different questions. DLinear has substantial low-frequency residual energy on both Weather and ETTm1, yet the corresponding probe correlations are 0.471 and 0.008. Low-frequency energy alone therefore does not establish that an error can be inferred from history. The probe provides separate evidence of history-dependent variation in the future-mean residual, without implying that every low-frequency component is predictable or that all backbones share a common training failure. Since test windows overlap, we report correlations and group means as descriptive statistics, not independent-sample significance tests.