SAR-LCRT: RISK-GATED RESIDUAL TRANSPORT FOR FROZEN TIME-SERIES FORECASTERS UNDER DELAYED FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate time-series forecasts support planning in energy, transportation, and environmental systems. Yet a trained model can repeatedly overestimate or underestimate future values, and retraining it after deployment may be costly. Correcting these errors presents two challenges: error patterns vary with the input, and the usefulness of a correction can be assessed only after the predicted time interval has elapsed. We propose **SAR-LCRT**, a framework that improves a trained forecaster without changing its parameters. Low-Rank Conditional Residual Transport (LCRT) learns patterns in prediction errors across future steps and variables, using validation data to construct input-dependent corrections. Structure-Adaptive Risk Control (SAR) then chooses whether to use a correction or retain the original forecast, based on observed outcomes of past forecasts. The framework compares all available corrections to guide selection. It separately tracks the gains and losses of the forecasts actually used and stops correction when the accumulated evidence no longer supports it. Across eight architectures and 13 datasets, SAR-LCRT matches or improves the original models' MSE on 86.1% of 416 forecasting tasks, based on three-seed means. On PEMS08, it reduces MSE by 11.06% on average across architectures and forecast horizons. These results demonstrate the value of combining structured error correction with delayed-feedback selection as a reusable addition to trained forecasters.

1 INTRODUCTION

Time-series forecasting uses past observations to predict future values, supporting decisions such as electricity scheduling and traffic management. Recent progress has largely come from better forecasting architectures (Nie et al., 2023; Liu et al., 2024) and training objectives (Wang et al., 2025b; 2026a; Shi et al., 2025; Qiu et al., 2025; Kudrat et al., 2025). Once a model is deployed, however, replacing or retraining it may be costly. This raises a practical question: *can we improve its predictions while keeping the trained model unchanged?*

A promising starting point is the residual: the difference between an observed value and its prediction. Repeated errors can contain useful patterns rather than only random noise. For example, a model may underestimate a rising series or make related errors across several variables. Learning these patterns offers a way to adjust its output without rebuilding the forecasting model. Existing test-time and online methods also address changing conditions through adaptation (Kim et al., 2025; Lau et al., 2025); here we focus on a correction layer whose forecasting parameters remain fixed after deployment.

Two challenges make this approach more than a constant bias adjustment. First, a useful correction must depend on the current input and capture error patterns shared across future steps and variables. Second, a correction that worked previously may become unsuitable as conditions change. Its full outcome is also delayed: a prediction for the next several time steps cannot be fully evaluated until those steps have passed. The system must therefore decide both *how to correct* a forecast and *when to use* that correction from the evidence available at the time.

We propose SAR-LCRT to address these challenges through separate correction and selection components (Figure 1). Low-Rank Conditional Residual Transport (LCRT) learns a compact represen-

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.

tation of shared error patterns and uses it to construct input-dependent corrections. Rather than treating every future step and variable separately, it represents their common error structure through a small number of shared patterns. Once learned, these patterns produce corrections for new inputs without repeatedly fitting new models. The original model and correction maps are fixed before testing. Structure-Adaptive Risk Control (SAR) chooses among their forecasts using completed past outcomes. It compares all available candidates, but tracks accumulated gains and losses only for the forecasts actually used. This distinction allows the system to learn which correction is useful while deciding when to return to the original model. Here risk control means using past errors to govern continued correction.

Our main contributions are:

1. A low-rank residual correction method that captures shared error structure across forecast steps and variables without retraining the backbone.
2. A delayed-feedback selection policy that separates candidate comparison from the loss accounting of issued forecasts and retains the original model as a fallback.
3. An evaluation across eight architectures and 13 datasets: SAR-LCRT matches or improves Base MSE on 358 of 416 tasks using three-seed means. On PEMS08, the average MSE reduction is 11.06% across architectures and horizons, with the largest task-level reduction of 24.57% achieved by MDMLP-EIA at horizon 48.

2 Related Work

Forecasting architectures and objectives. Forecasting models have been built in many ways: residual bases and hierarchical interpolation (Oreshkin et al., 2020; Challu et al., 2023), attention and decomposition (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022b), interactions across variables and time scales (Zhang & Yan, 2023; Chen et al., 2024), and frequency information or Koopman dynamics (Zhou et al., 2022a; Liu et al., 2023). Other designs rely on patches, convolutions, feature mixing, or lightweight networks (Nie et al., 2023; Luo & Wang, 2024; Wu et al., 2023; Wang et al., 2024; Zeng et al., 2023; Liu et al., 2024; Chen et al., 2025; An et al., 2026; Liu et al., 2026; Chen et al., 2026; Wang et al., 2026b). RevIN and Non-stationary Transformers handle distribution shifts (Kim et al., 2022; Liu et al., 2022), and TimesFM, Moirai, and Chronos pretrain on large time-series collections before transferring to new tasks (Das et al., 2024; Woo et al., 2024; Ansari et al., 2024). Training objectives have also been improved with label transformations, per-step weights, frequency-domain errors, and structural losses (Wang et al., 2025b;a; 2026a; Shi et al., 2025; Kudrat et al., 2025; Qiu et al., 2025). All of these improve the model itself or how it is trained. SAR-LCRT is different: it adds corrections to an already-trained model and decides when to use them, so it can be combined with any forecasting architecture.

Test-time and online forecasting. Some methods keep updating the forecaster itself online, such as FSNet and OneNet (Pham et al., 2023; Zhang et al., 2023), while TAFAS and DSOF update a small extra module as data changes, even when feedback is delayed (Kim et al., 2025; Lau et al., 2025). COSA keeps the forecaster frozen but updates a small output adapter online from the observed targets (Im & Kwon, 2026), and ADAPT-Z adjusts the input features online under delayed feedback (Huang et al., 2026). In contrast, SAR-LCRT fixes the backbone and every residual map before testing, and changes only which forecast it uses, once a full target has arrived. So it never learns parameters, features, or maps online; it only makes a discrete choice online. We compare against simple residual baselines (mean, EWMA, and ridge) under the same frozen checkpoints and complete-target feedback, so the differences come from the correction and selection design, not from the data.

Delayed decisions and sequential calibration. Delayed-feedback online learning studies decisions made before earlier losses arrive (Joulani et al., 2013), and conservative bandits bound cumulative loss against a baseline (Wu et al., 2016). SAR faces a similar decision problem, but it observes every candidate's loss once the target is complete. It tracks the losses of the forecasts actually selected during each correction episode, without a regret or uniform baseline guarantee. Its policy combines frozen residual candidates, direct candidate comparisons, and exits based on the forecasts it actually issued. Adaptive conformal inference, EnbPI, and KOWCPI study prediction-interval

coverage (Gibbs & Candès, 2021; Xu & Xie, 2021; Lee et al., 2025). SAR instead selects a point forecast; its HAC bounds guide candidate comparisons and do not guarantee prediction-interval coverage.

3 METHODOLOGY

SAR-LCRT fits two corrections to a frozen backbone and selects among Base and corrected forecasts online (Figure 1). The two corrections are Pure LCRT, which always applies one learned adjustment, and Adaptive LCRT, which can choose among several adjustment forms. After full targets arrive, candidate losses guide selection and issued-forecast losses control withdrawal, so the system can keep a correction while it is helping and fall back to Base when it is not. Fitted parameters stay fixed during testing.

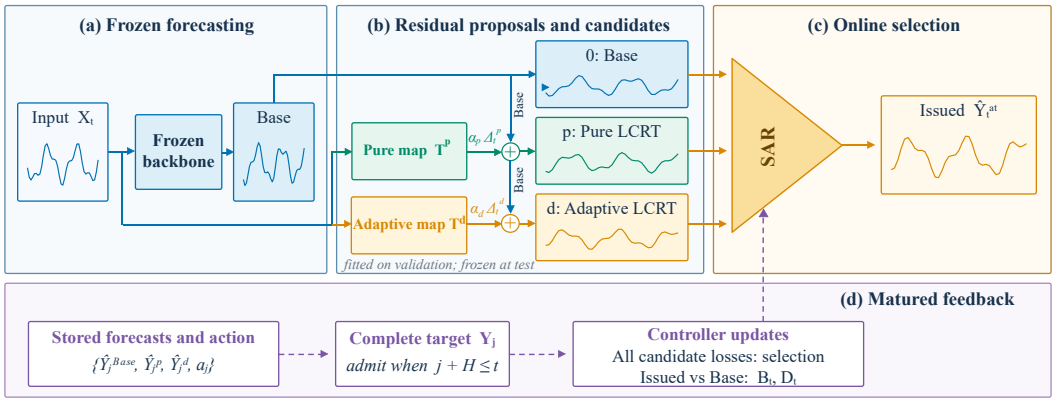


Figure 1: **Overview of SAR-LCRT. (a)** The frozen backbone produces Base. **(b)** Validation-fitted T^p, T^d produce the Pure LCRT and Adaptive LCRT candidates. **(c)** SAR outputs one candidate unchanged. **(d)** Once a full target arrives, all candidate losses guide selection; only issued-forecast losses update B_t and D_t for withdrawal. Solid arrows show current forecasts; dashed arrows show delayed feedback. The backbone and maps stay fixed at test; validation initializes a_0.

3.1 PROBLEM FORMULATION

At origin t, $\boldsymbol{X}_t = [\boldsymbol{x}_{t-L+1}, \ldots, \boldsymbol{x}_t] \in \mathbb{R}^{L\times C}$ contains L observations of C variables; $\boldsymbol{Y}_t = [\boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_{t+H}] \in \mathbb{R}^{H\times C}$ is the H-step target. The backbone parameters $\boldsymbol{\theta}$ are fixed:

$$\widehat{\boldsymbol{Y}}_t^{\mathrm{Base}} = f_{\boldsymbol{\theta}}(\boldsymbol{X}_t) \in \mathbb{R}^{H\times C}, \qquad \boldsymbol{R}_t = \boldsymbol{Y}_t - \widehat{\boldsymbol{Y}}_t^{\mathrm{Base}}, \tag{1}$$

LCRT predicts $\boldsymbol{R}_t$. Candidate indices $k \in \{0, p, d\}$ denote Base, Pure LCRT, and Adaptive LCRT, with $\widehat{\boldsymbol{Y}}_t^0 = \widehat{\boldsymbol{Y}}_t^{\mathrm{Base}}$ and loss $\ell_j^k = (HC)^{-1}\|\boldsymbol{Y}_j - \widehat{\boldsymbol{Y}}_j^k\|_F^2$. At step t, SAR only knows losses for forecasts whose full target has already arrived (i.e., $j + H \le t$).

3.2 LOW-RANK CONDITIONAL RESIDUAL TRANSPORT

LCRT predicts residuals with coefficients shared across steps and variables; *transport* here simply means adding a correction, not optimal transport. Features $\boldsymbol{\phi}(\boldsymbol{X}_t) \in \mathbb{R}^F$ include levels, variability, trends, differences, autocorrelation, frequencies, and channel summaries, plus an intercept $\phi_0 = 1$ (Appendix E.3). LCRT fits a linear map on the validation data $\mathcal{V}_{\mathrm{fit}}$, where $S \in \mathbb{R}^{F\times F}$ standardizes the non-intercept features:

$$\boldsymbol{A}^* = \underset{\boldsymbol{A}\in\mathbb{R}^{F\times HC}}{\arg\min} \sum_{t\in\mathcal{V}_{\mathrm{fit}}} \|\operatorname{vec}(\boldsymbol{R}_t) - \boldsymbol{A}^\top S^\top \boldsymbol{\phi}(\boldsymbol{X}_t)\|_2^2 + \lambda|\mathcal{V}_{\mathrm{fit}}|\|\boldsymbol{A}_{-0}\|_F^2. \tag{2}$$

Here $\boldsymbol{A}_{-0}$ excludes the intercept. We reshape $\boldsymbol{B} = S\boldsymbol{A}^*$ into $\mathcal{B} \in \mathbb{R}^{F\times H\times C}$ and keep only the top singular directions along forecast steps and channels, giving $\boldsymbol{U}_H \in \mathbb{R}^{H\times r_H}$ and $\boldsymbol{U}_C \in \mathbb{R}^{C\times r_C}$ with small ranks. Projecting onto these two bases yields

$$\Pi_r(\mathcal{B}) = \mathcal{B} \times_2 (\boldsymbol{U}_H\boldsymbol{U}_H^\top) \times_3 (\boldsymbol{U}_C\boldsymbol{U}_C^\top), \qquad \widehat{\boldsymbol{Y}}_t^k = \widehat{\boldsymbol{Y}}_t^{\mathrm{Base}} + \alpha_k \boldsymbol{\Delta}_t^k, \tag{3}$$

where $\times_2$ and $\times_3$ project along forecast steps and channels. For a spatial map, $\Delta_{t,hc}^k = \sum_f \phi_f(\boldsymbol{X}_t)[\Pi_r(\mathcal{B}^k)]_{fhc}$, and $\alpha_k > 0$ sets the correction size. Pure LCRT uses this spatial map whenever it can be fitted. Adaptive LCRT instead picks a spatial map ($C \geq 32$), a shared temporal map with 14 features per channel, or a blend of shared and channel-specific coefficients; temporal maps project only along forecast steps.

Time-ordered validation folds choose the map, ridge penalty, rank, blending weight, and positive step. Gaps between folds prevent fitting targets from leaking into later evaluation. The chosen coefficients are fitted once and frozen (Appendix E.3). SAR then selects the output:

$$a_t = \mathrm{SAR}(\mathcal{I}_t) \in \{0, p, d\}, \qquad \widetilde{\boldsymbol{Y}}_t = \begin{cases} \widehat{\boldsymbol{Y}}_t^{\mathrm{Base}}, & a_t = 0, \\ \widehat{\boldsymbol{Y}}_t^{p}, & a_t = p, \\ \widehat{\boldsymbol{Y}}_t^{d}, & a_t = d. \end{cases} \tag{4}$$

The three candidates are stored for later scoring. SAR never blends them or changes their correction sizes during testing.

3.3 Selection under Delayed Feedback

Validation blocks select the backbone, fit corrections, and assess candidates, separated by $L + H$- and H-origin gaps (Appendix D.1). Validation may only shrink correction steps and cannot set Pure LCRT's step to zero. It picks the step with the largest lower confidence bound (LCB), breaking ties by mean gain and then smaller step.

For recent origins $\mathcal{W}_t$ with complete targets, the normalized gain is

$$z_{j,t}^k = 100 \frac{\ell_j^{\mathrm{Base}} - \ell_j^k}{\max\{|\mathcal{W}_t|^{-1} \sum_{i\in\mathcal{W}_t} \ell_i^{\mathrm{Base}}, 10^{-12}\}}. \tag{5}$$

A positive gain means lower loss than Base; the denominator is a common, nonzero scale. Because nearby forecasts overlap, their losses are correlated, so we use Newey–West standard errors (Newey & West, 1987) to build lower and upper confidence bounds (LCB/UCB) on the mean gain.

To start, SAR needs a positive 90% LCB and gains in at least two-thirds of the validation blocks; otherwise it stays with Base. Online, it compares the last $8H$ completed forecasts, updates every H arrivals, needs at least $\max(8, 3H)$ observations, and checks gains over five consecutive blocks.

Entering a correction from Base needs a positive mean gain and two positive blocks. Switching from Pure LCRT to Adaptive LCRT additionally needs both candidates to qualify, a positive paired 90% LCB, and four positive blocks for $\ell_j^p - \ell_j^d$; the reverse needs an eligible Pure LCRT, a negative paired UCB, and four negative blocks. From Base, SAR prefers Adaptive LCRT when the paired test passes, then an eligible Pure LCRT, then an eligible Adaptive LCRT. An active candidate stays until a switch or a budget exit; the full rule order is in Appendices D.3 and E.4. These bounds guide the decision but do not guarantee future performance.

3.4 Risk Budget and Delay Reserve

The budget tracks the forecasts SAR actually issued. An episode starts when a correction is entered at τ and ends when SAR returns to Base. Its completed-forecast gain is

$$B_t = \sum_{j=\tau}^{t-H} (\ell_j^{\mathrm{Base}} - \ell_j^{a_j}). \tag{6}$$

An empty sum is zero; positive B_t means the issued forecasts have lower total loss than Base. Because the most recent forecasts still await feedback, SAR also keeps a delay reserve. Let $\delta_t^{(-1)}$ be the positive part of $\ell_j^{a_j} - \ell_j^{\mathrm{Base}}$ summed over $j = t-2H+1, \ldots, t-H$, and $\delta_t^{(-2)}$ use the preceding non-overlapping H origins; both blocks are clipped to $j \geq \tau$, and empty blocks contribute zero. The reserve is

$$D_t = \delta_t^{(-1)} + \delta_t^{(-2)} \mathbf{1}[\delta_t^{(-1)} > 0] \mathbf{1}[\delta_t^{(-2)} > 0]. \tag{7}$$

The preceding block counts only when both blocks have net losses. This is observed downside, not a bound on future losses. After H episode forecasts mature, SAR checks the budget before selection and exits if $B_t < D_t - 10^{-12}$ (the tolerance only handles roundoff); the first check is at $\tau + 2H - 1$. Switching keeps the account, and re-entry starts from zero. So continuing after a check means $B_t \geq D_t - 10^{-12}$ for completed forecasts, though forecasts still awaiting feedback may later erase that gain.

SAR adds no backbone pass or gradient update. A streaming controller needs $O(H)$ storage; the reference replay keeps full history (Appendix D.4). Appendix C.4 measures runtime.

4 Experiments

We ask two questions. First, does SAR-LCRT make an already trained forecasting model more accurate? Second, is the "selective" part of SAR-LCRT useful on top of simply correcting every forecast? To answer fairly, we compare the three methods on exactly the same setting: for each task, Base, Pure LCRT, and SAR-LCRT all read the same saved (already trained) model and are scored on the same test targets. In other words, the only thing that differs is the correction and selection step, not the underlying model. We first report results for Base and SAR-LCRT tasks, then check whether the SAR-LCRT improvements hold across different models, forecast lengths and seeds.

4.1 Experimental Setup

Backbones and datasets. We evaluate eight forecasting models: FACT, FreDN, MDMLP-EIA, VPNet, XLinear, SimpleTM, iTransformer, and PatchTST (Wang et al., 2026b; An et al., 2026; Zhang et al., 2026; Liu et al., 2026; Chen et al., 2026; 2025; Liu et al., 2024; Nie et al., 2023). We use ECL, ETTh1/2, ETTm1/2, Weather, PEMS03/04/07/08, PEMS-BAY, and LargeST-SD/GBA. Inputs contain 96 observations; prediction lengths are $\{96, 192, 336, 720\}$ for ECL/ETT/Weather and $\{12, 24, 36, 48\}$ for PEMS03/04/07/08, PEMS-BAY, and LargeST-SD/GBA. Each model–dataset–length combination is a task: $8 \times 13 \times 4 = 416$ tasks, each with seeds 2025–2027. LargeST uses multivariate forecasting rather than its original graph setup. Splits, training, and normalization are in Appendix D.1.

Comparison methods and testing. We compare three things. Base simply returns the original model's forecast. Pure LCRT always adds a learned correction to the forecast. SAR-LCRT, at each time step, picks one of three outputs: the original forecast (Base), the always-corrected forecast (Pure LCRT), or the adaptively-corrected forecast (Adaptive LCRT). All three start from the same saved model for each task and training run (seed), so their only difference is this correction-and-selection step. No model weights are updated during testing. The online methods never know the future in advance. They wait until the full prediction window has actually happened, and only then use its true results as feedback. How the corrections are fitted and how the initial choice is made are explained in Section 3.

Metrics and averaging. MSE averages squared prediction errors, giving larger misses more weight; MAE averages their absolute values. Lower values are better. For task q, seed s, metric $e \in \{\mathrm{MSE}, \mathrm{MAE}\}$, and method k, we average seed errors, calculate each task's reduction from Base, then average over tasks $\mathcal{Q}$:

$$\bar{e}_q^k = \tfrac{1}{3} \sum_s e_{qs}^k, \qquad I_k^e(q) = 100(1 - \bar{e}_q^k / \bar{e}_q^{\mathrm{Base}}), \qquad \overline{I}_k^e = |\mathcal{Q}|^{-1} \sum_q I_k^e(q). \tag{8}$$

Each task has equal weight. Averaging relative changes prevents datasets with larger raw error scales from dominating the overall comparison. Table 1 also averages four prediction lengths, giving 104 model–dataset pairs. W/E/L counts wins, equal values, and losses against the stated comparator, with tolerance 10^{-9} unless specified; residual-baseline comparisons use ± 0.1 percentage points. Confidence intervals use 20,000 bootstrap samples of whole datasets (seed 2701), keeping all tasks within each of the 13 datasets together. Tables 8 and 9 report seed variation in Appendix B.1.

Table 1: **Forecasting accuracy with and without SAR-LCRT.** Input length is 96. Errors average three seeds and horizons $\{96, 192, 336, 720\}$ for ETT/ECL/Weather or $\{12, 24, 36, 48\}$ for PEMS/LargeST. For SAR-LCRT versus Base at three decimals, red bold marks lower errors; underlined red bold marks equal values.

Dataset	Variant	FACT (2026)		FreDN (2026)		MDMLP-EIA (2026)		VPNet (2026)		XLinear (2026)		SimpleTM (2025)		iTransformer (2024)		PatchTST (2023)	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	Base	0.440	0.421	0.387	0.398	0.388	0.398	0.412	0.415	0.407	0.406	0.389	0.399	0.418	0.416	0.386	0.402
	SAR-LCRT	**0.437**	**0.421**	**0.384**	**0.398**	**0.386**	**0.398**	**0.411**	**0.414**	**0.405**	**0.406**	**0.386**	**0.398**	**0.416**	**0.415**	**0.384**	**0.401**
ETTm2	Base	0.277	0.323	0.278	0.322	0.278	0.324	0.345	0.362	0.280	0.323	0.282	0.326	0.299	0.340	0.294	0.335
	SAR-LCRT	**0.274**	**0.323**	**0.275**	**0.322**	**0.275**	**0.324**	**0.343**	**0.362**	**0.277**	**0.323**	**0.280**	**0.326**	**0.295**	**0.339**	**0.291**	**0.335**
ETTh1	Base	0.453	0.443	0.442	0.435	0.446	0.437	0.524	0.483	0.478	0.451	0.446	0.442	0.477	0.461	0.470	0.455
	SAR-LCRT	0.459	0.448	0.449	0.441	0.453	0.443	0.526	0.484	0.484	0.457	0.453	0.448	0.480	0.464	0.471	0.457
ETTh2	Base	0.380	0.403	0.380	0.403	0.376	0.400	0.444	0.435	0.386	0.404	0.388	0.405	0.402	0.416	0.387	0.409
	SAR-LCRT	**0.380**	**0.403**	**0.380**	**0.403**	**0.376**	**0.400**	**0.444**	**0.435**	**0.386**	**0.404**	**0.388**	**0.405**	**0.402**	**0.416**	**0.387**	**0.409**
ECL	Base	0.243	0.328	0.184	0.272	0.187	0.275	0.176	0.269	0.183	0.278	0.166	0.259	0.180	0.271	0.189	0.274
	SAR-LCRT	**0.243**	0.330	**0.184**	0.273	**0.187**	**0.275**	**0.176**	**0.269**	**0.183**	**0.278**	**0.166**	0.260	**0.180**	**0.271**	**0.189**	**0.274**
Weather	Base	0.260	0.286	0.248	0.275	0.246	0.273	0.246	0.276	0.243	0.273	0.244	0.272	0.260	0.280	0.256	0.280
	SAR-LCRT	**0.254**	0.296	**0.242**	0.281	**0.242**	0.282	**0.244**	0.283	**0.239**	0.278	**0.240**	0.278	**0.255**	0.292	**0.252**	0.293
PEMS03	Base	0.171	0.281	0.148	0.254	0.149	0.253	0.097	0.207	0.102	0.208	0.095	0.202	0.109	0.218	0.148	0.252
	SAR-LCRT	**0.145**	**0.262**	**0.131**	**0.241**	**0.129**	**0.239**	**0.093**	**0.203**	**0.098**	**0.204**	**0.091**	**0.198**	**0.103**	**0.212**	**0.131**	**0.240**
PEMS04	Base	0.200	0.304	0.168	0.274	0.180	0.280	0.100	0.210	0.100	0.211	0.102	0.213	0.132	0.237	0.192	0.290
	SAR-LCRT	**0.164**	**0.279**	**0.151**	**0.262**	**0.155**	**0.265**	**0.098**	**0.207**	**0.097**	**0.208**	**0.099**	**0.209**	**0.121**	**0.228**	**0.165**	**0.272**
PEMS07	Base	0.166	0.277	0.142	0.245	0.140	0.240	0.083	0.189	0.091	0.191	0.087	0.187	0.103	0.205	0.170	0.269
	SAR-LCRT	**0.141**	**0.257**	**0.137**	**0.241**	**0.133**	**0.236**	**0.080**	**0.185**	**0.090**	**0.190**	**0.085**	**0.183**	**0.095**	**0.199**	**0.145**	**0.255**
PEMS08	Base	0.191	0.291	0.180	0.275	0.181	0.270	0.113	0.210	0.128	0.228	0.112	0.210	0.130	0.226	0.168	0.262
	SAR-LCRT	**0.159**	**0.269**	**0.149**	**0.256**	**0.148**	**0.250**	**0.105**	**0.204**	**0.112**	**0.216**	**0.102**	**0.204**	**0.119**	**0.219**	**0.150**	**0.251**
PEMS-BAY	Base	0.531	0.349	0.515	0.346	0.541	0.360	0.417	0.299	0.483	0.307	0.410	0.293	0.490	0.320	0.564	0.358
	SAR-LCRT	**0.505**	**0.346**	**0.501**	**0.345**	**0.522**	**0.358**	**0.407**	**0.294**	**0.470**	**0.305**	**0.400**	**0.287**	**0.469**	**0.320**	**0.529**	**0.358**
LargeST-SD	Base	0.071	0.167	0.045	0.128	0.048	0.130	0.036	0.112	0.034	0.109	0.034	0.109	0.035	0.111	0.055	0.139
	SAR-LCRT	**0.070**	0.172	**0.045**	0.129	**0.048**	0.132	**0.036**	**0.112**	**0.033**	**0.108**	**0.033**	**0.107**	**0.035**	**0.110**	**0.055**	0.144
LargeST-GBA	Base	0.102	0.206	0.066	0.155	0.067	0.156	0.060	0.147	0.056	0.143	0.052	0.137	0.054	0.137	0.103	0.205
	SAR-LCRT	**0.094**	**0.199**	**0.066**	0.156	**0.067**	**0.156**	**0.060**	**0.146**	**0.055**	**0.142**	**0.052**	**0.136**	**0.053**	**0.137**	**0.091**	**0.194**

4.2 Forecasting Performance

Improvements across the evaluation grid. SAR-LCRT matches or lowers MSE in **96 of 104 model–dataset pairs (92.3%)** at the three-decimal precision of Table 1. Each entry averages four prediction lengths. Before that averaging, it matches or improves Base on 358 of 416 tasks (86.1%), using three-seed means. So the gain is not just an artifact of averaging horizons. Equal rounded entries mean the displayed errors are equal, not that the predictions are identical. A bootstrap check in Appendix A.2 shows the overall MSE reduction is clearly above zero, so SAR-LCRT helps on average across the datasets we tested.

A broad benefit with dataset-dependent strength. The largest gains are on the PEMS datasets, especially PEMS03/04/07/08. ETTm1, ETTm2, Weather, and LargeST-GBA also improve on average MSE. Dropping any single dataset family or backbone still leaves a positive average MSE reduction, and the reduction stays positive after excluding the four main PEMS datasets. So no single family is driving the whole result. Table 5 (in Appendix A.1) and Figure 2 (in Appendix A.2) break down the dataset pattern; Table 7 (in Appendix A.2) checks the averaging.

Table 2: **Forecasting errors and relative reductions on PEMS.** Each row covers 32 tasks (eight backbones and four prediction lengths), with three seeds per task. MSE/MAE columns report mean errors; gains average each task's percentage reduction from its own Base. ↓ marks an error reduction. Red bold marks lower errors.

Dataset	Tasks	Base MSE	SAR-LCRT MSE	MSE gain	Base MAE	SAR-LCRT MAE	MAE gain
PEMS03	32	0.127	**0.115**	**↓7.73%**	0.234	**0.225**	**↓3.57%**
PEMS04	32	0.147	**0.131**	**↓7.98%**	0.252	**0.241**	**↓3.81%**
PEMS07	32	0.123	**0.113**	**↓5.94%**	0.225	**0.218**	**↓2.67%**
PEMS08	32	0.150	**0.131**	**↓11.06%**	0.247	**0.234**	**↓4.73%**
PEMS-BAY	32	0.494	**0.475**	**↓3.40%**	0.329	**0.327**	**↓0.73%**

4.3 Performance on PEMS Benchmarks

Agreement across models and prediction lengths. Table 2 shows that SAR-LCRT lowers both mean MSE and mean MAE on all five evaluated PEMS datasets. Across PEMS03/04/07/08, the average MSE reduction is 8.18%, with the largest dataset-level reduction of 11.06% on PEMS08. The improvement in both metrics holds for all 128 model–horizon tasks on these four datasets and all 384 corresponding seed-level comparisons. Thus each backbone benefits at every tested prediction length, and the result persists across training seeds. These results show that SAR-LCRT improves forecasting on the PEMS benchmarks under both MSE and MAE, with consistent evidence across PEMS03/04/07/08.

Benefits persist through the test period. To check whether the gains last across the whole test period, we split the test origins of PEMS03 and PEMS08 into ten consecutive, non-overlapping blocks per dataset. For FACT at $H = 48$, SAR-LCRT achieves lower MSE than Base in every block after averaging over three seeds (block MSE in Figure 4; rolling errors in Figure 6, both in Appendix A.3). The gains therefore hold from the beginning to the end of both PEMS test periods, rather than coming from a single lucky stretch. Since the correction maps stay fixed during testing, this shows they keep producing useful forecasts when combined with SAR's online selection. Complete task and seed results are in Appendix B.

5 Ablation Studies

We run three methods on the same 416 tasks: Base (the original forecast), Pure LCRT (always add a learned correction), and SAR-LCRT (choose among Base, Pure LCRT, and Adaptive LCRT). All three share the same frozen checkpoint and test stream for each task and seed, so the comparison isolates the correction and selection parts. This answers two questions: first, does adding a learned residual map help a frozen forecaster; and second, does SAR's selection step help beyond always applying that map.

Table 3 answers both questions. Pure LCRT improves both MSE and MAE over Base for every one of the eight backbones, so a frozen forecaster still leaves useful patterns in its residuals that a single map can exploit. SAR-LCRT then improves on Pure LCRT for every backbone too: the extra MSE reduction ranges from 0.75% to 1.72%, and across all 416 tasks SAR-LCRT lowers or matches Pure LCRT's MSE on 288 of 416 tasks (about 69%). Averaged over all tasks, SAR-LCRT reduces MSE by 3.31% relative to Base and by 1.25% relative to Pure LCRT. So the learned map and the selection step each make a separate, measurable contribution.

The table also shows that the benefit is not driven by one architecture: every backbone improves in both MSE and MAE, and removing any single backbone still leaves a positive average MSE reduction (Appendix A.2). Selection does real work too: SAR uses a residual candidate for 87.2% of forecasts on average, so the improvement comes from choosing among corrections most of the time rather than from rarely intervening.

A closer look at the datasets shows where the gains concentrate (Table 5 in Appendix A.1; the heatmap in Figure 2, Appendix A.2). On the PEMS and LargeST-GBA datasets, SAR-LCRT low-

Table 3: **Framework ablation across eight frozen backbones.** Each backbone has 52 tasks (13 datasets and four lengths), with three seeds per task. MSE/MAE gains are mean percentage error reductions from Base; SAR-LCRT vs Pure LCRT reports the mean taskwise MSE reduction from Pure LCRT. ↓ denotes a reduction. W/E/L counts how often SAR-LCRT wins, equals, or loses against Pure LCRT on MSE. SAR(Act.) is the mean percentage of forecasts using a correction.

Backbone	Tasks	Pure LCRT vs Base		SAR-LCRT vs Base		SAR-LCRT vs Pure LCRT	W/E/L	SAR(Act.)
		MSE gain	MAE gain	MSE gain	MAE gain	MSE gain		
FACT	52	↓4.29%	↓1.38%	↓5.84%	↓1.88%	↓1.72%	36/1/15	88.2%
FreDN	52	↓2.29%	↓0.59%	↓3.32%	↓0.89%	↓1.11%	31/1/20	85.3%
MDMLP-EIA	52	↓2.55%	↓0.46%	↓3.86%	↓0.93%	↓1.42%	28/2/22	88.1%
VPNet	52	↓0.61%	↓0.03%	↓1.64%	↓0.60%	↓1.04%	37/0/15	85.1%
XLinear	52	↓0.80%	↓0.03%	↓1.98%	↓0.48%	↓1.21%	39/1/12	86.6%
SimpleTM	52	↓0.62%	↓0.13%	↓1.86%	↓0.64%	↓1.26%	40/0/12	86.5%
iTransformer	52	↓1.54%	↓0.19%	↓2.99%	↓0.62%	↓1.49%	38/0/14	89.4%
PatchTST	52	↓4.25%	↓1.09%	↓4.96%	↓1.17%	↓0.75%	31/3/18	88.4%
Overall	416	↓2.12%	↓0.49%	↓3.31%	↓0.90%	↓1.25%	280/8/128	87.2%

ers MSE by 6.54% on average, about twice its overall 3.31%, and SAR keeps a correction on for 99.2% of those forecasts, compared with 87.2% overall. So the larger average gains come exactly where the correction is reliable enough to be used almost constantly. On LargeST-SD, always applying Pure LCRT is worse than Base, but SAR-LCRT still improves it (Appendix A.1), showing the value of returning to Base when a correction stops paying off. The full per-dataset breakdown is in Appendix A.1.

6 How SAR-LCRT Works and What It Costs

Why the correction helps. A trained forecaster is not randomly wrong. It often under-predicts rising series, over-predicts falling ones, or repeats the same error across nearby steps and related variables. LCRT learns a mapping from the input features to the error the backbone will make, and adds the opposite as a correction. It fits this map once and keeps it fixed, so the backbone is never retrained. To keep the map small, it shares coefficients across forecast steps and variables instead of learning a separate number for every future position; the full construction is in Appendix E.3. This shared, low-rank structure is what lets one compact map improve many different forecasts.

Why selection is needed. A map that fits the validation errors well can still be wrong later, because the data may drift. SAR therefore keeps three choices at every step: use the original forecast (Base), always apply the correction (Pure LCRT), or apply the more flexible correction (Adaptive LCRT). Once the true future values arrive, SAR scores the three stored forecasts on a recent window and keeps the one that has been doing best. If a correction stops helping, SAR returns to Base; if it starts helping again, SAR can bring it back, all without fitting a new map. Learning the map and deciding to use it are thus two separate, fixed parts of the system.

Only the forecasts actually issued matter. Choosing a good candidate is not enough if that candidate is not the one delivered to the user. SAR keeps a running account of the losses of the forecasts it actually issued, measured against Base. When that account shows a correction is no longer paying off, SAR returns to Base. This is what makes the selection risk-gated: the system is judged by what it really output, not by its best unused candidate. Across the full evaluation this combination improves the average MSE, and the benefit still holds after removing any single dataset family (Table 6 in Appendix A.2).

Speed and memory. We time two example backbones, MDMLP-EIA and SimpleTM, on ETTh1 and PEMS08 using the same frozen checkpoint, one forecast at a time. Table 4 reports the timing and memory results.

Table 4: **SAR-LCRT speed and memory relative to Base.** All values are ratios relative to Base (Base = 1). Speed is the ratio of prediction times (Base time divided by SAR-LCRT time), so values below 1 mean slower. GPU memory is peak allocated tensor memory; CPU memory is peak process memory. Absolute timings and memory sizes are in Appendix C.4.

Backbone	Dataset	H	Speed Ratio	GPU Memory Ratio	CPU Memory Ratio
MDMLP-EIA	ETTh1	336	0.8599	1.0000	1.0004
MDMLP-EIA	PEMS08	48	0.4787	1.0000	1.0012
SimpleTM	ETTh1	336	0.8721	1.0000	1.0003
SimpleTM	PEMS08	48	0.5885	1.0000	1.0008

Two conclusions follow from Table 4. First, the only meaningful cost is time. SAR-LCRT is about 1.15× to 2.09× slower than Base, because it adds feature computation, map evaluation, and selection to every backbone pass. The relative slowdown is larger on PEMS08 (H=48), where Base is already fast (16–23 ms), so the fixed extra work of roughly 4–18 ms per forecast matters more. Second, memory is almost unchanged: peak GPU memory stays exactly the same (ratio 1.000), and CPU memory grows by only 0.30–1.15 MB (ratio 1.0003–1.0012); the absolute values are in Table 20 (Appendix C.4). The per-configuration latency, the one-time fitting cost (about 401–648 seconds, paid once per checkpoint), and the per-decision controller update (230–880 microseconds) are reported in Appendix C.4; the storage and computation bounds are in Appendix D.4.

In short, SAR-LCRT trades a modest, predictable slowdown for better forecasts, while leaving memory almost unchanged.

7 Discussion and Limitations

SAR-LCRT treats "improving a model" and "deciding when to use the improvement" as two separate steps. This is useful because it works with any already-trained forecaster: the backbone stays fixed, and a new correction map or selection rule can be added without retraining.

Our results suggest a practical way to think about deployment: the model's parameters are one lever, and the choice of which forecast to actually issue is another. Here the correction map is learned once on validation data, while the choice of output adapts to delayed test feedback, so the system can react to new data without re-learning the backbone. Across eight architectures this output-selection step kept helping after the forecaster was frozen, so a model and its deployment policy are worth evaluating together.

A few limits are worth noting. SAR-LCRT needs held-out validation data and complete target feedback, so it can only score a correction after the full prediction window has passed. On a few settings a correction can briefly raise error before the fallback takes over (Appendix C.2), and inference adds a modest amount of time (Appendix C.4). On average both MSE and MAE improve, though the MAE improvement is smaller (Appendix A.2).

8 Conclusion

We proposed SAR-LCRT to improve frozen forecasters through learned residual corrections and delayed-feedback selection. LCRT captures shared error patterns across forecast steps and variables, while SAR decides when to use or withdraw corrections without retraining the backbone. Across eight architectures and 13 datasets, SAR-LCRT matches or improves Base MSE on 358 of 416 tasks using three-seed means. PEMS08 has the largest dataset-level MSE reduction, averaging 11.06% across architectures and horizons; the largest task-level reduction is 24.57% for MDMLP-EIA on PEMS08 at horizon 48. On PEMS03/04/07/08, both metrics improve across all 128 task means and 384 seed-level comparisons, supporting consistent benefits across backbones, horizons, and seeds. These results support combining structured correction with online selection while retaining the original forecast as a fallback. Future work will explore richer corrections and their selection under distribution shifts and longer feedback delays.

AI USE STATEMENT

Generative AI tools were used for manuscript organization and language editing; literature discovery and metadata checks; critique of experimental design; interpretation of author-produced result summaries; LaTeX, table, and figure layout; code drafting and trace checks; and consistency review. They were not used to synthesize, alter, or replace reported observations; the experiments were executed by the authors. Numerical claims were recomputed from repository result files, method descriptions were checked against source code and documentation, and the authors reviewed all assisted work. The authors remain responsible for the manuscript, artifacts, and citation accuracy.

REPRODUCIBILITY STATEMENT

Section 3, Section 4, and Appendix D specify the algorithm, splits, purges, temporal information sets, aggregation, and configurations. Appendix B.1 reports exact three-seed dispersion without imputation; Appendix C.1 gives residual-baseline comparisons and reproduction checks. Appendix C.2 audits actions, B_t, D_t, candidate/issued losses, and maturation indices. Appendix D documents the aggregation order and numerical conventions needed to reconstruct the tables.

REFERENCES

Zhongde An, Jinhong You, Jiyanglin Li, Yiming Tang, Wen Li, Heming Du, and Shouguo Du. FreDN: Spectral disentanglement for time series forecasting via learnable frequency decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19623–19631, 2026. doi: 10.1609/aaai.v40i24.39042.

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

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. NHITS: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6989–6997, 2023. doi: 10.1609/aaai.v37i6.25854.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. SimpleTM: A simple baseline for multivariate time series forecasting. In *International Conference on Learning Representations*, 2025.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International Conference on Learning Representations*, 2024.

Xinyang Chen, Huidong Jin, Yu Huang, and Zaiwen Feng. XLinear: A lightweight and accurate MLP-based model for long-term time series forecasting with exogenous inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20325–20335, 2026. doi: 10.1609/aaai.v40i24.39121.

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. PMLR, 2024.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Xiannan Huang, Shuhan Qiu, Jiayuan Du, and Chao Yang. Online time series prediction using feature adjustment. In *International Conference on Learning Representations*, 2026.

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.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1453–1461. PMLR, 2013.

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.

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.

Dilfira Kudrat, Zongxia Xie, Yanru Sun, Tianyu Jia, and Qinghua Hu. Patch-wise structural loss for time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31841–31859. PMLR, 2025.

Ying-yee Ava Lau, Zhiwen Shao, and Dit-Yan Yeung. Fast and slow streams for online time series forecasting without information leakage. In *International Conference on Learning Representations*, 2025.

Jonghyeok Lee, Chen Xu, and Yao Xie. Kernel-based optimally weighted conformal time-series prediction. In *International Conference on Learning Representations*, 2025.

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.

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. doi: 10.52202/068431-0718.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with koopman predictors. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0538.

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.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.

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.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

Quang Pham, Chenghao Liu, Doyen Sahoo, and Steven C. H. Hoi. Learning fast and slow for online time series forecasting. In *International Conference on Learning Representations*, 2023.

Xiangfei Qiu, Xingjian Wu, Hanyin Cheng, Xvyuan Liu, Chenjuan Guo, Jilin Hu, and Bin Yang. DBLoss: Decomposition-based loss function for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 31546–31573, 2025. doi: 10.52202/085713-0936.

Tianyi Shi, Zhu Meng, Yue Chen, Siyang Zheng, Fei Su, Jin Huang, Changrui Ren, and Zhicheng Zhao. OLMA: One loss for more accurate time series forecasting. *arXiv preprint arXiv:2505.11567*, 2025. URL https://arxiv.org/abs/2505.11567.

Hao Wang, Licheng Pan, Zhichao Chen, Xu Chen, Qingyang Dai, Lei Wang, Haoxuan Li, and Zhouchen Lin. Time-o1: Time-series forecasting needs transformed label alignment. In *Advances in Neural Information Processing Systems*, volume 38, pp. 9591–9621, 2025a. doi: 10.52202/085713-0294.

Hao Wang, Licheng Pan, Yuan Shen, Zhichao Chen, Degui Yang, Yifei Yang, Sen Zhang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. FreDF: Learning to forecast in the frequency domain. In *International Conference on Learning Representations*, 2025b.

Hao Wang, Licheng Pan, Yuan Lu, Zhichao Chen, Tianqiao Liu, Shuting He, Zhixuan Chu, Qingsong Wen, Haoxuan Li, and Zhouchen Lin. Quadratic direct forecast for training multi-step time-series forecast models. In *International Conference on Learning Representations*, 2026a.

Huiqiang Wang, Jieming Shi, and Qing Li. FACT: Fine-grained across-variable convolution for multivariate time series forecasting. In *International Conference on Learning Representations*, 2026b.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024.

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. PMLR, 2024.

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.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1254–1262. PMLR, 2016.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11559–11569. PMLR, 2021.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Hu Zhang, Zhien Dai, Zhaohui Tang, and Yongfang Xie. MDMLP-EIA: Multi-domain dynamic MLPs with energy invariant attention for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28221–28229, 2026. doi: 10.1609/aaai.v40i33.40049.

Yi-Fan Zhang, Qingsong Wen, Xue Wang, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. OneNet: Enhancing time series forecasting models under concept drift by online ensembling. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3066.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

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. doi: 10.1609/aaai.v35i12.17325.

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. doi: 10.52202/068431-0921.

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. PMLR, 2022b.

Reading percentage results. As in Tables 2 and 3, $\downarrow x\%$ means the error went down by $x\%$ compared with the stated comparator, and $\uparrow x\%$ means it went up by $x\%$. Zero means no change at the displayed precision. Activation and coefficient-of-variation percentages have no arrow, because they describe how often a method is used or how much values vary, not an error change. Raw errors, standard deviations, counts, ratios, and signed loss accounts keep their original units.

A Results Across Datasets

A.1 Results for each dataset

Table 5: **Average MSE reductions by dataset.** Pure LCRT and SAR-LCRT show average percentage reductions from Base across tasks. SAR-LCRT vs Pure LCRT averages the percentage error reduction of SAR-LCRT relative to Pure LCRT; it is not the difference between the preceding columns. W/E/L compares SAR-LCRT with Pure LCRT (wins/equal/losses). Pos. is the percentage of tasks where SAR-LCRT improves Base. ↓ marks a reduction and ↑ an increase. Datasets are grouped after inspecting the results to summarize their patterns.

Dataset	Tasks	Pure LCRT vs Base	SAR-LCRT vs Base	SAR-LCRT vs Pure LCRT	W/E/L	Pos.
Larger gains: PEMS and LargeST-GBA						
LargeST-GBA	32	↓2.86%	↓3.15%	↓0.28%	17/2/13	100.0%
PEMS-BAY	32	↓2.18%	↓3.40%	↓1.24%	29/0/3	100.0%
PEMS07	32	↓4.04%	↓5.94%	↓1.99%	22/4/6	100.0%
PEMS03	32	↓5.67%	↓7.73%	↓2.20%	31/1/0	100.0%
PEMS04	32	↓5.86%	↓7.98%	↓2.30%	31/0/1	100.0%
PEMS08	32	↓5.93%	↓11.06%	↓5.53%	32/0/0	100.0%
Average		↓4.42%	↓6.54%	↓2.26%	162/7/23	100.0%
Moderate						
ETTm1	32	↓0.76%	↓0.64%	↑0.12%	9/1/22	84.4%
LargeST-SD	32	↑0.67%	↓0.90%	↓1.54%	32/0/0	71.9%
ETTm2	32	↓0.61%	↓1.16%	↓0.55%	27/0/5	100.0%
Weather	32	↓0.49%	↓1.94%	↓1.45%	22/0/10	100.0%
Average		↓0.30%	↓1.16%	↓0.85%	90/1/37	89.1%
Weak / negative						
ETTh1	32	↑0.24%	↑0.97%	↑0.72%	4/0/28	34.4%
ECL	32	0.00%	↓0.01%	0.00%	11/0/21	25.0%
ETTh2	32	↓0.03%	↓0.03%	0.00%	13/0/19	84.4%
Average		↑0.07%	↑0.31%	↑0.24%	28/0/68	47.9%

Table 5 reports the average improvement for each dataset. SAR-LCRT reduces MSE most on PEMS08: 11.06% relative to Base and 5.53% relative to Pure LCRT. ETTh1 has higher MSE than Base, while ECL and ETTh2 change little. On LargeST-SD, Pure LCRT raises MSE, whereas SAR-LCRT lowers it. So the overall result combines large PEMS improvements with smaller changes elsewhere. A small average change does not mean that SAR always selects Base: corrections are used for about 60% of forecasts on ECL and ETTh2.

Figure 2 shows the same pattern across the eight models. PEMS03/04/07/08 generally improve, ETTh1 has higher MSE, and ECL/ETTh2 remain close to Base. Each cell averages four prediction lengths, so the figure compares model–dataset results without identifying why datasets differ. Tables 11 and 12 report each prediction length separately.

A.2 Overall results and checks of the averaging procedure

Average results over all tasks. Across all 416 tasks, averaging three seeds within each task, SAR-LCRT reduces MSE by 3.31% relative to Base and by 1.25% relative to Pure LCRT. To check whether these gains are reliable, we recompute each gain many times by resampling whole datasets and report the middle 95% of the results as a range of plausible values. For MSE the ranges are [1.45%, 5.38%] and [0.48%, 2.18%], and both sit entirely above zero, so the MSE gains are reliable. The average MAE reduction from Base is 0.90%, but its range is [-0.30%, 2.10%], which includes zero, so the MAE gain is too small to be certain. These calculations include every dataset.

Table 6: **Removing one dataset family at a time.** Each row reports the average task-level error reduction after excluding the named family. ↓ marks a reduction and ↑ an increase.

Omitted family	Remaining tasks	MSE gain	MAE gain
ETT	288	↓4.68%	↓1.38%
ECL	384	↓3.58%	↓0.99%
Weather	384	↓3.42%	↓1.27%
PEMS03/04/07/08	288	↓1.14%	↑0.34%
PEMS-BAY	384	↓3.30%	↓0.92%
LargeST	352	↓3.54%	↓1.02%

To check whether one dataset family drives the result, we remove each family in turn and recompute the average. MSE still improves in every case. The four PEMS datasets contribute strongly, but they are not the whole story: without them, average MSE still improves, while average MAE rises. Table 6 keeps the complete estimates.

Table 7: **Changing task weights or excluding groups.** All calculations use the same eight-backbone, 416-task results. The last two rows give the range of average MSE reductions after removing each dataset or backbone in turn.

Analysis	MSE gain
Prespecified task macro	↓3.31%
Equal weight over the eight horizon values	↓3.10%
Leave one dataset out	↓2.66% to ↓3.66%
Leave one backbone out	↓2.94% to ↓3.54%

Table 7 checks whether the overall result depends on the averaging rule or on a single group. Giving each prediction length equal weight still yields lower MSE. Removing any one dataset or backbone also leaves a positive average reduction. The full-grid improvement therefore does not rely on one particular weighting or one favorable group; these are checks of the same results, not additional independent experiments.

Figure 3 shows larger improvements on the PEMS and LargeST datasets, which use shorter prediction lengths. Because the short-horizon and long-horizon groups use different datasets, we cannot tell how much of the gap comes from the data itself and how much from the shorter feedback delay. MAE usually rises in the long-horizon group. Frequent correction appears in both successful and unsuccessful tasks, so how often a forecast is corrected does not by itself explain performance (Section 7).

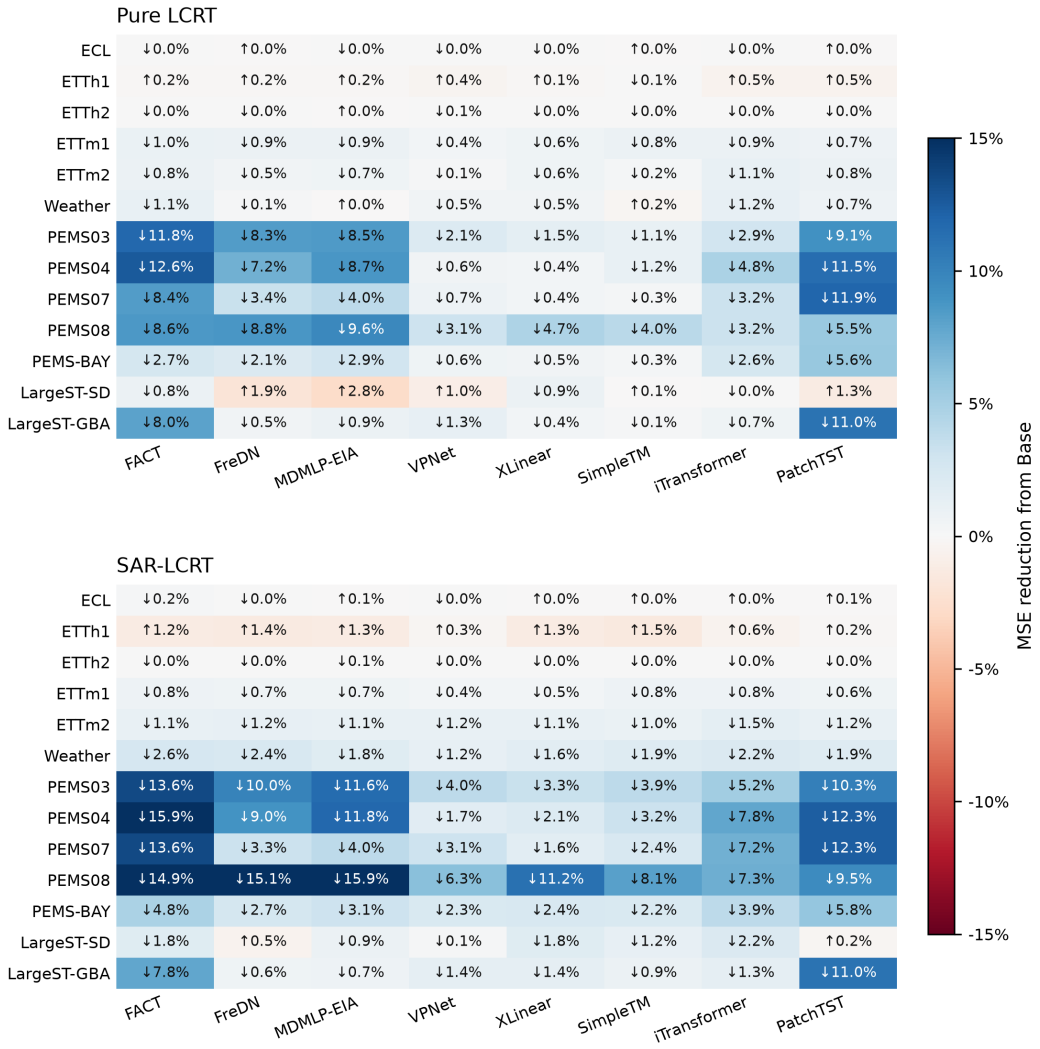


Figure 2: **MSE improvement for every dataset and backbone.** Each cell averages four prediction lengths. All eight backbones and 13 datasets are included. ↓ marks a reduction and ↑ an increase. The color scale shows signed percentage reductions.

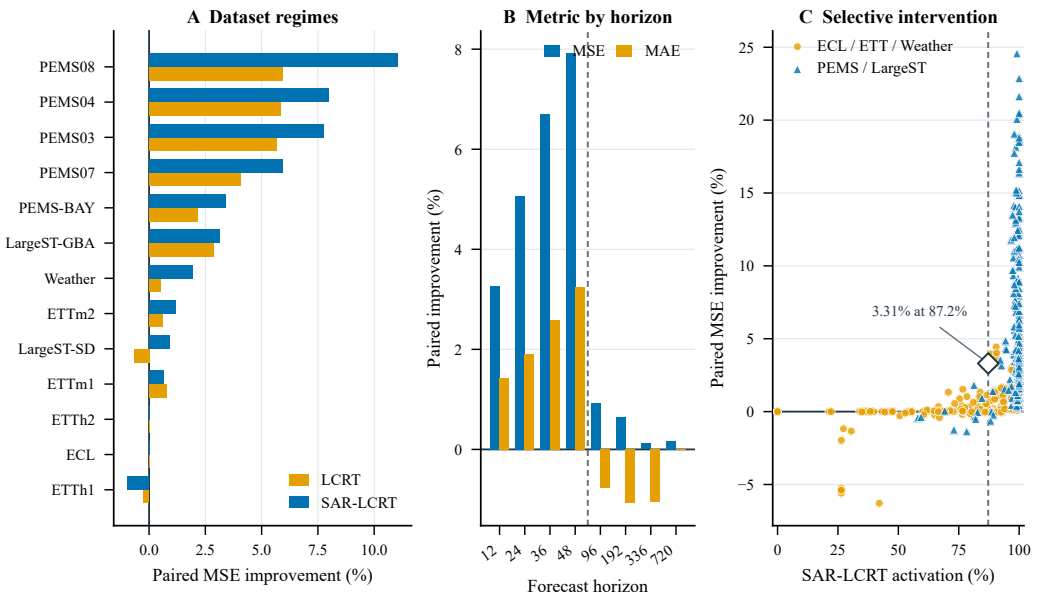


Figure 3: **How improvement varies across tasks. A:** Average MSE reduction for each dataset. **B:** Average MSE and MAE reductions for each prediction length; the divider separates PEMS/LargeST from ECL/ETT/Weather. **C:** Percentage of forecasts using correction versus MSE reduction. Each point is one task averaged over three seeds ($n = 416$); the diamond gives the averages across tasks.

A.3 How frozen FACT behaves during testing

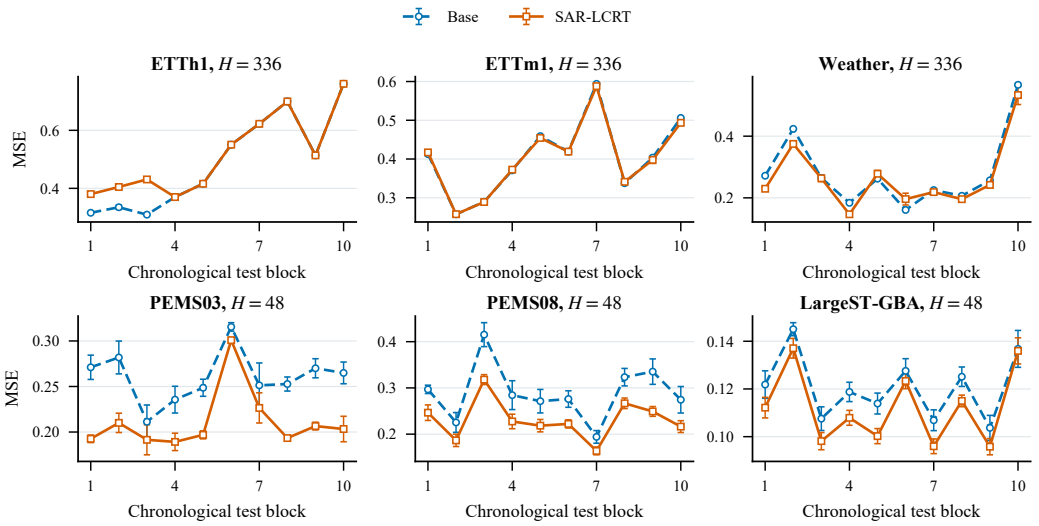


Figure 4: **Chronological deployment performance on frozen FACT.** Ten contiguous test blocks show the three-seed mean ± one sample SD. SAR-LCRT matches or lowers MSE in 51/60 blocks and lowers it in all 30 $H = 48$ PEMS03/PEMS08/LargeST-GBA blocks; ETTh1 ($H = 336$) has higher MSE in the first three, then falls back to Base. The horizontal axis is deployment time, not training iteration.

Figure 4 divides each of six FACT test runs into ten consecutive blocks. PEMS03, PEMS08, and LargeST-GBA improve across the blocks; ETTh1 initially has higher MSE and later returns to Base. Neighboring blocks are correlated, so they should not be counted as independent experiments.

The following figures use the same six datasets, frozen checkpoints, and recorded predictions. No model or correction map is refitted for these plots. Figure 5 documents checkpoint selection, so this diagnostic is included to distinguish backbone training from later deployment behavior. Rolling errors and forecast examples show what happens during testing.

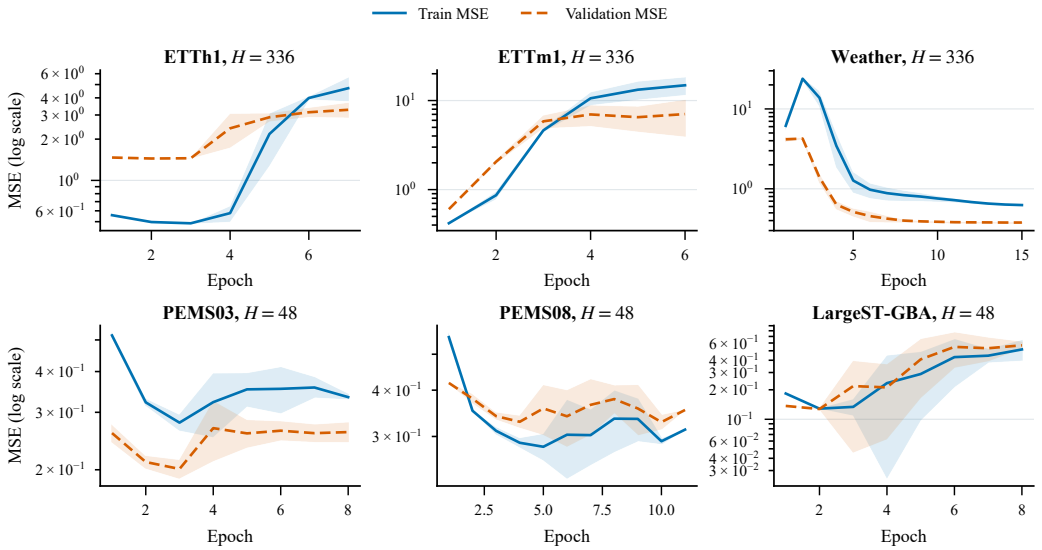


Figure 5: **Training and checkpoint selection for FACT.** Curves show mean training and validation-selection MSE over three seeds; shading shows one sample SD. These are backbone training curves. SAR-LCRT is fitted only after checkpoint selection. ETTh1 and ETTm1 errors do not decrease steadily, which illustrates why the selected checkpoint need not be the final epoch.

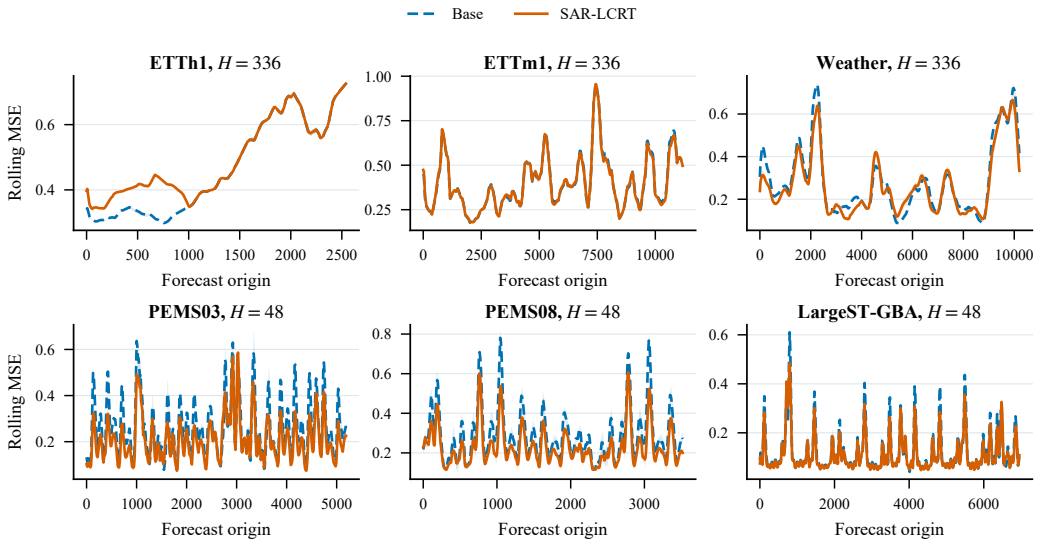


Figure 6: **Rolling test MSE for frozen FACT.** Each point averages errors over H consecutive forecast origins. Curves and shading show the three-seed mean and one sample SD. PEMS03/08 and LargeST-GBA have lower errors through testing. On ETTh1, early errors are higher, and the curves coincide after SAR returns to Base. Figure 4 summarizes the same runs.

Figure 7 illustrates how issuing Base or an adjusted candidate changes an individual prediction. It is included to make the outputs interpretable, not to estimate average accuracy. The examples are chosen by a fixed rule, regardless of their errors. On ETTh1 and ETTm1, SAR selects Base, so the

forecast curves overlap. On Weather, the Adaptive LCRT correction still leaves visible errors. In the PEMS08 panel, the Adaptive LCRT correction lowers much of the forecast, bringing parts of the predicted trajectory closer to the observed values. Together with the unchanged Base outputs, this illustrates the two available actions: adjust a forecast or leave it untouched. It does not establish average accuracy; that is measured by the full result tables.

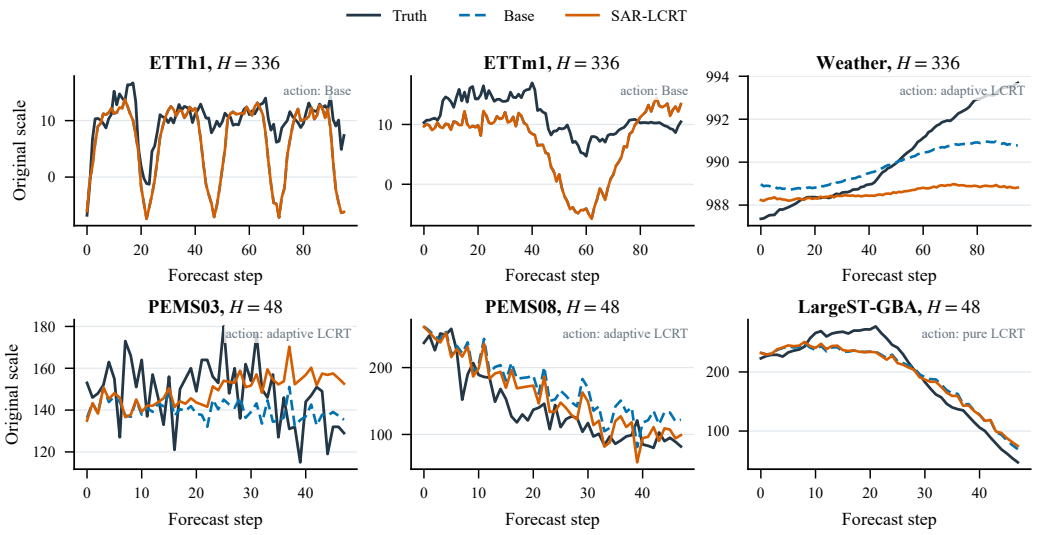


Figure 7: **FACT forecast examples selected by a fixed rule.** Each panel uses seed 2025, channel 0, and the origin at 60% of the test run. Selection does not inspect targets or errors. Values use the original scale, and each panel names the action taken by SAR. For $H = 336$, only the first 96 steps are shown. Base is the original model's forecast. These examples illustrate individual predictions; the tables measure average performance.

B Variation Across Seeds and Complete Results

B.1 Variation and Improvements Across Three Seeds

Table 8: **Variation and MSE improvements across three seeds.** For each backbone, SD columns give the median sample standard deviation across its 52 tasks. SAR CV is $100\,\mathrm{SD}/\mathrm{mean}$. A 3/3 win means lower SAR-LCRT MSE at all three matching seeds; $\geq$2/3 means at least two seeds improve. The comparison tolerance is 10^{-9}.

Backbone	Tasks	Base SD	SAR SD	SAR CV	3/3 wins	$\geq$2/3 wins
FACT	52	0.0035	0.0029	1.43%	46/52	48/52
FreDN	52	0.0012	0.0011	0.59%	38/52	42/52
MDMLP-EIA	52	0.0008	0.0006	0.32%	42/52	44/52
VPNet	52	0.0017	0.0016	1.22%	38/52	44/52
XLinear	52	0.0019	0.0018	1.70%	40/52	43/52
SimpleTM	52	0.0017	0.0015	0.89%	44/52	45/52
iTransformer	52	0.0016	0.0015	0.94%	43/52	44/52
PatchTST	52	0.0019	0.0016	0.88%	39/52	41/52
All	**416**	**0.0016**	**0.0015**	**0.82%**	**330/416**	**351/416**

The result files contain 2496 measured MSE/MAE standard deviations for all 416 tasks; no missing value is filled in. Across 1248 paired Base/SAR-LCRT comparisons, MSE improves, stays equal, or worsens in 1048/25/175 cases. SAR-LCRT improves in all three seeds for 330 tasks and in at

least two seeds for 351 tasks. These counts show how often the same improvement repeats across the three training runs; they are not a formal statistical test.

The repeated wins in Table 8 show that many improvements appear in all three runs, not just in one lucky seed.

Table 9: **Seed variation by dataset.** Each dataset contains 32 model–prediction-length tasks. Base/Pure LCRT/SAR medians summarize the sample MSE standard deviations calculated within each task. SAR P90 is their 90th percentile. The last columns give median SAR-LCRT MAE SD and MSE CV ($100\,\mathrm{SD}/\mathrm{mean}$). Values summarize seed variation, not differences between task means.

Dataset	Tasks	MSE SD				SAR-LCRT	
		Base med.	Pure LCRT med.	SAR med.	SAR P90	MAE SD med.	MSE CV
ETTm1	32	0.0028	0.0028	0.0029	0.0148	0.0022	0.67%
ETTm2	32	0.0032	0.0036	0.0035	0.0191	0.0026	1.18%
ETTh1	32	0.0053	0.0054	0.0048	0.0314	0.0024	1.02%
ETTh2	32	0.0073	0.0072	0.0073	0.0187	0.0036	1.70%
ECL	32	0.0006	0.0006	0.0007	0.0033	0.0011	0.38%
Weather	32	0.0017	0.0019	0.0021	0.0068	0.0022	0.77%
PEMS03	32	0.0007	0.0008	0.0006	0.0023	0.0007	0.59%
PEMS04	32	0.0006	0.0007	0.0006	0.0028	0.0006	0.53%
PEMS07	32	0.0009	0.0009	0.0008	0.0034	0.0009	0.82%
PEMS08	32	0.0014	0.0010	0.0010	0.0032	0.0010	0.78%
PEMS-BAY	32	0.0038	0.0037	0.0033	0.0136	0.0013	0.83%
LargeST-SD	32	0.0005	0.0005	0.0005	0.0015	0.0007	1.10%
LargeST-GBA	32	0.0006	0.0005	0.0005	0.0030	0.0008	0.92%

Table 9 summarizes seed variation for each dataset. Because different datasets have very different error scales, a raw standard deviation is hard to compare across them; we therefore also report the coefficient of variation, which divides the standard deviation by the mean. The median SAR-LCRT MSE coefficient of variation stays below 2% on every dataset. ETTh1 has the largest 90th-percentile MSE standard deviation, meaning a few of its tasks vary more across seeds. The released CSV gives every task's measured value.

B.2 Means and standard deviations for MDMLP-EIA

Table 10: **MDMLP-EIA results for each prediction length.** Each value is the mean ± sample SD of seeds 2025/2026/2027; no SD is estimated from other tasks. MSE Gain is the percentage reduction from Base, with higher values better; ↓ marks a reduction and ↑ an increase. MDMLP-EIA is shown because it is also used for supplementary decision-log checks, not because it has the smallest variation. All test forecasts are scored here; Table 16 scores only forecasts whose targets were complete by the final decision. Table 23 gives the remaining MAE means and SDs.

Dataset	H	MSE mean ± SD			MAE mean ± SD		MSE gain
		Base	Pure LCRT	SAR-LCRT	Base	SAR-LCRT	
ETTm1	96	0.321 ± 0.0014	0.317 ± 0.0016	0.317 ± 0.0013	0.360 ± 0.0004	0.359 ± 0.0005	↓1.27%
	192	0.367 ± 0.0009	0.362 ± 0.0009	0.363 ± 0.0009	0.385 ± 0.0007	0.383 ± 0.0006	↓0.97%
	336	0.399 ± 0.0017	0.395 ± 0.0008	0.398 ± 0.0010	0.406 ± 0.0011	0.407 ± 0.0007	↓0.20%
	720	0.465 ± 0.0018	0.465 ± 0.0018	0.465 ± 0.0016	0.443 ± 0.0003	0.443 ± 0.0008	↓0.17%
ETTm2	96	0.175 ± 0.0008	0.174 ± 0.0006	0.173 ± 0.0020	0.258 ± 0.0006	0.258 ± 0.0009	↓0.80%
	192	0.240 ± 0.0008	0.237 ± 0.0013	0.237 ± 0.0014	0.302 ± 0.0009	0.302 ± 0.0013	↓1.57%
	336	0.298 ± 0.0022	0.297 ± 0.0021	0.295 ± 0.0022	0.339 ± 0.0014	0.338 ± 0.0015	↓1.00%
	720	0.399 ± 0.0019	0.398 ± 0.0035	0.395 ± 0.0046	0.397 ± 0.0025	0.397 ± 0.0028	↓0.86%
ETTh1	96	0.375 ± 0.0008	0.375 ± 0.0008	0.375 ± 0.0008	0.396 ± 0.0003	0.396 ± 0.0003	↑0.06%
	192	0.433 ± 0.0012	0.432 ± 0.0012	0.432 ± 0.0012	0.426 ± 0.0011	0.426 ± 0.0010	↓0.30%
	336	0.479 ± 0.0030	0.484 ± 0.0030	0.505 ± 0.0024	0.449 ± 0.0008	0.471 ± 0.0005	↑5.39%
	720	0.498 ± 0.0123	0.498 ± 0.0123	0.498 ± 0.0123	0.479 ± 0.0068	0.479 ± 0.0068	0.00%

Continued on next page

Table 10 – *continued*

Dataset	H	MSE mean ± SD			MAE mean ± SD		MSE gain
		Base	Pure LCRT	SAR-LCRT	Base	SAR-LCRT	
ETTh2	96	0.285 ± 0.0029	0.284 ± 0.0029	0.284 ± 0.0029	0.336 ± 0.0008	0.336 ± 0.0008	↓0.06%
	192	0.369 ± 0.0008	0.369 ± 0.0008	0.369 ± 0.0007	0.390 ± 0.0006	0.390 ± 0.0006	↓0.04%
	336	0.416 ± 0.0070	0.416 ± 0.0067	0.416 ± 0.0067	0.427 ± 0.0034	0.427 ± 0.0031	↓0.10%
	720	0.434 ± 0.0012	0.435 ± 0.0016	0.434 ± 0.0012	0.448 ± 0.0010	0.448 ± 0.0011	↓0.03%
ECL	96	0.159 ± 0.0003	0.159 ± 0.0003	0.159 ± 0.0003	0.250 ± 0.0002	0.251 ± 0.0007	↑0.06%
	192	0.170 ± 0.0001	0.170 ± 0.0001	0.170 ± 0.0001	0.259 ± 0.0001	0.260 ± 0.0001	↑0.23%
	336	0.187 ± 0.0006	0.187 ± 0.0006	0.187 ± 0.0006	0.277 ± 0.0006	0.277 ± 0.0006	0.00%
	720	0.231 ± 0.0027	0.231 ± 0.0027	0.231 ± 0.0027	0.313 ± 0.0017	0.313 ± 0.0017	↑0.02%
Weather	96	0.164 ± 0.0009	0.161 ± 0.0008	0.161 ± 0.0007	0.208 ± 0.0013	0.220 ± 0.0013	↓1.48%
	192	0.210 ± 0.0012	0.208 ± 0.0009	0.208 ± 0.0002	0.250 ± 0.0006	0.270 ± 0.0002	↓1.06%
	336	0.266 ± 0.0010	0.273 ± 0.0018	0.254 ± 0.0038	0.291 ± 0.0004	0.297 ± 0.0058	↓4.44%
	720	0.344 ± 0.0002	0.343 ± 0.0001	0.344 ± 0.0002	0.343 ± 0.0006	0.343 ± 0.0006	↓0.17%
PEMS03	12	0.077 ± 0.0001	0.074 ± 0.0001	0.073 ± 0.0001	0.185 ± 0.0001	0.181 ± 0.0001	↓5.16%
	24	0.122 ± 0.0005	0.112 ± 0.0008	0.110 ± 0.0004	0.233 ± 0.0005	0.223 ± 0.0006	↓9.94%
	36	0.172 ± 0.0001	0.154 ± 0.0009	0.148 ± 0.0005	0.278 ± 0.0002	0.260 ± 0.0005	↓14.08%
	48	0.225 ± 0.0009	0.197 ± 0.0013	0.186 ± 0.0007	0.318 ± 0.0007	0.293 ± 0.0005	↓17.11%
PEMS04	12	0.095 ± 0.0003	0.092 ± 0.0004	0.091 ± 0.0001	0.203 ± 0.0005	0.200 ± 0.0002	↓3.91%
	24	0.148 ± 0.0005	0.138 ± 0.0008	0.134 ± 0.0008	0.257 ± 0.0004	0.247 ± 0.0007	↓9.68%
	36	0.208 ± 0.0008	0.185 ± 0.0002	0.176 ± 0.0003	0.308 ± 0.0013	0.288 ± 0.0004	↓15.17%
	48	0.268 ± 0.0008	0.229 ± 0.0010	0.219 ± 0.0003	0.354 ± 0.0009	0.324 ± 0.0002	↓18.50%
PEMS07	12	0.071 ± 0.0003	0.070 ± 0.0002	0.070 ± 0.0002	0.172 ± 0.0004	0.172 ± 0.0003	↓1.48%
	24	0.116 ± 0.0002	0.112 ± 0.0003	0.112 ± 0.0003	0.221 ± 0.0004	0.219 ± 0.0004	↓3.08%
	36	0.164 ± 0.0011	0.156 ± 0.0003	0.156 ± 0.0003	0.265 ± 0.0017	0.260 ± 0.0009	↓5.04%
	48	0.208 ± 0.0005	0.195 ± 0.0006	0.195 ± 0.0006	0.300 ± 0.0004	0.293 ± 0.0005	↓6.28%
PEMS08	12	0.090 ± 0.0004	0.086 ± 0.0004	0.084 ± 0.0003	0.193 ± 0.0004	0.188 ± 0.0005	↓6.87%
	24	0.146 ± 0.0002	0.133 ± 0.0003	0.126 ± 0.0001	0.247 ± 0.0002	0.233 ± 0.0001	↓13.16%
	36	0.211 ± 0.0013	0.185 ± 0.0009	0.171 ± 0.0006	0.298 ± 0.0008	0.273 ± 0.0006	↓19.07%
	48	0.279 ± 0.0019	0.241 ± 0.0050	0.210 ± 0.0016	0.344 ± 0.0011	0.307 ± 0.0007	↓24.57%
PEMS-BAY	12	0.319 ± 0.0026	0.315 ± 0.0030	0.314 ± 0.0024	0.262 ± 0.0017	0.262 ± 0.0013	↓1.62%
	24	0.488 ± 0.0003	0.480 ± 0.0006	0.478 ± 0.0000	0.338 ± 0.0001	0.341 ± 0.0002	↓2.21%
	36	0.623 ± 0.0009	0.599 ± 0.0022	0.596 ± 0.0014	0.396 ± 0.0005	0.394 ± 0.0007	↓4.32%
	48	0.733 ± 0.0018	0.698 ± 0.0002	0.702 ± 0.0008	0.442 ± 0.0009	0.436 ± 0.0004	↓4.26%
LargeST-SD	12	0.033 ± 0.0001	0.033 ± 0.0001	0.032 ± 0.0001	0.108 ± 0.0001	0.109 ± 0.0003	↓1.22%
	24	0.046 ± 0.0001	0.046 ± 0.0001	0.045 ± 0.0001	0.128 ± 0.0002	0.128 ± 0.0002	↓2.13%
	36	0.053 ± 0.0001	0.054 ± 0.0001	0.053 ± 0.0001	0.138 ± 0.0001	0.141 ± 0.0002	↓0.68%
	48	0.060 ± 0.0001	0.063 ± 0.0028	0.060 ± 0.0004	0.146 ± 0.0002	0.149 ± 0.0008	↑0.38%
LargeST-GBA	12	0.047 ± 0.0003	0.047 ± 0.0003	0.047 ± 0.0003	0.132 ± 0.0004	0.132 ± 0.0005	↓0.44%
	24	0.064 ± 0.0002	0.064 ± 0.0002	0.064 ± 0.0002	0.154 ± 0.0003	0.154 ± 0.0002	↓0.45%
	36	0.075 ± 0.0003	0.074 ± 0.0003	0.074 ± 0.0003	0.166 ± 0.0004	0.166 ± 0.0004	↓0.86%
	48	0.083 ± 0.0001	0.082 ± 0.0001	0.082 ± 0.0001	0.174 ± 0.0001	0.173 ± 0.0001	↓1.12%

Table 10 reports each MDMLP-EIA task separately. Each mean and standard deviation comes from three runs of the same model, dataset, and prediction length. Combining different architectures would mix differences between models with variation between seeds. The table includes all 13 datasets and four prediction lengths. Tables 8 and 9 summarize all backbones; the released CSV gives the complete task-level values.

B.3 Complete results for every prediction length

Tables 11 and 12 give all $8 \times 13 \times 4 = 416$ task means. Each task has adjacent Base, Pure LCRT, and SAR-LCRT rows. Thin lines separate prediction lengths and heavier lines separate datasets.

Table 11: **Complete Base/Pure LCRT/SAR-LCRT results (I).** Three-seed mean errors at each horizon H; input and horizon settings follow Table 1. For SAR-LCRT versus Base at three decimals, red bold marks lower errors; underlined red bold marks equal values.

Dataset	H	Variant	FACT		FreDN		MDMLP-EIA		VPNet	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	Base	0.328	0.367	0.320	0.357	0.321	0.360	0.343	0.370
		Pure LCRT	0.323	0.365	0.316	0.356	0.317	0.359	0.340	0.369
		SAR-LCRT	**0.324**	**0.366**	**0.316**	**0.356**	**0.317**	**0.359**	**0.339**	**0.370**

Continued on next page

Table 11 – *continued*

Dataset	H	Variant	FACT		FreDN		MDMLP-EIA		VPNet	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	192	Base	0.373	0.391	0.370	0.386	0.367	0.385	0.391	0.400
		Pure LCRT	0.368	0.389	0.365	0.384	0.362	0.383	0.389	0.399
		SAR-LCRT	**0.368**	**0.389**	**0.366**	**0.384**	**0.363**	**0.383**	**0.389**	**0.399**
ETTm1	336	Base	0.405	0.411	0.399	0.407	0.399	0.406	0.417	0.421
		Pure LCRT	0.401	0.410	0.396	0.405	0.395	0.405	0.416	0.421
		SAR-LCRT	**0.403**	**0.411**	**0.398**	**0.407**	**0.398**	0.407	**0.416**	**0.421**
ETTm1	720	Base	0.656	0.516	0.457	0.443	0.465	0.443	0.497	0.468
		Pure LCRT	0.656	0.516	0.457	0.443	0.465	0.443	0.497	0.468
		SAR-LCRT	**0.654**	0.519	0.458	**0.443**	**0.465**	**0.443**	**0.497**	**0.468**
ETTm2	96	Base	0.173	0.257	0.175	0.256	0.175	0.258	0.214	0.281
		Pure LCRT	0.173	0.257	0.175	0.256	0.174	0.258	0.215	0.282
		SAR-LCRT	**0.170**	**0.257**	**0.172**	**0.256**	**0.173**	**0.258**	**0.207**	**0.279**
ETTm2	192	Base	0.236	0.298	0.240	0.298	0.240	0.302	0.274	0.326
		Pure LCRT	0.233	0.298	0.237	0.300	0.237	0.302	0.275	0.330
		SAR-LCRT	**0.234**	0.299	**0.237**	0.299	**0.237**	**0.302**	**0.272**	0.328
ETTm2	336	Base	0.300	0.340	0.300	0.339	0.298	0.339	0.350	0.372
		Pure LCRT	0.298	0.339	0.299	0.338	0.297	0.338	0.347	0.372
		SAR-LCRT	**0.297**	**0.339**	**0.298**	**0.338**	**0.295**	**0.338**	**0.347**	**0.372**
ETTm2	720	Base	0.397	0.397	0.399	0.396	0.399	0.397	0.543	0.468
		Pure LCRT	0.394	0.396	0.396	0.395	0.398	0.397	0.543	0.468
		SAR-LCRT	**0.393**	**0.396**	**0.393**	**0.395**	**0.395**	**0.397**	**0.543**	**0.468**
ETTh1	96	Base	0.384	0.402	0.378	0.396	0.375	0.396	0.406	0.418
		Pure LCRT	0.384	0.402	0.378	0.396	0.375	0.396	0.405	0.418
		SAR-LCRT	**0.384**	**0.402**	**0.378**	**0.396**	**0.375**	**0.396**	**0.406**	**0.418**
ETTh1	192	Base	0.444	0.434	0.437	0.429	0.433	0.426	0.510	0.466
		Pure LCRT	0.442	0.433	0.436	0.428	0.432	0.426	0.510	0.466
		SAR-LCRT	**0.442**	**0.434**	**0.436**	**0.429**	**0.432**	**0.426**	**0.510**	**0.466**
ETTh1	336	Base	0.489	0.457	0.480	0.450	0.479	0.449	0.564	0.499
		Pure LCRT	0.494	0.465	0.485	0.458	0.484	0.458	0.574	0.505
		SAR-LCRT	0.515	0.478	0.507	0.473	0.505	0.471	0.571	0.505
ETTh1	720	Base	0.497	0.478	0.475	0.466	0.498	0.479	0.616	0.548
		Pure LCRT	0.497	0.478	0.475	0.466	0.498	0.479	0.616	0.548
		SAR-LCRT	**0.497**	**0.478**	**0.475**	**0.466**	**0.498**	**0.479**	**0.616**	**0.548**
ETTh2	96	Base	0.298	0.348	0.293	0.343	0.285	0.336	0.330	0.370
		Pure LCRT	0.298	0.348	0.293	0.343	0.284	0.336	0.330	0.370
		SAR-LCRT	**0.298**	**0.348**	**0.293**	**0.343**	**0.284**	**0.336**	0.331	**0.370**
ETTh2	192	Base	0.380	0.398	0.365	0.390	0.369	0.390	0.450	0.431
		Pure LCRT	0.379	0.398	0.365	0.390	0.369	0.390	0.450	0.431
		SAR-LCRT	**0.379**	**0.398**	**0.365**	**0.390**	**0.369**	**0.390**	**0.449**	**0.431**
ETTh2	336	Base	0.411	0.423	0.423	0.430	0.416	0.427	0.497	0.464
		Pure LCRT	0.411	0.423	0.423	0.429	0.416	0.427	0.496	0.463
		SAR-LCRT	**0.411**	**0.423**	**0.423**	**0.429**	**0.416**	**0.427**	**0.497**	**0.464**
ETTh2	720	Base	0.430	0.444	0.439	0.449	0.434	0.448	0.498	0.476
		Pure LCRT	0.430	0.444	0.439	0.449	0.435	0.449	0.498	0.476
		SAR-LCRT	**0.430**	0.445	**0.439**	**0.449**	**0.434**	**0.448**	**0.498**	**0.476**
ECL	96	Base	0.213	0.303	0.159	0.249	0.159	0.250	0.143	0.238
		Pure LCRT	0.213	0.307	0.159	0.249	0.159	0.250	0.143	0.238
		SAR-LCRT	**0.212**	0.307	**0.158**	0.250	**0.159**	0.251	**0.143**	**0.237**
ECL	192	Base	0.226	0.316	0.169	0.258	0.170	0.259	0.160	0.253
		Pure LCRT	0.226	0.317	0.169	0.258	0.170	0.259	0.160	0.253
		SAR-LCRT	**0.225**	0.318	**0.169**	**0.258**	**0.170**	0.260	**0.160**	0.254
ECL	336	Base	0.246	0.333	0.184	0.274	0.187	0.277	0.181	0.275
		Pure LCRT	0.246	0.333	0.184	0.274	0.187	0.277	0.181	0.275
		SAR-LCRT	**0.246**	**0.333**	**0.184**	**0.274**	**0.187**	**0.277**	**0.181**	**0.275**
ECL	720	Base	0.288	0.361	0.224	0.308	0.231	0.313	0.219	0.309
		Pure LCRT	0.287	0.361	0.224	0.308	0.231	0.313	0.219	0.309
		SAR-LCRT	**0.288**	**0.361**	**0.224**	**0.308**	**0.231**	**0.313**	**0.219**	**0.309**
Weather	96	Base	0.175	0.222	0.165	0.211	0.164	0.208	0.163	0.210
		Pure LCRT	0.172	0.223	0.163	0.213	0.161	0.211	0.161	0.211
		SAR-LCRT	**0.168**	0.231	**0.161**	0.219	**0.161**	0.220	**0.160**	0.219

Continued on next page

Table 11 – *continued*

Dataset	H	Variant	FACT		FreDN		MDMLP-EIA		VPNet	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	192	Base	0.227	0.265	0.209	0.250	0.210	0.250	0.208	0.252
		Pure LCRT	0.226	0.270	0.207	0.254	0.208	0.256	0.206	0.254
		SAR-LCRT	**0.224**	0.286	**0.203**	0.260	**0.208**	0.270	**0.205**	0.265
Weather	336	Base	0.282	0.304	0.267	0.292	0.266	0.291	0.265	0.294
		Pure LCRT	0.277	0.305	0.272	0.302	0.273	0.302	0.265	0.299
		SAR-LCRT	**0.268**	0.313	**0.256**	0.298	**0.254**	0.297	**0.261**	0.300
Weather	720	Base	0.358	0.354	0.350	0.348	0.344	0.343	0.350	0.348
		Pure LCRT	0.358	0.353	0.349	0.348	0.343	0.343	0.349	0.348
		SAR-LCRT	**0.358**	**0.353**	**0.349**	**0.348**	**0.344**	**0.343**	**0.350**	**0.348**
PEMS03	12	Base	0.088	0.200	0.079	0.187	0.077	0.185	0.065	0.172
		Pure LCRT	0.083	0.196	0.075	0.184	0.074	0.182	0.064	0.170
		SAR-LCRT	**0.082**	**0.195**	**0.075**	**0.184**	**0.073**	**0.181**	**0.063**	**0.167**
PEMS03	24	Base	0.137	0.253	0.121	0.233	0.122	0.233	0.086	0.196
		Pure LCRT	0.122	0.242	0.112	0.225	0.112	0.226	0.085	0.195
		SAR-LCRT	**0.120**	**0.239**	**0.111**	**0.225**	**0.110**	**0.223**	**0.084**	**0.193**
PEMS03	36	Base	0.197	0.308	0.171	0.278	0.172	0.278	0.108	0.221
		Pure LCRT	0.170	0.287	0.154	0.266	0.154	0.266	0.105	0.217
		SAR-LCRT	**0.165**	**0.285**	**0.150**	**0.262**	**0.148**	**0.260**	**0.103**	**0.215**
PEMS03	48	Base	0.260	0.362	0.220	0.316	0.225	0.318	0.130	0.241
		Pure LCRT	0.216	0.331	0.196	0.301	0.197	0.302	0.126	0.238
		SAR-LCRT	**0.211**	**0.328**	**0.188**	**0.295**	**0.186**	**0.293**	**0.123**	**0.236**
PEMS04	12	Base	0.105	0.219	0.094	0.204	0.095	0.203	0.075	0.180
		Pure LCRT	0.098	0.211	0.091	0.200	0.092	0.201	0.075	0.180
		SAR-LCRT	**0.097**	**0.211**	**0.090**	**0.200**	**0.091**	**0.200**	**0.074**	**0.178**
PEMS04	24	Base	0.163	0.277	0.140	0.253	0.148	0.257	0.092	0.201
		Pure LCRT	0.146	0.264	0.132	0.245	0.138	0.250	0.091	0.200
		SAR-LCRT	**0.140**	**0.260**	**0.130**	**0.244**	**0.134**	**0.247**	**0.090**	**0.198**
PEMS04	36	Base	0.230	0.333	0.193	0.299	0.208	0.308	0.108	0.220
		Pure LCRT	0.192	0.308	0.177	0.286	0.185	0.294	0.107	0.219
		SAR-LCRT	**0.183**	**0.302**	**0.172**	**0.284**	**0.176**	**0.288**	**0.106**	**0.218**
PEMS04	48	Base	0.300	0.387	0.246	0.341	0.268	0.354	0.124	0.238
		Pure LCRT	0.250	0.355	0.219	0.323	0.229	0.331	0.123	0.237
		SAR-LCRT	**0.235**	**0.345**	**0.213**	**0.320**	**0.219**	**0.324**	**0.121**	**0.234**
PEMS07	12	Base	0.086	0.197	0.073	0.177	0.071	0.172	0.060	0.160
		Pure LCRT	0.083	0.194	0.072	0.176	0.070	0.172	0.059	0.159
		SAR-LCRT	**0.081**	**0.192**	**0.072**	**0.176**	**0.070**	**0.172**	**0.058**	**0.156**
PEMS07	24	Base	0.140	0.256	0.117	0.226	0.116	0.221	0.075	0.181
		Pure LCRT	0.128	0.247	0.113	0.223	0.112	0.219	0.075	0.180
		SAR-LCRT	**0.121**	**0.240**	**0.114**	**0.223**	**0.112**	**0.219**	**0.073**	**0.177**
PEMS07	36	Base	0.194	0.306	0.166	0.269	0.164	0.265	0.091	0.200
		Pure LCRT	0.174	0.293	0.159	0.266	0.156	0.260	0.091	0.199
		SAR-LCRT	**0.161**	**0.281**	**0.159**	**0.266**	**0.156**	**0.260**	**0.088**	**0.196**
PEMS07	48	Base	0.246	0.350	0.213	0.307	0.208	0.300	0.105	0.215
		Pure LCRT	0.219	0.332	0.202	0.301	0.195	0.293	0.104	0.214
		SAR-LCRT	**0.201**	**0.315**	**0.202**	**0.300**	**0.195**	**0.293**	**0.101**	**0.211**
PEMS08	12	Base	0.096	0.206	0.091	0.198	0.090	0.193	0.074	0.177
		Pure LCRT	0.092	0.201	0.087	0.194	0.086	0.190	0.074	0.176
		SAR-LCRT	**0.089**	**0.198**	**0.085**	**0.192**	**0.084**	**0.188**	**0.072**	**0.173**
PEMS08	24	Base	0.157	0.265	0.145	0.252	0.146	0.247	0.100	0.201
		Pure LCRT	0.142	0.253	0.134	0.244	0.133	0.240	0.098	0.200
		SAR-LCRT	**0.135**	**0.248**	**0.127**	**0.238**	**0.126**	**0.233**	**0.094**	**0.196**
PEMS08	36	Base	0.221	0.319	0.209	0.303	0.211	0.298	0.125	0.222
		Pure LCRT	0.199	0.305	0.185	0.289	0.185	0.284	0.121	0.221
		SAR-LCRT	**0.181**	**0.292**	**0.172**	**0.280**	**0.171**	**0.273**	**0.116**	**0.215**
PEMS08	48	Base	0.290	0.373	0.274	0.348	0.279	0.344	0.153	0.240
		Pure LCRT	0.260	0.356	0.244	0.333	0.241	0.326	0.144	0.239
		SAR-LCRT	**0.231**	**0.337**	**0.212**	**0.314**	**0.210**	**0.307**	**0.139**	**0.234**
PEMS-BAY	12	Base	0.354	0.276	0.318	0.258	0.319	0.262	0.282	0.245
		Pure LCRT	0.346	0.275	0.314	0.259	0.315	0.262	0.281	0.245
		SAR-LCRT	**0.338**	**0.273**	**0.314**	0.259	**0.314**	**0.262**	**0.278**	**0.240**

Continued on next page

Table 11 – *continued*

Dataset	H	Variant	FACT		FreDN		MDMLP-EIA		VPNet	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS-BAY	24	Base	0.488	0.331	0.476	0.328	0.488	0.338	0.393	0.291
		Pure LCRT	0.480	0.334	0.467	0.328	0.480	0.342	0.391	0.291
		SAR-LCRT	**0.471**	**0.331**	**0.463**	**0.327**	**0.478**	0.341	**0.385**	**0.285**
PEMS-BAY	36	Base	0.595	0.376	0.589	0.378	0.623	0.396	0.467	0.320
		Pure LCRT	0.578	0.379	0.574	0.379	0.599	0.395	0.463	0.321
		SAR-LCRT	**0.563**	**0.372**	**0.570**	**0.378**	**0.596**	**0.394**	**0.454**	**0.313**
PEMS-BAY	48	Base	0.687	0.413	0.680	0.420	0.733	0.442	0.525	0.341
		Pure LCRT	0.662	0.414	0.658	0.418	0.698	0.436	0.521	0.343
		SAR-LCRT	**0.649**	**0.407**	**0.656**	**0.417**	**0.702**	**0.436**	**0.510**	**0.335**
LargeST-SD	12	Base	0.045	0.132	0.031	0.107	0.033	0.108	0.027	0.100
		Pure LCRT	0.045	0.136	0.031	0.108	0.033	0.110	0.027	0.099
		SAR-LCRT	**0.044**	0.134	**0.031**	**0.107**	**0.032**	0.109	**0.027**	**0.099**
LargeST-SD	24	Base	0.070	0.168	0.043	0.125	0.046	0.128	0.034	0.110
		Pure LCRT	0.069	0.173	0.043	0.128	0.046	0.131	0.035	0.112
		SAR-LCRT	**0.069**	0.173	**0.043**	0.126	**0.045**	**0.128**	**0.034**	**0.110**
LargeST-SD	36	Base	0.081	0.179	0.050	0.135	0.053	0.138	0.039	0.116
		Pure LCRT	0.079	0.185	0.051	0.139	0.054	0.143	0.039	0.118
		SAR-LCRT	**0.079**	0.185	**0.050**	0.137	**0.053**	0.141	**0.039**	0.117
LargeST-SD	48	Base	0.090	0.191	0.057	0.143	0.060	0.146	0.044	0.122
		Pure LCRT	0.089	0.199	0.058	0.149	0.063	0.156	0.045	0.125
		SAR-LCRT	**0.088**	0.195	**0.057**	0.145	**0.060**	0.149	**0.044**	0.124
LargeST-GBA	12	Base	0.070	0.173	0.046	0.132	0.047	0.132	0.043	0.128
		Pure LCRT	0.065	0.166	0.046	0.132	0.047	0.132	0.043	0.126
		SAR-LCRT	**0.065**	**0.166**	**0.046**	**0.132**	**0.047**	**0.132**	**0.042**	**0.125**
LargeST-GBA	24	Base	0.102	0.207	0.063	0.153	0.064	0.154	0.059	0.146
		Pure LCRT	0.094	0.200	0.063	0.153	0.064	0.154	0.058	0.146
		SAR-LCRT	**0.094**	**0.200**	**0.063**	**0.153**	**0.064**	**0.154**	**0.058**	**0.146**
LargeST-GBA	36	Base	0.116	0.223	0.073	0.164	0.075	0.166	0.067	0.155
		Pure LCRT	0.106	0.214	0.073	0.164	0.074	0.166	0.066	0.154
		SAR-LCRT	**0.107**	**0.215**	**0.072**	**0.164**	**0.074**	**0.166**	**0.066**	**0.154**
LargeST-GBA	48	Base	0.121	0.222	0.082	0.173	0.083	0.174	0.072	0.158
		Pure LCRT	0.112	0.216	0.081	0.172	0.082	0.173	0.072	0.157
		SAR-LCRT	**0.112**	**0.216**	**0.081**	**0.173**	**0.082**	**0.173**	**0.072**	**0.158**

Table 11 covers FACT, FreDN, MDMLP-EIA, and VPNet. Reporting each prediction length makes patterns visible that horizon averages can hide, including the shared ETTh1 difficulty at $H = 336$. The highlighting compares SAR-LCRT only with Base in the same task and metric. Red bold means lower error; red bold with an underline means equal error at three decimals. Base and Pure LCRT are always plain, as are SAR-LCRT values above Base. Highlighting therefore does not mean that SAR-LCRT beats Pure LCRT. Equal displayed errors also do not imply identical predictions or statistical equivalence: rounding and averaging can hide small differences.

Table 12: **Complete Base/Pure LCRT/SAR-LCRT results (II).** Three-seed mean errors at each horizon H; settings and highlighting follow Table 11.

Dataset	H	Variant	XLinear		SimpleTM		iTransformer		PatchTST	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	Base	0.327	0.361	0.320	0.356	0.348	0.376	0.324	0.362
		Pure LCRT	0.322	0.359	0.315	0.354	0.343	0.374	0.320	0.360
		SAR-LCRT	**0.322**	**0.359**	**0.315**	**0.354**	**0.343**	**0.374**	**0.320**	**0.360**
ETTm1	192	Base	0.385	0.392	0.369	0.384	0.393	0.401	0.365	0.387
		Pure LCRT	0.382	0.391	0.365	0.383	0.389	0.399	0.361	0.385
		SAR-LCRT	**0.382**	**0.391**	**0.365**	**0.383**	**0.389**	**0.400**	**0.362**	**0.385**
ETTm1	336	Base	0.420	0.412	0.402	0.408	0.423	0.420	0.395	0.408
		Pure LCRT	0.419	0.412	0.399	0.406	0.419	0.418	0.392	0.406
		SAR-LCRT	0.421	0.413	**0.401**	**0.407**	**0.421**	**0.419**	**0.394**	**0.407**

Continued on next page

Table 12 – *continued*

Dataset	H	Variant	XLinear		SimpleTM		iTransformer		PatchTST	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	720	Base	0.495	0.459	0.466	0.446	0.510	0.467	0.462	0.453
		Pure LCRT	0.495	0.459	0.466	0.446	0.510	0.467	0.462	0.453
		SAR-LCRT	**0.495**	0.460	**0.465**	**0.446**	**0.510**	**0.467**	**0.462**	**0.453**
ETTm2	96	Base	0.172	0.254	0.176	0.258	0.192	0.275	0.186	0.268
		Pure LCRT	0.171	0.254	0.176	0.258	0.190	0.275	0.185	0.268
		SAR-LCRT	**0.169**	**0.254**	**0.174**	**0.258**	**0.187**	**0.274**	**0.183**	**0.268**
ETTm2	192	Base	0.240	0.299	0.245	0.304	0.258	0.316	0.255	0.312
		Pure LCRT	0.237	0.301	0.245	0.307	0.253	0.315	0.252	0.313
		SAR-LCRT	**0.237**	0.300	**0.242**	0.305	**0.254**	**0.316**	**0.251**	0.313
ETTm2	336	Base	0.298	0.336	0.304	0.342	0.322	0.356	0.314	0.349
		Pure LCRT	0.297	0.336	0.302	0.342	0.320	0.355	0.312	0.349
		SAR-LCRT	**0.296**	**0.336**	**0.301**	**0.342**	**0.319**	**0.355**	**0.312**	**0.349**
ETTm2	720	Base	0.412	0.402	0.404	0.400	0.424	0.412	0.421	0.412
		Pure LCRT	0.409	0.401	0.403	0.400	0.421	0.411	0.416	0.410
		SAR-LCRT	**0.407**	**0.401**	**0.402**	**0.400**	**0.421**	**0.411**	**0.416**	**0.410**
ETTh1	96	Base	0.385	0.398	0.380	0.400	0.407	0.419	0.395	0.408
		Pure LCRT	0.385	0.398	0.380	0.400	0.407	0.419	0.395	0.407
		SAR-LCRT	**0.385**	**0.398**	**0.380**	**0.400**	**0.407**	**0.419**	0.396	**0.407**
ETTh1	192	Base	0.447	0.431	0.437	0.435	0.457	0.449	0.462	0.448
		Pure LCRT	0.446	0.430	0.435	0.434	0.456	0.448	0.461	0.447
		SAR-LCRT	**0.446**	**0.431**	**0.436**	**0.435**	**0.456**	**0.449**	**0.461**	**0.447**
ETTh1	336	Base	0.480	0.444	0.479	0.457	0.504	0.472	0.510	0.473
		Pure LCRT	0.485	0.453	0.479	0.457	0.513	0.480	0.521	0.484
		SAR-LCRT	0.506	0.467	0.509	0.480	0.514	0.481	0.516	0.479
ETTh1	720	Base	0.599	0.531	0.486	0.478	0.539	0.505	0.510	0.493
		Pure LCRT	0.599	0.531	0.486	0.478	0.543	0.508	0.510	0.493
		SAR-LCRT	**0.599**	**0.531**	**0.486**	**0.478**	0.542	0.507	**0.510**	**0.492**
ETTh2	96	Base	0.314	0.351	0.297	0.344	0.317	0.357	0.300	0.347
		Pure LCRT	0.314	0.350	0.297	0.344	0.317	0.356	0.300	0.347
		SAR-LCRT	**0.314**	**0.350**	**0.297**	**0.344**	**0.317**	**0.357**	**0.300**	**0.347**
ETTh2	192	Base	0.371	0.391	0.370	0.392	0.411	0.415	0.380	0.400
		Pure LCRT	0.371	0.391	0.370	0.392	0.411	0.415	0.380	0.400
		SAR-LCRT	**0.371**	**0.391**	**0.370**	**0.392**	**0.411**	**0.415**	**0.380**	**0.400**
ETTh2	336	Base	0.415	0.426	0.438	0.434	0.454	0.447	0.428	0.437
		Pure LCRT	0.414	0.426	0.438	0.434	0.454	0.447	0.427	0.437
		SAR-LCRT	**0.414**	**0.426**	**0.438**	**0.434**	**0.454**	**0.447**	**0.427**	**0.437**
ETTh2	720	Base	0.444	0.448	0.445	0.451	0.427	0.446	0.439	0.452
		Pure LCRT	0.444	0.448	0.445	0.451	0.427	0.446	0.439	0.452
		SAR-LCRT	**0.444**	**0.448**	**0.445**	**0.451**	**0.427**	**0.446**	**0.439**	**0.452**
ECL	96	Base	0.156	0.254	0.138	0.233	0.149	0.241	0.164	0.250
		Pure LCRT	0.156	0.254	0.138	0.233	0.149	0.241	0.164	0.250
		SAR-LCRT	**0.156**	**0.254**	**0.138**	**0.233**	**0.149**	**0.241**	**0.164**	0.251
ECL	192	Base	0.166	0.263	0.155	0.248	0.165	0.257	0.173	0.259
		Pure LCRT	0.166	0.263	0.155	0.248	0.165	0.257	0.173	0.260
		SAR-LCRT	**0.166**	**0.263**	0.156	0.249	**0.165**	**0.257**	**0.173**	0.260
ECL	336	Base	0.184	0.280	0.170	0.264	0.181	0.273	0.189	0.276
		Pure LCRT	0.184	0.280	0.170	0.264	0.181	0.273	0.189	0.276
		SAR-LCRT	**0.184**	**0.280**	**0.170**	**0.264**	**0.181**	**0.273**	**0.189**	**0.276**
ECL	720	Base	0.224	0.315	0.200	0.293	0.226	0.311	0.230	0.311
		Pure LCRT	0.224	0.314	0.200	0.293	0.226	0.311	0.230	0.311
		SAR-LCRT	**0.224**	**0.315**	**0.200**	**0.293**	**0.226**	**0.311**	**0.230**	**0.311**
Weather	96	Base	0.156	0.202	0.159	0.203	0.175	0.214	0.176	0.218
		Pure LCRT	0.156	0.203	0.159	0.204	0.171	0.216	0.173	0.220
		SAR-LCRT	**0.155**	0.211	**0.156**	0.213	**0.169**	0.224	**0.171**	0.232
Weather	192	Base	0.206	0.249	0.206	0.248	0.225	0.258	0.219	0.256
		Pure LCRT	0.203	0.252	0.204	0.251	0.222	0.265	0.217	0.262
		SAR-LCRT	**0.202**	0.260	**0.201**	0.257	**0.222**	0.281	**0.216**	0.276
Weather	336	Base	0.266	0.294	0.264	0.292	0.281	0.299	0.276	0.297
		Pure LCRT	0.264	0.300	0.271	0.303	0.277	0.302	0.276	0.305
		SAR-LCRT	**0.256**	0.296	**0.256**	0.297	**0.270**	0.314	**0.268**	0.316

Continued on next page

Table 12 – *continued*

Dataset	H	Variant	XLinear		SimpleTM		iTransformer		PatchTST	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	720	Base	0.346	0.347	0.347	0.346	0.358	0.350	0.353	0.347
		Pure LCRT	0.345	0.347	0.346	0.346	0.358	0.349	0.353	0.346
		SAR-LCRT	**0.345**	**0.347**	**0.346**	**0.346**	**0.358**	**0.349**	**0.353**	**0.346**
PEMS03	12	Base	0.065	0.167	0.062	0.166	0.068	0.173	0.078	0.185
		Pure LCRT	0.065	0.167	0.062	0.165	0.067	0.172	0.075	0.182
		SAR-LCRT	**0.063**	**0.166**	**0.061**	**0.164**	**0.066**	**0.171**	**0.074**	**0.182**
PEMS03	24	Base	0.088	0.196	0.082	0.190	0.094	0.204	0.123	0.233
		Pure LCRT	0.087	0.195	0.082	0.189	0.092	0.201	0.113	0.226
		SAR-LCRT	**0.085**	**0.193**	**0.080**	**0.187**	**0.090**	**0.200**	**0.112**	**0.224**
PEMS03	36	Base	0.113	0.222	0.106	0.216	0.122	0.233	0.171	0.276
		Pure LCRT	0.111	0.220	0.105	0.215	0.118	0.229	0.152	0.263
		SAR-LCRT	**0.109**	**0.218**	**0.102**	**0.211**	**0.114**	**0.226**	**0.150**	**0.261**
PEMS03	48	Base	0.142	0.245	0.127	0.237	0.152	0.261	0.219	0.316
		Pure LCRT	0.138	0.241	0.125	0.235	0.145	0.256	0.191	0.297
		SAR-LCRT	**0.136**	**0.240**	**0.120**	**0.230**	**0.140**	**0.252**	**0.187**	**0.294**
PEMS04	12	Base	0.072	0.177	0.075	0.180	0.086	0.190	0.100	0.209
		Pure LCRT	0.072	0.176	0.074	0.179	0.084	0.189	0.095	0.204
		SAR-LCRT	**0.071**	**0.175**	**0.073**	**0.178**	**0.082**	**0.187**	**0.095**	**0.204**
PEMS04	24	Base	0.092	0.202	0.094	0.204	0.116	0.223	0.159	0.266
		Pure LCRT	0.091	0.202	0.093	0.203	0.111	0.221	0.144	0.255
		SAR-LCRT	**0.090**	**0.200**	**0.091**	**0.201**	**0.108**	**0.217**	**0.142**	**0.254**
PEMS04	36	Base	0.109	0.223	0.112	0.226	0.147	0.254	0.221	0.317
		Pure LCRT	0.109	0.222	0.111	0.224	0.139	0.249	0.192	0.298
		SAR-LCRT	**0.107**	**0.220**	**0.108**	**0.221**	**0.134**	**0.243**	**0.188**	**0.296**
PEMS04	48	Base	0.125	0.241	0.127	0.242	0.179	0.281	0.288	0.367
		Pure LCRT	0.125	0.241	0.126	0.240	0.165	0.274	0.237	0.334
		SAR-LCRT	**0.121**	**0.237**	**0.122**	**0.236**	**0.158**	**0.267**	**0.234**	**0.333**
PEMS07	12	Base	0.059	0.155	0.060	0.156	0.064	0.162	0.078	0.185
		Pure LCRT	0.059	0.155	0.059	0.156	0.063	0.161	0.075	0.182
		SAR-LCRT	**0.059**	**0.155**	**0.059**	**0.155**	**0.062**	**0.160**	**0.075**	**0.181**
PEMS07	24	Base	0.082	0.182	0.079	0.178	0.091	0.193	0.137	0.247
		Pure LCRT	0.082	0.182	0.078	0.178	0.088	0.191	0.125	0.239
		SAR-LCRT	**0.081**	**0.181**	**0.077**	**0.175**	**0.085**	**0.188**	**0.123**	**0.237**
PEMS07	36	Base	0.103	0.203	0.097	0.197	0.117	0.221	0.195	0.294
		Pure LCRT	0.103	0.203	0.096	0.197	0.112	0.218	0.168	0.281
		SAR-LCRT	**0.102**	**0.202**	**0.094**	**0.194**	**0.107**	**0.213**	**0.168**	**0.281**
PEMS07	48	Base	0.120	0.222	0.113	0.215	0.139	0.244	0.270	0.351
		Pure LCRT	0.119	0.223	0.113	0.215	0.133	0.240	0.214	0.320
		SAR-LCRT	**0.117**	**0.221**	**0.109**	**0.210**	**0.125**	**0.233**	**0.214**	**0.320**
PEMS08	12	Base	0.078	0.181	0.073	0.173	0.079	0.178	0.090	0.194
		Pure LCRT	0.076	0.179	0.071	0.172	0.078	0.177	0.087	0.191
		SAR-LCRT	**0.073**	**0.175**	**0.069**	**0.170**	**0.076**	**0.175**	**0.085**	**0.190**
PEMS08	24	Base	0.113	0.218	0.099	0.200	0.115	0.215	0.140	0.243
		Pure LCRT	0.108	0.214	0.096	0.198	0.111	0.212	0.133	0.238
		SAR-LCRT	**0.100**	**0.207**	**0.092**	**0.195**	**0.107**	**0.208**	**0.128**	**0.234**
PEMS08	36	Base	0.145	0.245	0.125	0.224	0.147	0.242	0.193	0.285
		Pure LCRT	0.137	0.242	0.119	0.221	0.141	0.239	0.181	0.278
		SAR-LCRT	**0.127**	**0.232**	**0.113**	**0.217**	**0.134**	**0.234**	**0.171**	**0.272**
PEMS08	48	Base	0.174	0.267	0.151	0.246	0.179	0.267	0.248	0.326
		Pure LCRT	0.164	0.264	0.141	0.243	0.171	0.263	0.230	0.316
		SAR-LCRT	**0.148**	**0.250**	**0.133**	**0.236**	**0.161**	**0.256**	**0.215**	**0.306**
PEMS-BAY	12	Base	0.301	0.248	0.271	0.237	0.299	0.248	0.320	0.260
		Pure LCRT	0.301	0.248	0.271	0.237	0.295	0.249	0.311	0.259
		SAR-LCRT	**0.294**	**0.246**	**0.267**	**0.234**	**0.291**	**0.247**	**0.309**	**0.260**
PEMS-BAY	24	Base	0.434	0.295	0.381	0.282	0.448	0.306	0.510	0.339
		Pure LCRT	0.433	0.295	0.380	0.282	0.440	0.309	0.482	0.343
		SAR-LCRT	**0.427**	**0.294**	**0.374**	**0.278**	**0.432**	**0.306**	**0.480**	0.342
PEMS-BAY	36	Base	0.553	0.329	0.458	0.314	0.563	0.348	0.659	0.397
		Pure LCRT	0.546	0.329	0.456	0.315	0.545	0.351	0.620	0.398
		SAR-LCRT	**0.537**	**0.325**	**0.446**	**0.308**	**0.534**	**0.346**	**0.621**	0.398

Continued on next page

Table 12 – *continued*

Dataset	H	Variant	XLinear		SimpleTM		iTransformer		PatchTST	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS-BAY	48	Base	0.642	0.358	0.531	0.339	0.649	0.379	0.766	0.436
		Pure LCRT	0.638	0.361	0.529	0.339	0.621	0.383	0.702	0.431
		SAR-LCRT	**0.623**	**0.355**	**0.514**	**0.329**	**0.621**	**0.379**	**0.704**	**0.432**
LargeST-SD	12	Base	0.023	0.090	0.025	0.094	0.028	0.099	0.036	0.114
		Pure LCRT	0.023	0.091	0.024	0.094	0.027	0.099	0.037	0.119
		SAR-LCRT	**0.023**	**0.090**	**0.024**	**0.093**	**0.027**	**0.097**	**0.036**	0.117
LargeST-SD	24	Base	0.031	0.105	0.032	0.107	0.034	0.109	0.052	0.137
		Pure LCRT	0.031	0.106	0.032	0.106	0.034	0.109	0.053	0.143
		SAR-LCRT	**0.031**	**0.105**	**0.032**	**0.105**	**0.033**	**0.108**	**0.052**	0.141
LargeST-SD	36	Base	0.038	0.116	0.036	0.113	0.038	0.116	0.061	0.147
		Pure LCRT	0.037	0.115	0.036	0.113	0.038	0.116	0.061	0.156
		SAR-LCRT	**0.036**	**0.114**	**0.036**	**0.112**	**0.037**	**0.115**	**0.061**	0.153
LargeST-SD	48	Base	0.042	0.124	0.042	0.120	0.042	0.121	0.071	0.158
		Pure LCRT	0.042	0.126	0.043	0.120	0.043	0.125	0.071	0.169
		SAR-LCRT	**0.042**	**0.124**	**0.042**	**0.119**	**0.041**	**0.121**	**0.070**	0.166
LargeST-GBA	12	Base	0.041	0.123	0.040	0.122	0.041	0.120	0.073	0.178
		Pure LCRT	0.041	0.122	0.040	0.121	0.040	0.120	0.064	0.166
		SAR-LCRT	**0.040**	**0.122**	**0.040**	**0.120**	**0.040**	**0.120**	**0.064**	**0.166**
LargeST-GBA	24	Base	0.053	0.141	0.050	0.135	0.052	0.136	0.102	0.206
		Pure LCRT	0.053	0.141	0.050	0.135	0.051	0.135	0.090	0.195
		SAR-LCRT	**0.053**	**0.140**	**0.050**	**0.134**	**0.051**	**0.135**	**0.090**	**0.195**
LargeST-GBA	36	Base	0.061	0.151	0.057	0.143	0.059	0.144	0.115	0.215
		Pure LCRT	0.061	0.150	0.057	0.143	0.058	0.144	0.103	0.206
		SAR-LCRT	**0.060**	**0.150**	**0.057**	**0.142**	**0.058**	**0.144**	**0.103**	**0.206**
LargeST-GBA	48	Base	0.067	0.157	0.063	0.149	0.064	0.149	0.121	0.221
		Pure LCRT	0.067	0.157	0.063	0.148	0.063	0.149	0.109	0.211
		SAR-LCRT	**0.066**	**0.156**	**0.062**	**0.148**	**0.063**	**0.149**	**0.109**	**0.211**

Table 12 reports XLinear, SimpleTM, iTransformer, and PatchTST. These models show the same broad pattern: traffic tasks improve more often, while ETTh1 remains difficult. Together, the two tables retain all 416 tasks, including results above Base error.

C Residual Baselines, Decision Logs, and Runtime

C.1 Additional residual baselines

These extra comparisons check whether a few simple residual corrections give the same benefit under the same delayed-feedback setup. They add context for the main framework comparison in Table 3, but use a 96-task subset rather than the full 416-task ablation. Tables 13 and 14 separate "reproducing a known setting" from "tuning it on validation", and Table 15 shows the subset's results are not driven by one backbone.

The residual-baseline comparison uses FACT, FreDN, MDMLP-EIA, and SimpleTM on ETTh1, ETTm1, Weather, PEMS03, PEMS08, and LargeST-GBA. Four prediction lengths per pair give 96 tasks, each with three seeds. All methods share the same frozen checkpoint, test data, and complete-target feedback schedule. Here a win or loss is only counted when the Base-relative MSE gain differs by more than ± 0.1 percentage points; smaller differences are treated as equal (the main 416-task analysis uses the tighter 10^{-9} tolerance).

Reproducing and tuning the residual baselines. We repeated 18 checks of the original Mean, EWMA, and Ridge runs and recovered the archived results. The original EWMA uses update weight 0.05, with a half-life of about 13.5 completed forecasts (older forecasts lose half their influence after that many arrivals), and was not tuned on validation data. Table 13 reproduces that configuration and should not be read as the best possible EWMA setting.

We then choose Mean/EWMA/Ridge settings by validation MSE, predicting each validation point before using its available feedback for later updates. This is possible for 252 of 288 task–seed runs; the other 36 runs use a preset default because too little validation history is available. All three

Table 13: **Reproduction of the original residual baselines.** Values are average percentage error reductions from Base over 96 tasks. The 18 repeated checks cover three methods, two tasks, and three seeds. ↓ marks a reduction and ↑ an increase.

Method	MSE gain	MAE gain	W/E/L
Online EWMA	↑9.46%	↑8.20%	0/0/96
Online Mean Residual	↑0.73%	↑0.73%	21/52/23
Online Ridge	↑1.41%	↑4.94%	67/1/28

methods still have higher average error than Base (Table 14). Ridge improves 70 tasks, but larger errors on some remaining tasks outweigh those gains. Tuned EWMA has higher MSE on all 96 tasks, with especially large errors on ETTh1 and Weather. This result concerns the tested settings, not every possible EWMA or Ridge design. The export includes individual seeds, 16,632 validation trials, all 288 parameter selections, configurations, and source hashes (Appendix E).

Table 14: **Always-on residual corrections with validation-selected settings.** Values are average percentage error reductions from Base. Worst is the least favorable task's MSE reduction. Settings are chosen before testing using sequential validation predictions when enough history is available. ↓ marks a reduction and ↑ an increase.

Method	MSE gain	MAE gain	W/E/L	Worst MSE gain
Validated Mean	↑1.95%	↑1.88%	41/6/49	↑26.58%
Validated EWMA	↑16.47%	↑8.40%	0/0/96	↑237.25%
Validated Ridge	↑0.98%	↑3.43%	70/3/23	↑75.79%

Table 15: **Same-protocol online methods by backbone.** Improvements are task-macro percentages over paired Base; each backbone contains 24 task means. Every online method receives only fully matured feedback. Red bold marks the largest average gain within each backbone. Method(Act.) is the percentage of forecasts using the method's residual estimate. ↓ marks a reduction and ↑ an increase.

Backbone	Method	MSE gain	MAE gain	W/E/L	Method(Act.)
FACT	Base	0.00%	0.00%	0/24/0	–
	Online Mean Residual	↑0.37%	↑0.45%	11/8/5	100.00%
	Online EWMA	↑9.91%	↑8.06%	0/0/24	100.00%
	Online Ridge	↓0.05%	↑3.91%	17/0/7	100.00%
	Pure LCRT	↓5.03%	**↓2.05%**	20/3/1	100.00%
	SAR-LCRT	**↓6.39%**	↓1.99%	21/2/1	90.03%
FreDN	Base	0.00%	0.00%	0/24/0	–
	Online Mean Residual	↑0.77%	↑0.79%	3/14/7	100.00%
	Online EWMA	↑10.03%	↑8.82%	0/0/24	100.00%
	Online Ridge	↑0.98%	↑4.79%	17/0/7	100.00%
	Pure LCRT	↓3.08%	↓0.96%	19/3/2	100.00%
	SAR-LCRT	**↓4.57%**	**↓1.21%**	20/3/1	89.85%
MDMLP-EIA	Base	0.00%	0.00%	0/24/0	–
	Online Mean Residual	↑0.77%	↑0.77%	2/16/6	100.00%
	Online EWMA	↑9.77%	↑8.71%	0/0/24	100.00%
	Online Ridge	↑0.90%	↑5.17%	15/1/8	100.00%
	Pure LCRT	↓3.27%	↓0.81%	19/3/2	100.00%
	SAR-LCRT	**↓4.89%**	**↓1.16%**	21/2/1	88.61%
SimpleTM	Base	0.00%	0.00%	0/24/0	–
	Online Mean Residual	↑1.03%	↑0.92%	5/14/5	100.00%
	Online EWMA	↑8.13%	↑7.21%	0/0/24	100.00%
	Online Ridge	↑3.84%	↑5.92%	18/0/6	100.00%
	Pure LCRT	↓1.00%	↓0.10%	18/5/1	100.00%
	SAR-LCRT	**↓2.35%**	**↓0.31%**	21/2/1	89.59%

SAR-LCRT lowers average MSE on each of the four backbones, from 2.35% on SimpleTM to 6.39% on FACT. It also has the largest average MSE reduction among the methods in each backbone's comparison. Ridge shows why the number of improved tasks and the average improvement answer

different questions: it helps many tasks, yet raises average MSE on three backbones and raises MAE overall.

C.2 Decision logs and feedback timing

The figures and tables in this subsection are checks on the implementation. They let a reader verify, on recorded runs, that the delayed-feedback rules in Section 3 are followed. Table 16 and Figure 8 connect outcomes to the loss account; Table 17 checks each seed's timing and choices. Figure 9 shows that how often a correction is used does not by itself determine accuracy. These cases do not estimate the whole-grid gains.

Table 16: **MDMLP-EIA results with feedback available by the final decision.** Errors score $T - H$ forecast origins, excluding the final H. On ETTh1/$H = 336$, this uses 2,209 of 2,545 origins, whereas Table 10 uses all 2,545. SAR(Act.) uses all T forecast origins, including those still awaiting feedback. Margin is the final $B_t - D_t$.

Dataset	H	Base MSE	Issued MSE	MSE gain	SAR(Act.)	Margin
ETTh1	336	0.4427	0.4725	↑6.72%	26.4%	-65.7
PEMS08	48	0.2785	0.2102	↓24.50%	99.1%	+234.1

These two runs show a success and a failure of the controller. PEMS08/$H = 48$ ends with a positive risk balance $B_t - D_t = +234.1$ and 24.50% lower MSE, while ETTh1/$H = 336$ ends at -65.7 with 6.72% higher MSE. The balance is updated only when a target arrives, so losses from forecasts that are still waiting for feedback can later push a previously positive balance down.

Read Figure 8 from top to bottom. The top row compares Base and the forecast that was actually issued. The middle row adds up the issued-forecast losses relative to Base (values below zero mean less total error). The bottom row shows $B_t - D_t$, the gain left after the reserve, using only the feedback available at each step; its shaded band shows how many seeds are currently using a correction. The first exit check happens only after H forecasts in the episode have complete targets, and earlier losses can still arrive afterward, so a briefly negative balance is not a rule violation.

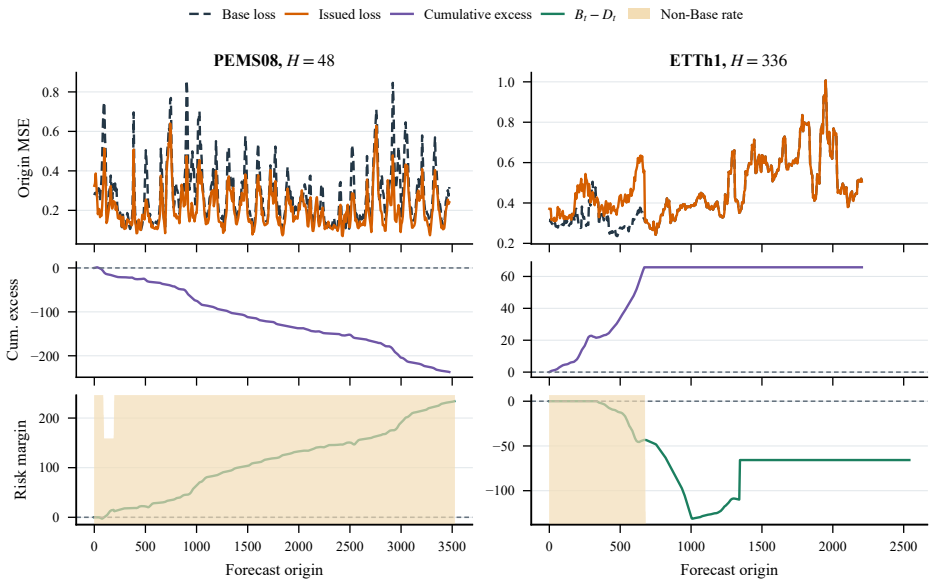


Figure 8: **Realized risk in a success and a failure case.** MDMLP-EIA on PEMS08/$H = 48$ gains 24.50% and ends with positive risk margin 234.1; ETTh1/$H = 336$ has a -6.72% paired gain and ends at -65.7. Base/issued losses, cumulative excess, the intervention region, and $B_t - D_t$ show the delayed fallback. Curves average the three seed traces at each aligned origin.

Table 17: **Per-seed timing check for the MDMLP-EIA logs.** Origins counts all forecasts; Matured counts those with complete targets by the final decision. SAR(Act.) is the percentage of all Origins using a residual candidate, whereas MSE gain scores the Matured subset. Entry/Exit count transitions into/out of correction; Margin is the final $B_t - D_t$.

Dataset	H	Seed	Origins	Matured	SAR(Act.)	Entry	Exit	MSE gain	Margin
ETTh1	336	2025	2545	2209	26.4%	1	1	↑6.92%	-67.0
ETTh1	336	2026	2545	2209	26.4%	1	1	↑6.52%	-64.3
ETTh1	336	2027	2545	2209	26.4%	1	1	↑6.73%	-65.9
PEMS08	48	2025	3524	3476	97.3%	2	1	↓23.54%	+217.7
PEMS08	48	2026	3524	3476	100.0%	1	0	↓24.62%	+237.4
PEMS08	48	2027	3524	3476	100.0%	1	0	↓25.33%	+247.0

The six runs contain 18207 forecast decisions. By the final decision, targets are complete for 17055 forecasts; the remaining 1152 are the final H forecasts of each run. A code check finds 0 places where the controller would have seen feedback before its target was complete. In other words, these logs never use future information early. This confirms the timing is correct, not that the predictions are accurate.

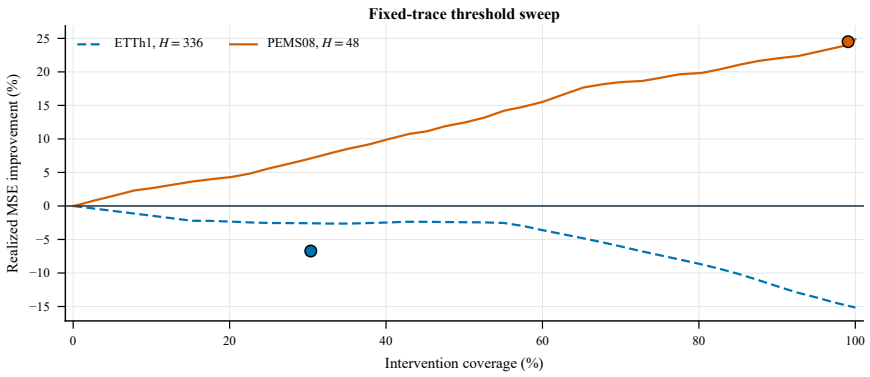


Figure 9: **Error versus the fraction of forecasts using correction.** Each curve applies confidence thresholds to a fixed MDMLP-EIA decision log. Both axes use only origins whose targets were complete by the final decision. Circles mark the deployed policies on this same subset. These are after-the-run checks, not results of selecting new policies on test data.

The coverage in Figure 9 uses the completed-target subset, whereas SAR(Act.) in Tables 16 and 17 uses all forecast origins. In each ETTh1 run, 671 forecasts use correction: $671/2545 = 26.4\%$ over the full run and $671/2209 = 30.4\%$ over the completed-target subset. The two percentages describe the same decisions with different denominators.

For each recorded run, the threshold sweep changes how often a correction is retained. On PEMS08, every tested nonzero rate lowers MSE; the deployed policy corrects 99.1% of forecasts and lowers MSE by 24.50%. On ETTh1, every tested nonzero rate raises MSE; the deployed policy corrects 30.4% of forecasts and raises MSE by 6.72%. Using correction less often therefore does not recover a benefit in this ETTh1 case. These are checks of fixed logs, not results from retraining or selecting a new policy.

C.3 Case-specific checks of decisions and delayed accounting

The following supplementary checks use the existing MDMLP-EIA anchor logs for PEMS08/$H = 48$ and ETTh1/$H = 336$, with seeds 2025–2027. The archived file contains 18,207 decisions. Candidate and issued losses are available for 17,055 origins whose targets are complete by the final decision; the remaining 1,152 are the last H origins of each run. We do not impute them. All

actions are recorded online. Comparing the stored candidate losses after the run does not change those actions or supply feedback early.

Continued assessment after withdrawal. For PEMS08 seed 2025, the controller's recorded choice changes from Adaptive LCRT to Base, then to Pure LCRT, then back to Adaptive LCRT, at origins 95, 191, and 239; the latest feedback available at those points reaches origins 47, 143, and 191. It left because its risk budget ran out, re-entered after Pure LCRT passed its entry test, and later switched to Adaptive LCRT on paired evidence. Seeds 2026 and 2027 stay with Adaptive LCRT through the first 300 origins. This records what the controller did, not that those transitions were optimal.

Two ways of scoring the same issued forecasts. All ETTh1 runs start with Adaptive LCRT and return to Base at $2H - 1 = 671$. By then the first $H = 336$ issued forecasts have complete targets, but the forecasts from 336 to 670 are still waiting. Table 18 separates the loss already known at the switch from the loss of those still incomplete forecasts, which arrives only later. After origin 671 every issued forecast is simply Base. So the curve of issued losses is flat after 671, even though the online account keeps receiving delayed losses from earlier origins.

Table 18: **ETTh1 loss accounting at the first return to Base.** Summed issued-minus-Base MSE, not percentages. "Known" scores origins 0–335; "Pending" scores 336–670 after their targets arrive; "Later" scores 671–2208. Losses are never admitted before their full targets arrive.

Seed	Known at exit	Pending at exit	Later issued forecasts
2025	22.117	44.895	0.000
2026	20.892	43.437	0.000
2027	21.786	44.115	0.000

The implementation checks the issued-loss budget before choosing a candidate and scores all stored candidates once their targets arrive. An audit of the 17,055 scored origins finds no record that uses feedback early and no mismatch between the issued loss and the candidate that was actually chosen. These checks confirm what the controller did and when it could use the evidence; they do not isolate the value of low-rank projection or prove the thresholds are optimal. The bundled trace file and CSVs keep the data needed to rebuild the accounting table.

C.4 Measured latency and memory on selected configurations

Table 19: **Measured runtime and memory on two backbones.** Timing uses 10 warm-ups and 50 repeats. p50 is median prediction time and p95 is its 95th percentile, both in milliseconds; each displayed timing is averaged over three seed runs. Time Ratio is SAR p50 divided by Base p50, the reciprocal of Speed Ratio in Table 4. ΔGPU/CPU are memory increases over Base. Calib. is one-time correction fitting in seconds; Update is the controller-state update time in microseconds.

Backbone	Dataset	H	Base p50	SAR p50	SAR p95	Time Ratio	ΔGPU (MB)	ΔCPU (MB)	Calib. (s)	Update (μs)
MDMLP-EIA	ETTh1	336	37.27	43.34	81.24	1.16×	+0.000	+0.33	401.3	443.5
MDMLP-EIA	PEMS08	48	16.49	34.45	64.49	2.09×	+0.000	+1.15	597.3	230.3
SimpleTM	ETTh1	336	25.74	29.52	84.14	1.15×	+0.000	+0.30	456.5	336.1
SimpleTM	PEMS08	48	23.27	39.55	86.95	1.70×	+0.000	+0.82	647.9	879.9

The benchmark uses FP32 and one forecast at a time, with seeds 2025–2027. Each run has ten warm-ups and 50 timed predictions; every number below is the mean of the three per-seed values. The time ratio divides the mean SAR-LCRT p50 by the mean Base p50, and ranges from 1.15 to 2.09 across the four settings. Peak GPU memory does not grow, and CPU memory grows by only 0.30–1.15 MB. Fitting the correction takes 401–648 s once per checkpoint and is not counted in per-forecast time; updating the controller takes 230–880 μs. The larger ratios on PEMS08 mostly reflect its faster Base model: SAR-LCRT takes 34.45–39.55 ms in mean p50 versus 16.49–23.27 ms for Base. These numbers cover two backbones and two tasks on one machine, so they may differ on other hardware.

Table 20: **Absolute memory measurements underlying Table 4.** Each value is the mean of three per-seed peak measurements in MB. GPU values refer to allocated tensor memory, not reserved device memory; CPU values refer to process RSS. Keeping both denominators visible allows the memory ratios to be checked without combining GPU and CPU memory.

Backbone	Dataset	H	Base GPU	SAR-LCRT GPU	Base CPU	SAR-LCRT CPU
MDMLP-EIA	ETTh1	336	40.605	40.605	863.384	863.712
MDMLP-EIA	PEMS08	48	75.812	75.812	966.591	967.743
SimpleTM	ETTh1	336	40.634	40.634	868.557	868.854
SimpleTM	PEMS08	48	43.877	43.877	967.246	968.061

D Protocol and Implementation Details

D.1 Data splits, training, and validation blocks

Data and optimization. ETT uses its standard time-ordered split. ECL, Weather, and PEMS-BAY use 70/10/20 train/validation/test splits; PEMS03/04/07/08 and LargeST use 60/20/20. Each channel is normalized with training statistics, except LargeST, which uses one global scale. Training minimizes normalized MSE with AdamW: learning rate 5×10^{-4}, weight decay 10^{-4}, cosine decay, at most 30 epochs, early-stopping patience 5, gradient clipping at 1.0, effective batch size 32, and deterministic FP32 execution.

Validation partitions. The first 25% of validation forecast origins is used to select the backbone checkpoint. A gap of $L + H$ origins separates this block from correction fitting. Of the remaining origins, 70% are assigned to fitting, with at least 48 fitting origins and 32 later evaluation origins required. An additional H-origin gap ensures that fitting targets are available before the fitted candidates are evaluated. If a valid split cannot be formed, Pure LCRT keeps a conservative step and SAR selects Base.

D.2 Correction and selection settings

Features and candidate settings. Spatial features summarize input levels, recent changes, trends, successive differences, lag-one dependence, three frequency bands, and differences between channels. They also include an intercept and up to eight discrete cosine transform coefficients for each channel-mean, standard-deviation, and trend vector. Temporal maps use 14 features per channel. The exact formulas are in Appendix E.3. Validation folds expand over time and select the ridge penalty $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, spatial rank $r \in \{2, 4, 8\}$, temporal rank $r \in \{1, 2, 4\}$, coefficient-blending weight $\eta \in \{0, 0.25, 0.5, 1\}$, and step $\alpha \in \{0.01, 0.025, 0.05, 0.1, 0.2, 0.35, 0.5, 0.75\}$. Candidates are ranked by mean gain, then worst-fold gain, then smaller step, rank, and ridge penalty.

Final correction size and initial action. After fitting the final coefficients, a separate validation block chooses a positive correction size no larger than the inner-fold choice. It considers the positive grid $\{0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.35, 0.5, 0.75\}$ together with the retained inner-fold step as specified in Appendix E.3. It first maximizes the lower confidence bound (LCB), then mean gain, and finally prefers the smaller step.

Online selection uses up to $8H$ completed forecasts, requires at least $\max(8, 3H)$, and checks gains across five time blocks. The Newey–West bandwidth is at least $H - 1$, capped by the available history. Entry requires a positive mean gain and two positive blocks. Switching from Pure LCRT to Adaptive LCRT additionally requires both candidates to be entry-eligible, a positive paired 90% LCB, and four positive paired blocks.

The method starts with the validation-selected action. Each correction episode starts with zero validation credit.

Order of prediction and feedback. At time t, the frozen backbone and correction maps produce three forecasts from $\boldsymbol{X}_t$. The controller also receives stored losses only for origins j with $j + H \leq$

t. All three candidate losses may enter the comparison window, but only the loss of the forecast actually selected at j enters the episode account. SAR then selects one current forecast and stores the candidates and action for later scoring. Constructing candidates and receiving past feedback can occur in either software order: neither uses the current forecast's future targets. An earlier action is never changed afterward.

D.3 Selection thresholds and why they are used

SAR uses different conditions to choose the initial action, start correction, switch between candidates, and return to Base. Table 21 lists these conditions. Time windows scale with the prediction length H, and gains are measured relative to a reference forecast. Controller constants are fixed across tasks; correction parameters and step sizes are selected using validation data, not test results.

Table 21: **Operational SAR thresholds and their roles.** H is the forecast horizon, n the available evidence length, and b the number of chronological summary blocks. Settings describe the evaluated controller; design rationale is not a claim of optimality.

Decision or quantity	Setting / condition	Intended role
Feedback eligibility	Admit j only if $j + H \le t$	Exclude incomplete target losses.
Validation initialization	At least 32 origins and 2 blocks; positive 90% LCB and $\ge 2/3$ positive blocks	Audit the exact frozen candidate before assigning the initial action.
Online evidence	Latest $8H$ matured origins; reports refresh each H arrivals; availability requires $\max(8, 3H)$ origins	Limit stale history while retaining several overlapping horizons.
Summary blocks	Online: 5 blocks; validation: $b = \min(3, \max(1, \lfloor n/16 \rfloor))$	Check that evidence is spread across time rather than concentrated in one short interval.
Online HAC bandwidth	$L = \min(n-1, \max(H-1, L_{\text{auto}}))$	$L_{\text{auto}} = \min(n-1, \max(1, \lfloor 4(n/100)^{2/9} \rfloor))$; covers overlap.
Confidence level	LCB $= \bar{z} - t_{0.9,b-1}\widehat{\text{SE}}_{\text{HAC}}$; UCB uses $+$	One-sided 90% screening, not an anytime guarantee.
Entry from Base	Positive mean gain and ≥ 0.4 positive blocks (2/5); one refreshed report suffices	Responsive re-entry; the loss budget separately monitors harm.
Pure LCRT to Adaptive LCRT	Both entry-eligible; paired Adaptive-LCRT-advantage LCB > 0 and $\ge 2/3$ positive blocks (4/5)	Demand stronger evidence before replacing a useful Pure LCRT proposal.
Adaptive LCRT to Pure LCRT	Pure LCRT entry-eligible; paired UCB < 0 and $\ge 2/3$ negative blocks (4/5)	Return to Pure LCRT when the paired evidence favors it.
Exit to Base	At least H matured losses in the episode; $B_t < D_t - 10^{-12}$	Give realized loss precedence over candidate entry/switch evidence.
Risk credit and reserve	Zero validation credit; reserve one adverse horizon, or two when consecutive	Avoid subsidizing test losses with validation gains; respond to persistent downside.

Why entry and replacement differ. Entry requires positive mean gain and two positive blocks out of five (40%). It does not require a majority. This lets the controller start a correction on limited recent evidence while the budget separately checks its actual losses. Replacing Pure LCRT with Adaptive LCRT requires a direct comparison of their losses and stronger agreement across blocks. Two-thirds means four of five online blocks. Validation normally uses three blocks and requires two positive blocks, or both when only two blocks are available. These are design choices intended to make entry more responsive than switching, not proven optimal fractions.

The zero-gain threshold asks whether the candidate improves its reference. The one-sided 90% bound also accounts for uncertainty in the mean. We do not establish that 90% is better than 95%. Exit instead uses accumulated losses of the forecasts actually output, starting with no credit from

validation. The reserve uses one recent horizon of net loss and adds the preceding horizon only if both have net losses. This distinguishes repeated recent losses from a single adverse block.

Dependence and timing. For $n > 1$, the automatic bandwidth is $L_{\text{auto}} = \min(n - 1, \max(1, \lfloor 4(n/100)^{2/9} \rfloor))$. Validation uses this bandwidth; online comparisons also require at least $H - 1$, capped at $n - 1$, to cover overlapping forecast targets. The standard error uses Bartlett-weighted HAC covariances. The critical value is the Student-t 0.9 quantile with $b - 1$ degrees of freedom, where b is the number of blocks. These bounds guide decisions; they do not give a simultaneous 90% guarantee for all repeated decisions. The $8H$ window limits old data, and updating comparisons every H arrivals reduces repeated computation.

The budget is checked before selection at each origin once an episode has H completed forecasts. With origins starting at zero, the first possible exit check is at $2H - 1$, while candidate comparisons first have $3H$ completed forecasts at $4H - 1$. Before the exit check becomes available, correction can continue with a negative partial balance. Once active, a candidate is not stopped simply because it fails the entry condition: exit uses the budget, and switching uses paired evidence. No two-update waiting period is required for exit. Re-entry starts a new episode with zero validation credit; 10^{-12} is only a numerical tolerance.

Table 22 is a reproduction guide: it records the unit and averaging order behind each class of result, allowing the reported tables to be checked against their source data.

Table 22: **Raw-to-paper aggregation and audit trail.** "Task" denotes one backbone–dataset–horizon combination; pp denotes percentage points.

Reported object	Statistical unit	Operation	Convention
Forecast error	Task–seed	Mean over the task's test origins and channels; then mean over three matched seeds	Raw precision
Paired gain	Task	$100(1 - \bar{e}^{k}/\bar{e}^{\text{Base}})$ after seed averaging	Two decimals
Primary macro and CI	416 tasks; 13 clusters	Task mean; 20,000 dataset-clustered bootstrap replicates (seed 2701)	95% interval
Table 1 count	104 dataset–backbone cells	Four-horizon averages; counts and highlighting compare displayed values	Three decimals
Residual-baseline W/E/L	96 tasks	Paired gains within ± 0.1 pp count as equal	Fixed tolerance
Trace maturity	Issued origin	Admit record j only when $j + H \leq t$; the final H origins per case remain censored	Exact count

D.4 Memory and computation

For feature width F and rank r, storing the spatial, pooled temporal, and channel-specific maps takes $O(Hr + Cr + Fr^2)$, $O(Fr + Hr)$, and $O(CFr + Hr)$ space, respectively. Computing FFT features costs $O(LC \log L)$. Applying a spatial map costs $O(Fr^2 + Hr^2 + HCr)$; applying a temporal map costs $O(CFr + CHr)$.

A streaming controller could keep the $8H$ comparison window, pending losses, episode sums, and two recent loss blocks in $O(H)$ space. HAC comparisons cost $O(H^2)$ every H origins, averaging $O(H)$ work per origin; running sums update in $O(1)$. The reference Python replay does not implement this compact storage: it keeps $O(T)$ history and constructs completed-loss arrays at each origin, adding $O(T)$ work when the history length is T. Exporting full trajectories also uses $O(T)$ space. The measured times in Table 19 describe this implementation, not the compact streaming bounds.

The $O(H)$ controller-state bound concerns scalar loss histories and episode accounts. If all three candidate prediction tensors are retained until their complete targets arrive, each of $O(H)$ outstanding origins stores $O(HC)$ values, giving $O(H^2C)$ additional storage. This distinguishes pending prediction storage from the scalar loss arrays used by the offline reference replay. The reported GPU and CPU measurements apply only to the measured implementation and configurations. In short, the added time comes from the correction and selection steps, while the added memory comes mostly from holding recent forecasts until their targets arrive.

E Additional Implementation Details

E.1 How the recorded results produce the tables

The residual-baseline export keeps Base and the three baseline methods for 288 task–seed runs, giving 1,152 method–seed records. At each step the controller can only use feedback for forecasts whose targets have already arrived, so the final decision can use only the first $T - H$ forecasts. Full-task scoring later also includes the last H forecasts once their targets arrive, but those late losses never affect earlier decisions.

For MDMLP-EIA on ETTh1/$H = 336$, Base/SAR-LCRT MSE averaged over three seeds is $0.479162/0.504996$ on all 2,545 origins and $0.442719/0.472483$ on the 2,209 origins with complete feedback. Tables 10 and 16 thus score different parts of the same runs. Their MSE improvements are approximately -5.39% and -6.72%, respectively.

E.2 Validation settings for the residual baselines

Mean uses windows $\{4H, 8H, 16H, \text{all}\}$. EWMA uses half-lives $\{0.5H, H, 2H, 4H, 8H, 16H\}$, with update weight $1 - 2^{-1/h}$ for half-life h. Ridge combines penalties $\{0.001, 0.01, 0.1\}$ with the Mean windows. It uses 14 temporal features, leaves the intercept unpenalized, and scales the penalty by the number of channel rows. This online baseline does not standardize features further; frozen LCRT standardizes them on its fitting block.

For n validation origins, scoring starts at $\max(3H, \lfloor n/4 \rfloor)$ and divides the rest into three consecutive blocks. Each prediction is scored before its available residual can update later predictions. Settings minimize the mean of the three block MSEs; exactly equal values follow the fixed trial order. If fewer than three scoring origins remain, defaults are Mean with all history, EWMA half-life $4H$, and Ridge penalty 0.01 with all history. For each method, 252 of 288 runs use validation selection and 36 use defaults. The export keeps 16,632 trial–block records and 288 parameter-selection records covering all three estimators, plus the residual-baseline code and source hashes. These are the tested grids, not an exhaustive comparison of all possible designs.

After choosing settings, we create a fresh estimator and process the available validation calibration suffix in time order. This suffix follows checkpoint selection and its $L + H$ gap; it is not limited to LCRT's fitting block. Only validation residuals initialize the estimator. If the suffix is unavailable, the estimator starts empty and outputs zero correction until its first eligible update. Test updates likewise wait for the full target of each forecast.

E.3 Feature definitions and correction fitting

Let x_{lc} denote the input at position $l = 1, \ldots, L$ and channel $c = 1, \ldots, C$. Feature standard deviations below use the population denominator, unlike the sample SDs reported across seeds. For $L = 96$, recent and quarter windows both have length 24. Define channel mean μ_c, standard deviation σ_c, last value $v_c = x_{Lc}$, recent mean u_c, level shift $u_c - \mu_c$, and quarter trend $q_c = \text{mean}_{l=L-23}^{L} x_{lc} - \text{mean}_{l=1}^{24} x_{lc}$. Let $d_{lc} = x_{l+1,c} - x_{lc}$ for $l = 1, \ldots, L-1$ and $z_{lc} = x_{lc} - \mu_c$. Lag-one dependence is

$$\rho_c = \frac{(L-1)^{-1} \sum_{l=1}^{L-1} z_{l+1,c} z_{lc}}{\max(L^{-1} \sum_{l=1}^{L} z_{lc}^2, 10^{-6})}. \tag{9}$$

For real FFT power P_{kc} of z_{lc}, the three band ratios sum power over bins 1–3, 4–12, and 13–$\lfloor L/2 \rfloor$, dividing each by $\max(\sum_{k=1}^{\lfloor L/2 \rfloor} P_{kc}, 10^{-6})$. DC is excluded. These quantities form the 14-element temporal feature vector:

$$\psi_c = (1, v_c, \mu_c, \sigma_c, u_c, u_c - \mu_c, q_c, \text{mean}_l\, d_{lc}, \text{mean}_l\, |d_{lc}|, \text{sd}_l\, d_{lc}, \rho_c, b_{1c}, b_{2c}, b_{3c})^\top \in \mathbb{R}^{14}. \tag{10}$$

The spatial vector contains an intercept and 15 scalars: global input mean and standard deviation; channel-averaged last value, recent mean, level shift, and quarter trend; global mean absolute first difference and standard deviation of signed differences; mean channel lag-one dependence; three band ratios computed after averaging FFT power over channels; and cross-channel standard deviations of μ_c, σ_c, q_c. We also project these last three channel vectors onto a discrete cosine transform

(DCT) basis. The orthonormal DCT basis is $Q_{c0} = C^{-1/2}$ and $Q_{ck} = \sqrt{2/C}\cos[\pi(c-1/2)k/C]$ for $k = 1, \ldots, \min(8, C) - 1$, using the original channel ordering. Thus $F = 16 + 3\min(8, C)$: 37 for seven-channel ETT and 40 when $C \geq 8$. NaNs are replaced by zero and infinite feature values by signed 10^4.

We compute feature means and scales on each fitting block and keep them fixed. Feature variance is floored at 10^{-8}; the intercept has mean zero and scale one. Equation 2 fits the standardized features $S^\top \phi$. Coefficients are converted back to the original feature scale before low-rank projection. For a shared temporal map, let N be the number of fitting origins and $r_{tc} \in \mathbb{R}^H$ the residual sequence of channel c. We fit

$$A_P^* = \operatorname*{arg\,min}_{A \in \mathbb{R}^{14 \times H}} \sum_{t,c} \|r_{tc} - A^\top S_P^\top \psi_{tc}\|_2^2 + \lambda NC \|A_{-0}\|_F^2, \qquad B_P = S_P A_P^*. \tag{11}$$

The channel-specific fit uses its own standardizer and N rows:

$$A_c^* = \operatorname*{arg\,min}_{A \in \mathbb{R}^{14 \times H}} \sum_{t} \|r_{tc} - A^\top S_c^\top \psi_{tc}\|_2^2 + \lambda N \|A_{-0}\|_F^2, \qquad B_c = S_c A_c^*. \tag{12}$$

The channel-specific version blends the shared and individual coefficients: $B_c^{(\eta)} = (1-\eta)B_P + \eta B_c$. For the shared map, we truncate the SVD of $B_P \in \mathbb{R}^{14 \times H}$. For the blended map, we stack the coefficients into a $14C \times H$ matrix and truncate its SVD, giving a common horizon basis across channels. Predictions use $\Delta_{t,:,c} = \widetilde{B}_c^\top \psi_{tc}$. In the shared map, $\widetilde{B}_c$ is the same for every channel. Rank is capped by the matrix dimensions. Spatial projection uses randomized SVD with the run seed and the library's default iteration setting; temporal projection computes the full deterministic SVD before truncation.

After fitting the final coefficients, step selection considers the inner-fold step, $\min(0.001, \alpha_{\text{inner}})$, and grid steps between these values. It chooses the largest available validation HAC LCB, then the largest mean gain, then the smallest step. If no report is available, the same ordering is applied to default report values. This changes only the positive step, not the fitted coefficients, and uses no test targets. The resulting candidate is then checked for initial-action eligibility. Because the same validation block selects the step and initializes the action, its LCB guides selection rather than certifying performance after selection.

E.4 Decision order and episode start times

At time t, receive losses only for $j + H \leq t$. Let P and A mean that Pure LCRT and Adaptive LCRT pass their entry conditions. R_+ means both pass, Adaptive LCRT has a positive paired LCB over Pure LCRT, and at least two-thirds of blocks favor Adaptive LCRT. R_- means Pure LCRT passes, the paired UCB is negative, and at least two-thirds of blocks favor Pure LCRT. With five online blocks, two-thirds requires four. Decisions follow this order:

1. If correction is active and H forecasts in the episode have complete targets, exit when $B_t < D_t - 10^{-12}$. This check precedes candidate selection.
2. If too few completed forecasts are available for comparison, retain the initial action at the first origin or the preceding action thereafter.
3. From Base, choose Adaptive LCRT if R_+, otherwise Pure LCRT if P, otherwise Adaptive LCRT if A, otherwise Base. One refreshed eligible report suffices.
4. From Pure LCRT, switch to Adaptive LCRT only if R_+. From Adaptive LCRT, switch to Pure LCRT only if R_-. Otherwise retain the active branch.
5. Starting correction from Base sets the episode start to t. Switching between corrections does not reset it. The account uses the action actually selected for each past forecast.

The configured number of validation-prior horizons is zero, so the budget receives no validation credit even though the code stores a validation LCB. Feedback from before re-entry can still inform candidate comparisons, but does not enter the new episode account. For an episode starting at τ, the first exit check is at $\tau + 2H - 1$. With the numerical tolerance, continuing correction after a check gives $B_t \geq D_t - 10^{-12}$ rather than an exact infinite-precision nonnegative balance. The

reserve uses the latest two non-overlapping completed-loss blocks, restricted to the episode. Unused harm-confirmation fields in the code do not delay exit.

E.5 Complete MDMLP-EIA MAE means and standard deviations

Table 23: **Complete MDMLP-EIA MAE results accompanying Table 10.** Errors are means ± sample SD over three seeds, scoring every test forecast. Gain is the percentage reduction from Base: ↓ marks a reduction and ↑ an increase. All SDs are measured.

Dataset	H	Base MAE	Pure LCRT MAE	SAR-LCRT MAE	MAE gain
ECL	96	0.250±0.0002	0.250±0.0002	0.251±0.0007	↑0.30%
ECL	192	0.259±0.0001	0.259±0.0001	0.260±0.0001	↑0.34%
ECL	336	0.277±0.0006	0.277±0.0006	0.277±0.0006	0.00%
ECL	720	0.313±0.0017	0.313±0.0017	0.313±0.0017	↑0.01%
ETTh1	96	0.396±0.0003	0.396±0.0003	0.396±0.0003	↑0.03%
ETTh1	192	0.426±0.0011	0.426±0.0010	0.426±0.0010	↓0.02%
ETTh1	336	0.449±0.0008	0.458±0.0007	0.471±0.0005	↑4.93%
ETTh1	720	0.479±0.0068	0.479±0.0068	0.479±0.0068	0.00%
ETTh2	96	0.336±0.0008	0.336±0.0008	0.336±0.0008	↓0.02%
ETTh2	192	0.390±0.0006	0.390±0.0006	0.390±0.0006	↓0.01%
ETTh2	336	0.427±0.0034	0.427±0.0031	0.427±0.0031	↓0.08%
ETTh2	720	0.448±0.0010	0.449±0.0008	0.448±0.0011	↑0.01%
ETTm1	96	0.360±0.0004	0.359±0.0006	0.359±0.0005	↓0.36%
ETTm1	192	0.385±0.0007	0.383±0.0006	0.383±0.0006	↓0.36%
ETTm1	336	0.406±0.0011	0.405±0.0006	0.407±0.0007	↑0.12%
ETTm1	720	0.443±0.0003	0.443±0.0003	0.443±0.0008	↓0.02%
ETTm2	96	0.258±0.0006	0.258±0.0005	0.258±0.0009	↑0.10%
ETTm2	192	0.302±0.0009	0.302±0.0013	0.302±0.0013	↑0.25%
ETTm2	336	0.339±0.0014	0.338±0.0015	0.338±0.0015	↓0.15%
ETTm2	720	0.397±0.0025	0.397±0.0028	0.397±0.0028	↓0.10%
LargeST-GBA	12	0.132±0.0004	0.132±0.0004	0.132±0.0005	↑0.08%
LargeST-GBA	24	0.154±0.0003	0.154±0.0003	0.154±0.0002	↑0.15%
LargeST-GBA	36	0.166±0.0004	0.166±0.0004	0.166±0.0004	↑0.08%
LargeST-GBA	48	0.174±0.0001	0.173±0.0001	0.173±0.0001	↓0.26%
LargeST-SD	12	0.108±0.0001	0.110±0.0001	0.109±0.0003	↑1.29%
LargeST-SD	24	0.128±0.0002	0.131±0.0002	0.128±0.0002	↑0.24%
LargeST-SD	36	0.138±0.0001	0.143±0.0002	0.141±0.0002	↑1.91%
LargeST-SD	48	0.146±0.0002	0.156±0.0052	0.149±0.0008	↑1.81%
PEMS-BAY	12	0.262±0.0017	0.262±0.0015	0.262±0.0013	↑0.21%
PEMS-BAY	24	0.338±0.0001	0.342±0.0001	0.341±0.0002	↑0.89%
PEMS-BAY	36	0.396±0.0005	0.395±0.0010	0.394±0.0007	↓0.59%
PEMS-BAY	48	0.442±0.0009	0.436±0.0003	0.436±0.0004	↓1.36%
PEMS03	12	0.185±0.0001	0.182±0.0001	0.181±0.0001	↓2.21%
PEMS03	24	0.233±0.0005	0.226±0.0002	0.223±0.0006	↓4.24%
PEMS03	36	0.278±0.0002	0.266±0.0009	0.260±0.0005	↓6.26%
PEMS03	48	0.318±0.0007	0.302±0.0010	0.293±0.0005	↓7.88%
PEMS04	12	0.203±0.0005	0.201±0.0005	0.200±0.0002	↓1.57%
PEMS04	24	0.257±0.0004	0.250±0.0005	0.247±0.0007	↓3.93%
PEMS04	36	0.308±0.0013	0.294±0.0005	0.288±0.0004	↓6.55%
PEMS04	48	0.354±0.0009	0.331±0.0010	0.324±0.0002	↓8.44%
PEMS07	12	0.172±0.0004	0.172±0.0003	0.172±0.0003	↓0.37%
PEMS07	24	0.221±0.0004	0.219±0.0004	0.219±0.0004	↓0.86%
PEMS07	36	0.265±0.0017	0.260±0.0009	0.260±0.0009	↓1.74%
PEMS07	48	0.300±0.0004	0.293±0.0005	0.293±0.0005	↓2.37%
PEMS08	12	0.193±0.0004	0.190±0.0006	0.188±0.0005	↓2.75%
PEMS08	24	0.247±0.0002	0.240±0.0002	0.233±0.0001	↓5.51%
PEMS08	36	0.298±0.0008	0.284±0.0007	0.273±0.0006	↓8.16%
PEMS08	48	0.344±0.0011	0.326±0.0027	0.307±0.0007	↓10.61%
Weather	96	0.208±0.0013	0.211±0.0013	0.220±0.0013	↑5.60%
Weather	192	0.250±0.0006	0.256±0.0029	0.270±0.0002	↑7.75%
Weather	336	0.291±0.0004	0.302±0.0004	0.297±0.0058	↑2.18%
Weather	720	0.343±0.0006	0.343±0.0006	0.343±0.0006	↓0.08%