TRACE: RETRIEVAL-AUGMENTED CORRECTION OF TIME-SERIES FORECASTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval augmented time series forecasting improves predictions by drawing on relevant historical cases. Existing methods typically retrieve past windows similar to the current observation and reuse the trajectories that followed them. However, a strong forecaster may already capture much of the useful structure in those trajectories, so transferring them directly can overwrite parts of an otherwise accurate prediction. We instead ask what historical information remains useful *after* a forecast has already been produced. We introduce TRACE, a post-hoc correction framework that retrieves a forecaster's past errors to identify what it tends to miss in similar situations. TRACE reduces unreliable corrections when retrieved errors disagree and uses validation data to determine whether a correction should be applied and how much historical trajectory information should be retained, without modifying the forecasting backbone. Across 585 settings covering 16 datasets and 13 forecasting backbones, TRACE reduces the average long term MSE by 15.6% and the average short term traffic MAE by 23.1% relative to uncorrected forecasts. Further analysis shows that retrieving past errors becomes increasingly favorable as the underlying forecaster improves, while borrowing more of a historical trajectory is mainly useful at shorter forecasting horizons. These results suggest that as forecasting models become stronger, retrieval is increasingly valuable for correcting recurring mistakes rather than replacing an already informative forecast.

1 INTRODUCTION

Modern time series forecasters capture increasingly complex temporal structure through frequency decomposition, patch based representations, cross variable attention, multiscale mixing, exogenous conditioning, and large scale pretraining (Zhou et al., 2022a; Nie et al., 2023b; Wu et al., 2023; Liu et al., 2024a; Wang et al., 2024a;d; Ansari et al., 2024b). Retrieval offers a complementary way to exploit historical information at inference time. Existing retrieval augmented forecasting methods retrieve related time series and historical windows (Yang et al., 2022; Tire et al., 2024), learn retrieval and prompting strategies (Zhang et al., 2024), augment foundation models with retrieved information (Ning et al., 2025), and develop specialized mechanisms for selecting and incorporating historical cases (Han et al., 2025; Zhou et al., 2026; Lee et al., 2026; Kang et al., 2026). These methods primarily address two questions: *which historical cases should be retrieved*, and *how should the retrieved information be incorporated into a forecasting model*. In this work, we ask a complementary question: *given a retrieved historical case, what information from that case actually complements the forecast already produced?*

A retrieved case can closely resemble the current lookback window while its continuation contains information that the forecaster has already captured. The trajectory that followed a historical window may therefore contain both information missing from the current forecast and patterns already represented correctly in it. Reusing the entire historical trajectory can consequently overwrite useful parts of the current prediction rather than supply only what it is missing, as illustrated in Figure 1. Similarity to the current lookback window is therefore not sufficient to determine which information from the corresponding historical continuation should be transferred. Recent work has highlighted a related problem under distributional change: similar historical windows do not necessarily lead to equally relevant continuations (Zhou et al., 2026). Our focus is complementary. Even when a historical case is relevant, its usefulness depends on what the forecaster has already captured. This distinction separates *case relevance*, whether a historical case resembles the current observations,

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

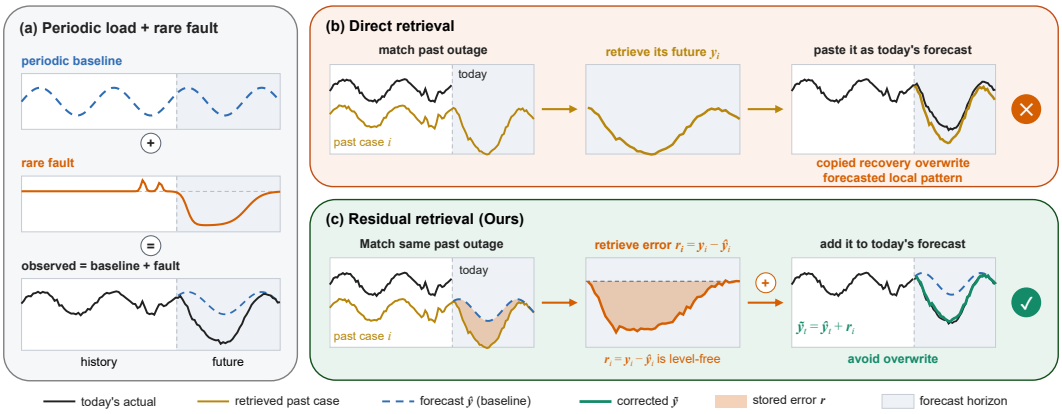

Figure 1: **Retrieve the error, not the input.** (a) An observed series is the sum of a dominant periodic component and an infrequent event. (b) Retrieving on the input matches a past window's *routine* structure, which carries almost all of the window's energy; copying the future that followed then injects the old level and phase on top of a forecast that already had them right. (c) Subtracting the forecaster's own prediction from that past outcome leaves only what it missed, turning the retrieved quantity into a correction that can be added to today's forecast.

from *forecast complementarity*, whether that case contains information still missing from the current prediction.

Historical forecasting records provide a direct way to identify such complementary information. For each past forecasting case, we know what the model predicted and, once the target period has elapsed, what actually occurred. Their difference records precisely what that forecaster failed to capture on that case. Instead of transferring the entire trajectory following a similar historical window, we can therefore retrieve past forecast errors and use them to correct mistakes that recur in similar situations. Crucially, these errors are forecaster dependent. Two models can observe the same historical window yet produce different predictions, leaving different errors behind. Historical forecast errors thus characterize not only the observed data, but also the characteristic information missed by the particular forecasting model being corrected.

This view also reveals a direct connection between historical trajectories and historical forecast errors. When a retrieved historical continuation is used in place of the current forecast, the resulting change can be decomposed into two components. The first is the historical forecast error, which captures what the forecaster missed on the retrieved case. The second is the difference between the forecast made for that historical case and the forecast made for the current case, which we call *trajectory drift*. Using the full historical trajectory retains both components, whereas using only the historical forecast error retains the first. Intermediate choices retain different amounts of trajectory drift. This decomposition makes an otherwise implicit design choice explicit: *how much information should be borrowed from a retrieved trajectory beyond what is already captured by its forecast error?* Motivated by this decomposition, we introduce TRACE, a post hoc framework for correcting forecasts from an existing time series model. TRACE maintains a chronological memory of the model's past forecast errors and retrieves corrections from relevant forecasting situations. Retrieval considers both the observed history and the forecast produced by the model, allowing retrieved cases to reflect what has been observed as well as what the forecaster expects to happen next. When retrieved errors disagree, TRACE attenuates the correction rather than forcing uncertain information into the prediction. Using held out validation data, it further determines how much trajectory drift should be retained and whether correction is beneficial at all. When historical evidence does not support a reliable correction, TRACE can use a simpler correction or preserve the original forecast. The forecasting backbone itself remains unchanged.

We evaluate TRACE across 585 settings covering 16 real world benchmarks and 13 forecasting architectures. To isolate the effect of retrieval, our main comparison fixes the base forecasts across all methods and varies only which historical information is retrieved and how it is used. Under this controlled comparison, TRACE reduces average long term MSE by 15.6% and average short term

traffic MAE by 23.1% relative to the uncorrected forecasts. In separate end to end evaluations, MSE reductions reach 41.3% for individual forecasting backbones. The experiments also reveal a consistent shift in which historical information is most useful. As the underlying forecaster becomes more accurate, its past forecast errors become increasingly favorable relative to full historical trajectories. Meanwhile, retaining more trajectory drift is most beneficial at shorter forecasting horizons and becomes less useful as the horizon increases. We further apply the same correction framework to ensembles, AutoML systems, and zero shot and fine tuned foundation models, demonstrating that the approach is not tied to a particular forecasting architecture.

Our contributions are threefold. **(i) Retrieved information as a design dimension:** we distinguish between retrieving a historical case that resembles the current observations and retrieving information that complements an existing forecast. **(ii) Novel forecast-aware retrospective retrieval framework:** we decompose trajectory based correction into historical forecast error and trajectory drift, and develop TRACE to adapt how much of each should be retained while suppressing unreliable corrections. **(iii) Comprehensive empirical analysis:** we evaluate TRACE over 585 settings spanning 16 benchmarks and 13 forecasting architectures under fixed base forecasts, compare it with conventional retrieval objects and published retrieval operators, and extend the same interface to ensembles, AutoML systems, and zero shot and fine tuned foundation models.

2 PRELIMINARIES

Notation. For multivariate forecasting with d variables, the task at origin t is to predict the next H steps from the most recent L observations. Let $\mathbf{x}_t \in \mathbb{R}^{L\times d}$ be the observed context, $\mathbf{y}_t \in \mathbb{R}^{H\times d}$ its realized future, and $f : \mathbb{R}^{L\times d} \to \mathbb{R}^{H\times d}$ a forecaster with prediction $\widehat{\mathbf{y}}_t = f(\mathbf{x}_t)$. We assume f has already been trained and deployed: we may query it, but its parameters are never updated. The quantity this work centers on is the *model-relative residual* $\mathbf{r}_t = \mathbf{y}_t - \widehat{\mathbf{y}}_t$, defined jointly by the data and the particular forecaster and observable only after the whole target window has elapsed. A retrieval-augmented forecaster additionally has a memory $\mathcal{M}_t$ of past items and emits a corrected forecast $\widetilde{\mathbf{y}}_t$; methods differ in whether those items hold prior contexts with the futures that followed or, as we propose, the same forecaster's prior residuals. Throughout, *causal* means free of lookahead: since $\mathbf{r}_i$ needs the complete future at origin i, an item may enter $\mathcal{M}_t$ only if $i + H < t$.

Problem 1 (Retrieval-augmented correction of a trained forecast) *Given a trained forecaster f, a query context $\mathbf{x}_t$ with base forecast $\widehat{\mathbf{y}}_t = f(\mathbf{x}_t)$, and a memory $\mathcal{M}_t$ restricted to items whose targets were fully observed before t, construct an additive correction c_t so that $\widetilde{\mathbf{y}}_t = \widehat{\mathbf{y}}_t + c_t$ is more accurate than $\widehat{\mathbf{y}}_t$ with respect to $\mathbf{y}_t$, without updating f and without observing any part of $\mathbf{y}_t$.*

Problem 1 isolates the question of Section 1: the base forecast, the data, and the information boundary are fixed, so any accuracy difference is attributable to what $\mathcal{M}_t$ stores and how it is used.

3 METHOD

TRACE has three components, sketched in Figure 2: a chronological memory of a trained forecaster's own errors (Section 3.1), a model-conditioned retriever that turns them into a reliability-shrunk correction whose *object* is itself tunable (Section 3.2), and a validation gate that decides whether to apply it (Section 3.3).

3.1 BUILDING MEMORY OF FORECAST ERRORS

Residuals as retrieval targets. Existing works tend to score past windows by context similarity and reuse the futures that followed (Tire et al., 2024; Han et al., 2025; Lee et al., 2026; Kang et al., 2026; Zhou et al., 2026). Storing $\mathbf{r}_i$ instead of $\mathbf{y}_i$ fixes three properties of that object. Raw-window distance is dominated by the periodic and trend energy the base forecast already predicts, so the short prefix announcing an unusual continuation rarely decides the neighborhood, whereas a residual has that explained component subtracted out. A retrieved future is an absolute trajectory carrying its own level, phase, and covariate response, so injecting it overwrites parts of $\widehat{\mathbf{y}}_t$ that were already correct, while a residual can only move the forecast. And an input neighborhood is identical for every forecaster, whereas memory indexed to one forecaster suppresses the correction when that backbone

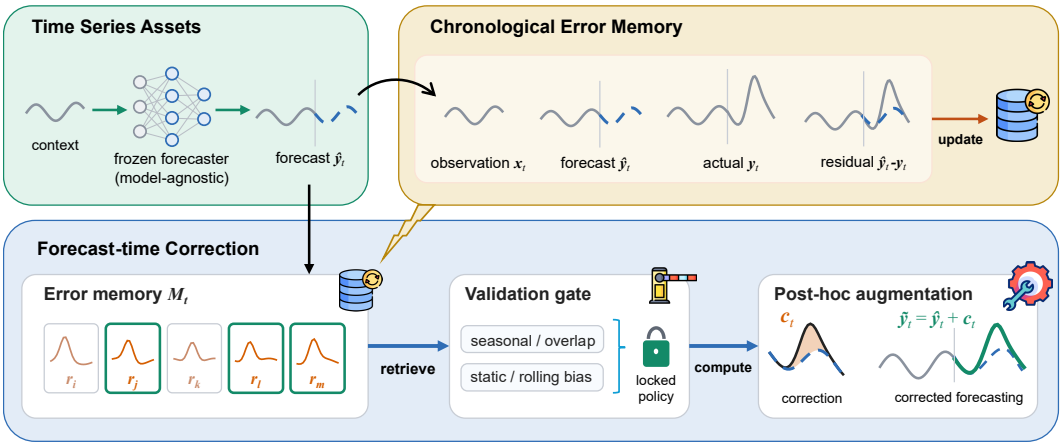

Figure 2: **TRACE separates validation-time calibration from forecast-time correction.** On a held-out validation split, the forecaster's own predictions are differenced against realized targets to build an error memory, and one correction policy is locked. At forecast time the backbone is never updated: the locked policy either retrieves agreeing past errors and adds them to the current forecast, or falls back to a simpler correction, or leaves the forecast unchanged. The actual continuation is never available at forecast time.

has no recurring error. Appendix Figure 4 shows all three effects on measured ETTh1 forecasts: neighbors with matching contexts have futures at visibly different levels, but their residuals share one error shape.

Unified correction. Retrieving a future and retrieving an error are not alternatives; they are the two ends of a single additive correction. Anchoring neighbor i's future to the query endpoint, as input retrieval does, and subtracting the base forecast gives an exact decomposition,

$$\big(\mathbf{y}_i + \boldsymbol{\delta}_{ti}\big) - \widehat{\mathbf{y}}_t \;=\; \mathbf{r}_i + \mathbf{g}_{ti}, \qquad \mathbf{g}_{ti} = \widehat{\mathbf{y}}_i + \boldsymbol{\delta}_{ti} - \widehat{\mathbf{y}}_t, \qquad \boldsymbol{\delta}_{ti} = \mathbf{1}_H \big(\mathbf{x}_{t,L} - \mathbf{x}_{i,L}\big)^\top \in \mathbb{R}^{H\times d}, \quad (1)$$

where $\mathbf{x}_{t,L}$ is the final context step and $\mathbf{1}_H$ is the H-dimensional vector of ones. The left side is what input retrieval adds to the forecast and $\mathbf{r}_i$ is what error retrieval adds; they differ by the drift $\mathbf{g}_{ti}$ between the forecaster's trajectory in the neighbor's state and in the query's. Input retrieval hard-codes a full drift term and prior error retrieval hard-codes none, and neither can be right everywhere, because $\mathbf{g}_{ti}$ captures the trajectory mismatch that can overwrite level, phase, and rhythm. Section 3.2 therefore exposes its weight instead of fixing it.

Chronological error memory. Memory is built from predictions on a held-out validation split rather than from training data, whose residuals are optimistically biased by the fit that produced them. Each item stores the context, base forecast, calendar marks, and residual, and is admitted only after its complete target has elapsed; during calibration an earlier validation prefix supplies memory and a later suffix supplies queries, so parameters are never chosen on the origins that memory explains.

3.2 Model-conditioned Error Retrieval

Model conditioned descriptor. The retrieved items should reflect not only the state of the series, but the state of the forecaster as well. TRACE therefore concatenates normalized shapes, first differences, summary statistics, slopes, and calendar marks over the last $\ell \in \{\min(48, L), L\}$ context steps, optionally augmented with the same features from the base forecast $\widehat{\mathbf{y}}$. Including $\widehat{\mathbf{y}}$ makes the index model-conditioned: the retrieved neighbors vary across backbones as different futures are predicted from the same observed context. High-dimensional series are pooled into at most 16 contiguous channel groups for descriptor construction, and features are standardized with memory statistics and ℓ_2-normalized. This descriptor construction requires no learned encoder, hence one implementation applies unchanged to every forecaster we study.

Neighbor weighting and shrinkage. Let $\mathcal{E}_t$ denote memory items whose complete targets precede origin t, and let $\mathcal{N}_k(t) \subseteq \mathcal{E}_t$ be the top-k items under cosine similarity of descriptors $\boldsymbol{\phi} \in \mathbb{R}^p$. A softmax turns the similarities into weights scaled with temperature τ:

$$s_{ti} = \boldsymbol{\phi}_t^\top \boldsymbol{\phi}_i, \qquad w_{ti} = \frac{\exp((s_{ti} - s_{t,\max})/\tau)}{\sum_{j \in \mathcal{N}_k(t)} \exp((s_{tj} - s_{t,\max})/\tau)}, \qquad \boldsymbol{\mu}_t = \sum_{i \in \mathcal{N}_k(t)} w_{ti}\, \mathbf{r}_i, \tag{2}$$

where $s_{t,\max} = \max_{j \in \mathcal{N}_k(t)} s_{tj}$. The weighted mean $\boldsymbol{\mu}_t$ estimates the recurring error at origin t. The mean is trustworthy only where the retrieved errors agree, and agreement may vary across the horizon and channels. TRACE therefore computes

$$\begin{aligned} \boldsymbol{\sigma}_t^2 &= \sum_{i \in \mathcal{N}_k(t)} w_{ti} \big(\mathbf{r}_i - \boldsymbol{\mu}_t\big)^{\odot 2}, \qquad k_{\text{eff}} = \Big(\sum_{i \in \mathcal{N}_k(t)} w_{ti}^2 \Big)^{-1}, \\ \boldsymbol{\rho}_t &= \frac{\boldsymbol{\mu}_t^{\odot 2}}{\boldsymbol{\mu}_t^{\odot 2} + \lambda \boldsymbol{\sigma}_t^2 / k_{\text{eff}} + \epsilon}, \qquad c_t^{\text{ret}} = \alpha\, \boldsymbol{\rho}_t \odot \boldsymbol{\mu}_t, \qquad \epsilon = 10^{-8}. \end{aligned} \tag{3}$$

Here $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2, \boldsymbol{\rho}_t \in \mathbb{R}^{H \times d}$, and the division and Hadamard operations are elementwise. Components on which neighbors agree keep $\boldsymbol{\rho}_t$ near one; a small mean with large disagreement is attenuated toward zero, returning that component to the original prediction. Scale α and shrinkage strength λ come from validation (Appendix C).

Error-augmented corrections. The same neighborhood and the same weights also aggregate the drift term of Equation 1, which lets a single expression interpolate between the two objects:

$$c_t = \alpha \Big(\boldsymbol{\rho}_t \odot \boldsymbol{\mu}_t \; + \; \beta\, \boldsymbol{h} \odot \textstyle\sum_{i \in \mathcal{N}_k(t)} w_{ti}\, \mathbf{g}_{ti} \Big), \qquad \beta \in [0, 1], \tag{4}$$

where $\boldsymbol{h} \in \mathbb{R}^{H \times d}$ repeats the chosen horizon profile across channels, either flat or rising linearly with lead time. Setting $\beta = 0$ recovers pure error retrieval, and $\beta = 1$ with $\alpha = 1$, a flat profile, and no shrinkage reproduces the endpoint-aligned analog mean that input-retrieval systems emit, so both published objects are literal special cases. Intermediate β is the case neither can express: keep the base trajectory but pull it partway towards the neighborhood's. It costs one extra weighted mean over the same index, and $\mathbf{g}_{ti}$ vanishes on its own when the forecaster is accurate and its neighbors match, which is exactly the regime where overwriting would hurt most.

3.3 Validation-Based Correction Construction

Candidate corrections. Error retrieval rests on a testable assumption: that this forecaster's errors recur in retrievably similar states. Where they do not, the right correction is a simpler one or none at all, and which case holds varies sharply across backbones and datasets. Besides the retrieved correction, the candidate pool therefore holds **identity** ($c_t = 0$), a **seasonal blend** repeating a recent period, an **overlap blend** of earlier forecasts of the same target with age decay, a **static bias** from the mean validation residual, and a **causal rolling bias** over recent elapsed residuals (Appendix C). The last two share error memory with retrieval but discard its state-conditioning, which is what makes the comparison informative.

Selection rule. For the metric set $\mathcal{Q}$ a task reports, candidate policy p is scored by its worst relative validation error on a query window $\mathcal{W}$, $S(p, \mathcal{W}) = \max_{q \in \mathcal{Q}} \mathcal{L}_q(\widetilde{\mathbf{y}}^{(p)}, \mathbf{y}; \mathcal{W}) / \mathcal{L}_q(\widehat{\mathbf{y}}, \mathbf{y}; \mathcal{W})$, so a candidate must not degrade any reported metric to be preferred. Adding the dial multiplies the candidate pool sevenfold, which makes the minimizer of a single score optimistically biased, so TRACE scores each candidate on the worse of the two chronological halves of the query window, $S(p) = \max\{S(p, \mathcal{W}_1), S(p, \mathcal{W}_2)\}$: a correction that helps early and hurts late can no longer be selected. The minimizing family and its parameters are locked and applied once to test; ties go to identity. Because the identity candidate scores exactly one, the gate can always refuse to act.

Nothing in the rule privileges one retrieval object: β is chosen exactly as α, k, or a seasonal period is, by whichever candidate survives both halves of the query window. Section 4.2 reports where that leaves the dial and what a horizon-dependent prior on it would add.

Table 1: **Long-term forecasting.** Each entry averages 13 backbones and horizons $\{96, 192, 336, 720\}$ over the *same* 364 base forecasts; ANALOG-KNN and RESIDUAL-KNN pin Equation 4 at $\beta = 1$ and $\beta = 0$. Lower is better; **best** and second best per column group.

Dataset	Uncorrected	Retrieve inputs / futures			Retrieve errors (ours)	
MSE / MAE	Base	ANALOG-KNN	RAFT	SARAF	RESIDUAL-KNN	TRACE
ETTh1	0.529 0.497	0.508 0.487	0.502 0.486	0.501 0.485	0.519 0.494	**0.481 0.471**
ETTh2	0.597 0.515	0.431 0.438	0.441 0.442	0.443 0.444	0.544 0.493	**0.427 0.436**
ETTm1	0.481 0.453	0.442 0.437	0.455 0.441	0.455 0.441	0.447 0.440	**0.428 0.431**
ETTm2	0.385 0.394	0.295 0.344	0.305 0.352	0.305 0.352	0.364 0.383	**0.291 0.342**
Weather	0.321 0.338	0.286 0.319	0.292 0.322	0.293 0.323	0.289 0.324	**0.272 0.310**
Electricity	0.242 0.329	0.207 0.305	0.213 0.308	0.212 0.308	0.216 0.310	**0.205 0.301**
Traffic	0.597 0.361	0.574 0.354	0.579 0.356	0.575 0.356	**0.550** 0.343	0.556 **0.342**
Average	0.450 0.413	0.392 0.383	0.398 0.387	0.398 0.387	0.418 0.398	**0.380 0.376**

4 EXPERIMENTS

We evaluate TRACE on long-term and short-term forecasting over 16 real-world benchmarks and 13 base forecasting models. Section 4.1 reports the main results, and Section 4.2 analyzes which correction each deployment ends up using. Full configurations in Appendix C.

Datasets. For long-term forecasting, we use seven benchmarks at horizons $\{96, 192, 336, 720\}$, including four ETT subsets, Weather, Electricity, and Traffic (Wang et al., 2024c). For short-term forecasting, we use the PEMS traffic networks datasets (Chen et al., 2001) at horizons $\{6, 12, 24\}$ and the EPF electricity markets datasets (Lago et al., 2021) at horizon 24. Across all datasets, models, and prediction horizons, we evaluate 585 settings in total: 364 long-term, 156 PEMS, and 65 EPF (Appendix Table 5).

Base models and setup. We correct 13 forecasters implemented in `TSLib` (Wang et al., 2024c) with their default configurations: TIMEXER, TIMEMIXER, PATTN, ITRANSFORMER, TIMESNET, PATCHTST, DLINEAR, FRETS, FEDFORMER, NON-STATIONARY TRANSFORMER, LIGHTTS, INFORMER, and AUTOFORMER (Wang et al., 2024d;b; Tan et al., 2024; Liu et al., 2024b; Wu et al., 2023; Nie et al., 2023a; Zeng et al., 2022; Yi et al., 2023; Zhou et al., 2022a; Liu et al., 2022; Zhang et al., 2022; Zhou et al., 2021; Wu et al., 2022). Each backbone is trained once and then kept fixed. For each setting, validation predictions and labels are used to build the error memory and select one correction policy, which is locked before test evaluation. Test labels never enter the selection process (Appendix D). All methods use the same base forecasters, validation split, memory boundary, and evaluation rule with fixed hyperparameters (Appendix A). Long-term and EPF settings report normalized MSE and MAE, while PEMS reports inverse-scaled MAE, RMSE, and MAPE. A *strict win* requires improvements on every reported metric. We report both mean and median paired improvements as baseline error levels vary across settings.

Compared systems. Equation 4 lets us compare retrieval objects rather than implementations. ANALOG-KNN pins it at $\beta = 1$, injecting from the same neighborhood and descriptor the endpoint-aligned futures that input-retrieval systems emit, and RESIDUAL-KNN pins it at $\beta = 0$. RAFT (Han et al., 2025) and SARAF (Zhou et al., 2026) contribute two published operators on the same fixed forecasts, while TRACE chooses α, β, and the horizon profile per setting on validation.

4.1 MAIN RESULTS

Long-term forecasting. As shown in Table 1, TRACE reduces the average MSE from 0.450 to 0.380 and the average MAE from 0.413 to 0.376 across all long-term forecasting settings. It achieves the lowest MSE and MAE on six of the seven datasets, except Traffic where RESIDUAL-KNN is lower on MSE. Compared with ANALOG-KNN, TRACE further reduces average MSE from 0.392 to 0.380 and MAE from 0.383 to 0.376.

Table 2: **Short-term forecasting.** Same protocol and columns on the PEMS networks (13 backbones, horizons $\{6, 12, 24\}$) and the five EPF markets (horizon 24); PEMS RMSE is in Appendix A. Lower is better; **best** and second best per column group.

Dataset	Uncorrected	Retrieve inputs / futures			Retrieve errors (ours)	
Metrics	**Base**	ANALOG-KNN	RAFT	SARAF	RESIDUAL-KNN	TRACE
PEMS traffic flow, MAE / MAPE, inverse-scaled (156 settings)						
PEMS03	20.781 0.208	17.097 0.168	19.855 0.196	19.886 0.196	18.224 0.182	**16.275 0.160**
PEMS04	26.982 0.189	22.541 0.149	25.911 0.179	25.972 0.180	23.664 0.161	**21.467 0.140**
PEMS07	30.087 0.139	24.210 0.104	28.865 0.133	28.940 0.133	26.224 0.118	**23.006 0.098**
PEMS08	23.086 0.148	18.069 0.114	22.005 0.139	22.049 0.139	19.654 0.127	**16.878 0.108**
Average	25.234 0.171	20.479 0.134	24.159 0.162	24.212 0.162	21.942 0.147	**19.407 0.126**
EPF electricity price, MSE / MAE (65 settings)						
NP	0.314 0.328	0.295 0.311	0.294 0.311	0.291 0.309	0.301 0.320	**0.284 0.303**
PJM	0.121 0.224	0.110 0.216	0.110 0.216	0.109 0.216	0.114 0.216	**0.103 0.207**
BE	0.453 0.305	0.451 0.296	0.447 0.298	0.448 0.300	0.425 0.294	**0.425 0.284**
FR	0.465 0.258	0.469 0.242	0.455 0.248	0.457 0.246	0.453 0.244	**0.450 0.225**
DE	0.526 0.459	0.506 0.449	0.506 0.449	0.505 0.449	0.520 0.451	**0.481 0.433**
Average	0.376 0.315	0.366 0.303	0.362 0.304	0.362 0.304	0.363 0.305	**0.349 0.290**

Short-term forecasting. Table 2 shows a wider margin on shorter horizons, where TRACE performs best on both reported metrics for all nine datasets. On PEMS, TRACE reduces the average MAE from 25.234 to 19.407 and MAPE from 0.171 to 0.126, with median paired reductions of 17.77% MAE, 15.95% RMSE, and 20.14% MAPE. On EPF, it reduces the average MSE from 0.376 to 0.349 and MAE from 0.315 to 0.290, with median paired reductions of 4.57% and 5.44%. In comparison, RAFT and SARAF provide little improvement on PEMS. Their input-based retrieval mainly returns traffic patterns that have been captured by the backbone, resulting in near-zero median gains.

Performance across settings. Across the 585 settings, TRACE improves all reported metrics in 541 cases over the backbone forecast, including 326 out of 364 long-term settings, all 156 short-term traffic settings, and 59 out of 65 short-term electricity settings. The corresponding counts are 490 for RESIDUAL-KNN, 478 for ANALOG-KNN, 384 for SARAF, and 375 for RAFT. SARAF and RAFT return forecasts identical to the base in 121 and 131 settings, respectively, compared with 50 for RESIDUAL-KNN and 3 for TRACE. In direct setting-wise comparisons, TRACE improves every reported metric over ANALOG-KNN in 453 settings, versus 54 in the opposite direction, and over RESIDUAL-KNN in 457 settings, versus 63 in the opposite direction. Among the five most accurate backbones, TRACE achieves strict improvements in 190/225 settings, compared with 170 for RESIDUAL-KNN, 157 for ANALOG-KNN, 80 for SARAF, and 77 for RAFT.

End-to-end trained forecasters. The comparison above holds base forecasts fixed to isolate the retrieval object. On the separate primary matrix, where each backbone is trained and corrected end to end, the error-only form of TRACE—Equation 4 at $\beta \equiv 0$ with single-window selection—improves every reported metric in 562/585 settings (96.1%): 355/364 long-term, 151/156 PEMS, and 56/65 EPF, with mean paired reductions of 8.5% MSE, 7.5% MAE, 10.6% RMSE, and 12.0% MAPE. Gains scale with how much error there is to remove, from 1.1% long-term MSE for TIMEXER to 41.3% for INFORMER, as an additive layer should, and no backbone is left behind (Appendix Table 17). The selection rule was fixed on a predetermined 123-setting development subset, and on the 462 held-out settings the strict-win rate is 437/462 (Appendix E). A locked correction also survives the passage of time: replaying TIMEMIXER at horizon 96 on a timeline beginning after the standard test split, with labels released only as their targets elapse, improves both metrics on all four ETT datasets and three seeds (Appendix H).

Ensembles, AutoML, and foundation models. Because TRACE consumes only predictions and targets, the same interface corrects systems that are not single models. Appendix Table 34 reports 81/94 strict wins over greedy forward selection, a portfolio and a zero-shot ensemble, AUTOGLUON's



Figure 3: **Retrieval object gains vs. forecasters.** RESIDUAL-KNN minus ANALOG-KNN median worst-metric gain (percentage points, symmetric-log axis) per backbone, ordered from most to least accurate by long-term base MSE.

`high_quality` predictor, and CHRONOS-BOLT in fine-tuned and zero-shot form (Caruana et al., 2004; Feurer et al., 2022; Shchur et al., 2023; Ansari et al., 2024b), from 12/16 to 16/16 per system, with the retrieval family selected in 70 settings and 65 of those strict wins: multi-component systems and an off-distribution foundation model still leave error structure their own machinery does not remove.

4.2 FURTHER ANALYSIS

Retrieval object and backbone strength. Ordering the 13 backbones by long-term base MSE, the advantage of retrieving errors over retrieving futures falls as base error rises in both task families, with Spearman ρ of -0.94 and -0.77 and a sign change at TIMESNET (Figure 3). The mechanism is the drift term of Equation 1: a weak forecaster's prediction holds little level, phase, or rhythm worth keeping, so replacing it with an analogous future is an improvement, whereas an accurate forecaster's prediction already holds all of that and only its residual is worth transferring. INFORMER, whose long-term base MSE is three times the best backbone's, is what pulls the ETTh2 and ETTm2 averages towards ANALOG-KNN.

Retrieval object and forecast horizon. Validation puts non-zero weight on the drift term in 435 of 585 settings, and where it declines with the horizon: 50/52, 50/52, and 112/117 settings at horizons 6, 12, and 24, then 69/91, 65/91, 64/91, and 25/91 at 96 through 720. A neighbor's trajectory stays usable for a few hours of traffic flow and is least usable a month ahead, where the weight survives in barely a quarter of settings. It is also rarely extreme, as 299 of those 435 settings use $\beta \in \{0.25, 0.5\}$ rather than the full overwrite input retrieval performs, so the selected operating point is one neither published object can express. Retrieval of either kind is selected in 449 of the 585 settings, taking all 156 PEMS and 63 of 65 EPF settings and 230 of 364 long-term ones, where a seasonal blend takes 75 and an overlap blend 46 (Appendix Figure 5).

Robustness-gain tradeoff. Scoring candidates on the worse of the two chronological validation halves instead of the whole window raises strict wins from 533 to 552 with the dial pinned shut. Opening the dial then trades 11 of those wins for magnitude everywhere: the median worst-metric gain rises from 9.73% to 15.69% on PEMS, from 1.77% to 3.56% on EPF, and from 3.64% to 4.17% in long-term MSE. Restricting $\beta > 0$ to horizons of at most 24 steps unless it clears the best $\beta = 0$ candidate by 0.05 recovers those 11 wins, reaching 555/585, but gives back the long-term magnitude (median MSE 3.64% again). We report the free dial, because the choice between per-setting accuracy and per-setting win rate is a deployment preference rather than a property of the method.

Failure modes of correction. Of the 585 settings, 44 are not strict wins: 38 long-term and 6 EPF, none on PEMS, concentrated on ETTm1 (17), ETTh1 (8), ETTh2 (7), and EPF-FR (6), and on the most accurate backbones, where the base trajectory is the thing worth protecting. Only 7 degrade every metric and 3 leave the forecast untouched; the other 34 trade one metric for another, the largest single loss being 6.9% MSE on an EPF-FR setting that gains 16.4% MAE. Gains can also be thin, with the worst metric improving by at most one point in 68 of the 541 strict wins, and selecting on a single validation window fails harder: 15 of its 52 non-strict settings degrade every metric.

5 Related Work

Time-series forecasting. Modern time-series forecasting spans several modeling families. Transformer-based methods (Wu et al., 2022; Zhou et al., 2022b) were developed to capture long-horizon temporal patterns and separate trend and seasonal variations. Later works reconsider how temporal and multivariate time series should be represented. (Nie et al., 2023a) patches time series as tokens, while (Liu et al., 2024b) treats each channel as token to capture cross-channel dependencies. Simpler forecasting architectures (Zeng et al., 2022; Yi et al., 2023; Wang et al., 2024b) have also been explored. These methods replace attention mechanisms with linear prediction over decomposed trend and seasonal components, frequency-domain MLPs, and multiscale modeling that separates and combines patterns at different temporal resolutions. More recently, large-scale pretraining has further enabled zero-shot forecasters (Ansari et al., 2024a; Das et al., 2024; Liu et al., 2024c; Woo et al., 2024). On top of these forecasters, TRACE serves as a model agnostic post-hoc correction framework that uses their historical forecast errors without changing their backbone.

Post-hoc Forecast Refinement and Error Correction. Post-hoc methods improve existing forecasts without correcting the forecasting backbone. Early efforts (Mouatadid et al., 2023; Jobst et al., 2024) focused on correcting systematic bias and modeling structured forecast errors. Recent works extend post-hoc correction to neural forecasters. One direction identifies forecasts that need correction and adjusts them using contextual information. For example, (Liu et al., 2025) uses local covariates and globally retrieved historical time series instances to revise selected forecasts. Another direction trains a separate correction module on top of a frozen forecaster. (Liang et al., 2026) introduces a lightweight adapter, a shallow MLP or low-rank head, for residual correction. (Nguyen et al., 2026) decomposes the base forecast into trend and seasonal components and learns to correct them separately. Our method, instead, treats realized historical forecast errors as retrieval memory and uses them to construct corrections for similar forecasting situations.

6 Conclusion

We asked what a retrieval-augmented time-series forecaster should retrieve, and answered that it should retrieve its own errors—and that the futures its predecessors retrieve are the same correction with one term forced on. TRACE realizes that answer as a post-hoc layer over a chronological memory of a forecaster's validation residuals, with shrinkage, a dial over the retrieval object, and a gate that can leave the forecast untouched. With all 585 base forecasts fixed it reaches 541 strict wins against 490 for pure error retrieval and 478 for the analog control, and takes the best average in every task family.

Scope and limitations. Error memory pays off when a forecaster's mistakes recur and borrowing a neighbor's trajectory pays off when they do not; Figure 3 locates that crossover by base accuracy and Section 4.2 by horizon. Leaving the dial free costs 11 strict wins against pinning it shut, and a horizon-dependent prior on it recovers those wins while giving back most of the long-term magnitude, so which of the two to deploy is a choice we report rather than settle. The end-to-end and additional-system results use the error-only configuration, so what the dial adds is measured only with base forecasts fixed. The primary matrix uses one seed and the prequential study one backbone; descriptor, pooling, and shrinkage ablations remain open.

AI use statement

Generative AI tools assisted with literature discovery, code, plotting, and manuscript editing. The authors verified all citations, analyses, experimental measurements, and final text.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon

Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024a. URL https://arxiv.org/abs/2403.07815.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024b.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 18, 2004.

Chao Chen, Karl F. Petty, Alexander Skabardonis, Pravin Pratap Varaiya, and Zhanfeng Jia. Freeway performance measurement system: Mining loop detector data. *Transportation Research Record*, 2001.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. Auto-sklearn 2.0: Hands-free automl via meta-learning. *Journal of Machine Learning Research*, 23 (261):1–61, 2022. URL http://jmlr.org/papers/v23/21-0992.html.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2505.04163.

David Jobst, Annette Möller, and Jürgen Groß. Time series based ensemble model output statistics for temperature forecasts postprocessing, 2024. URL https://arxiv.org/abs/2402.00555.

Junhyeok Kang, Jun Seo, Soyeon Park, Sangjun Han, Seohui Bae, Hyeokjun Choe, and Soonyoung Lee. Channel-wise retrieval for multivariate time series forecasting. *arXiv preprint arXiv:2604.05543*, 2026. URL https://arxiv.org/abs/2604.05543.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

Seunghan Lee, Jaehoon Lee, Jun Seo, Sungdong Yoo, Minjae Kim, Tae Yoon Lim, Dongwan Kang, Hwanil Choi, SoonYoung Lee, and Wonbin Ahn. Not all retrievals are useful: Cross-attention for input-aware RAG in time series forecasting. *arXiv preprint arXiv:2603.14709*, 2026. URL https://arxiv.org/abs/2603.14709.

Daojun Liang, Qi Li, Yinglong Wang, Jing Chen, Hu Zhang, Xiaoxiao Cui, Qizheng Wang, and Shuo Li. The forecast after the forecast: A post-processing shift in time series, 2026. URL https://arxiv.org/abs/2601.20280.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Rethinking the stationarity in time series forecasting. *NeurIPS*, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting, 2024b. URL https://arxiv.org/abs/2310.06625.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models, 2024c. URL https://arxiv.org/abs/2402.02368.

Zhiding Liu, Mingyue Cheng, Guanhao Zhao, Jiqian Yang, Qi Liu, and Enhong Chen. Improving time series forecasting via instance-aware post-hoc revision, 2025. URL https://arxiv.org/abs/2505.23583.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Judah Cohen, Miruna Oprescu, Ernest Fraenkel, and Lester Mackey. Adaptive bias correction for improved subseasonal forecasting. *Nature Communications*, 14(1), 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38874-y. URL http://dx.doi.org/10.1038/s41467-023-38874-y.

Minh Hoang Nguyen, Dai Do, Huu Hiep Nguyen, Dung Nguyen, Kien Do, and Hung Le. Reviving error correction in modern deep time-series forecasting, 2026. URL https://arxiv.org/abs/2605.21088.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023a. URL https://arxiv.org/abs/2211.14730.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *ICLR*, 2023b.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Yiming Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. TS-RAG: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecasters. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.07649.

Oleksandr Shchur, Ali Caner Turkmen, Nick Erickson, Huibin Shen, Alexander Shirkov, Tony Hu, and Bernie Wang. Autogluon–timeseries: Automl for probabilistic time series forecasting. In *International Conference on Automated Machine Learning*, pp. 9–1. PMLR, 2023.

Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Kutay Tire, Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2411.08249*, 2024. URL https://arxiv.org/abs/2411.08249.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a.

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, 2024b. URL https://arxiv.org/abs/2405.14616.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*, 2024c.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *arXiv preprint arXiv:2402.19072*, 2024d.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, 2022. URL https://arxiv.org/abs/2106.13008.

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

Sitan Yang, Carson Eisenach, and Dhruv Madeka. MQRetNN: Multi-horizon time series forecasting with retrieval augmentation. In *8th SIGKDD International Workshop on Mining and Learning from Time Series*, 2022. URL https://arxiv.org/abs/2207.10517.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Defu Lian, Ning An, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting, 2023. URL https://arxiv.org/abs/2311.06184.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting?, 2022. URL https://arxiv.org/abs/2205.13504.

Huanyu Zhang, Chang Xu, Yi-Fan Zhang, Zhang Zhang, Liang Wang, Tieniu Tan, and Jiang Bian. TimeRAF: Retrieval-augmented foundation model for zero-shot time series forecasting. *arXiv preprint arXiv:2412.20810*, 2024. URL https://arxiv.org/abs/2412.20810.

Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures. *arXiv preprint arXiv:2207.01186*, 2022.

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.

Shiqiao Zhou, Holger Schöner, Zipeng Wu, Edouard Fouché, IAG Wilson, and Shuo Wang. Stationarity-aware retrieval-augmented time series forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3817813.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. *ICML*, 2022a.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting, 2022b. URL https://arxiv.org/abs/2201.12740.

Appendix Contents

The appendix follows the paper's questions rather than the order in which experiments were run. Each entry below states what its evidence is for.

Section	Purpose
Appendix A, p. 12	Matched fixed-forecast retrieval controls and per-backbone values
Appendix B, p. 15	Datasets, models, training scope, and hardware
Appendix C, p. 16	Exact validation search space and selection rule
Appendix D, p. 18	Information boundaries and no-lookahead checks
Appendix E, p. 19	Policy attribution, failures, and hardware sensitivity
Appendix F, p. 19	Primary-matrix results by horizon, backbone, and cell
Appendix G, p. 32	Ensemble, AutoML, and foundation-model results
Appendix H, p. 33	Three-seed forward-transfer study on later observations

A Retrieval comparisons and controls

The matched fixed-forecast comparison spans the same 585 backbone–dataset–horizon cells for every method. RAFT uses multiscale Pearson retrieval and softmax aggregation (Han et al., 2025); SARAF adds time alignment, stationarity-controlled relevance–diversity selection, and Gaussian aggregation (Zhou et al., 2026). These rows evaluate retrieval operators on common base forecasts, not the original learned forecasting or fusion layers.

ANALOG-KNN transfers a normalized historical future, whereas RESIDUAL-KNN transfers the corresponding error from the same base model; in the notation of Equation 4 they pin $\beta = 1$ and

Table 3: **Long-term fixed-forecast comparison.** Mean over all 7 datasets and 4 horizons for each base model. $\Delta = \text{method} - \text{Base}$; lower is better. Bold marks the lowest error and corresponding delta.

System	TXER	TMIX	PATTN	ITRANS.	TNET	PTST	DLIN	FRETS	FEDF.	NONSTAT	LTS	INF.	AUTOF
Long-term, MSE (all datasets and horizons)													
Base	0.335	0.352	0.360	0.356	0.380	0.350	0.421	0.405	0.422	0.521	0.452	1.053	0.447
RAFT	0.334	0.350	0.358	0.353	0.373	0.348	0.400	0.390	0.405	0.455	0.417	0.576	0.416
Δ	−0.001	−0.002	−0.003	−0.003	−0.007	−0.002	−0.021	−0.015	−0.018	−0.066	−0.035	−0.476	−0.031
SARAF	0.334	0.349	0.357	0.353	0.372	0.347	0.399	0.390	0.404	0.455	0.416	0.581	0.416
Δ	−0.001	−0.003	−0.003	−0.003	−0.008	−0.003	−0.022	−0.015	−0.019	−0.066	−0.036	−0.472	−0.031
TRACE	**0.331**	**0.346**	**0.352**	**0.345**	**0.363**	**0.343**	**0.375**	**0.376**	**0.382**	**0.421**	**0.397**	**0.511**	**0.397**
Δ	**−0.004**	**−0.006**	**−0.008**	**−0.011**	**−0.017**	**−0.007**	**−0.046**	**−0.029**	**−0.041**	**−0.100**	**−0.055**	**−0.542**	**−0.050**
Long-term, MAE (all datasets and horizons)													
Base	0.341	0.349	0.356	0.356	0.363	0.354	0.406	0.393	0.409	0.448	0.429	0.732	0.428
RAFT	0.341	0.348	0.355	0.356	0.362	0.353	0.394	0.385	0.399	0.412	0.407	0.509	0.409
Δ	0.000	0.000	0.000	−0.001	−0.002	−0.001	−0.012	−0.008	−0.010	−0.036	−0.023	−0.223	−0.018
SARAF	0.341	0.348	0.355	0.355	0.361	0.353	0.394	0.385	0.399	0.412	0.407	0.513	0.410
Δ	0.000	0.000	0.000	−0.001	−0.002	−0.001	−0.012	−0.008	−0.010	−0.036	−0.023	−0.219	−0.018
TRACE	**0.339**	**0.346**	**0.352**	**0.350**	**0.356**	**0.350**	**0.377**	**0.376**	**0.384**	**0.396**	**0.391**	**0.477**	**0.395**
Δ	**−0.001**	**−0.003**	**−0.003**	**−0.006**	**−0.007**	**−0.003**	**−0.028**	**−0.018**	**−0.025**	**−0.052**	**−0.038**	**−0.255**	**−0.033**

$\beta = 0$. The full TRACE row selects β, the horizon profile, and the alternative simple corrections on validation. All methods select parameters on the same chronological validation suffix; retrieval memory ends at least one horizon earlier, and prediction APIs receive no test labels. The TRACE row was produced by a later run than the comparison systems; its error-only configuration reproduces the earlier run's selected candidate in all 585 cells and every strict-win classification, with a maximum relative metric difference of 1.6×10^{-5} from thread-level float reordering, so the rows remain paired cell by cell. Tables 3–4 give the per-backbone absolute errors and paired deltas behind main Tables 1 and 2, including the PEMS RMSE column omitted there. The accompanying machine-readable results retain all 585 per-cell records, so duplicating them as 90 multi-page tables is unnecessary.

Table 4: **PEMS and EPF fixed-forecast comparison.** Dataset–horizon mean by base model within each task. $\Delta = \text{method} - \text{Base}$; lower is better. Bold marks the lowest error and corresponding delta.

System	TXer	TMix	PAttn	iTrans.	TNet	PTST	DLin	FreTS	FEDf.	NonStat	LTS	Inf.	AutoF
PEMS, MAE (all datasets and horizons)													
Base	23.27	18.32	24.51	22.17	22.19	25.71	30.23	22.83	31.00	21.63	23.71	23.70	38.76
RAFT	23.27	18.32	24.51	22.17	21.80	25.72	29.78	22.83	27.25	21.00	23.16	22.47	31.80
Δ	0.00	0.00	0.00	0.00	−0.40	+0.01	−0.45	0.00	−3.76	−0.63	−0.55	−1.23	−6.96
SARAF	23.27	18.32	24.51	22.17	21.84	25.71	29.86	22.83	27.33	21.06	23.21	22.57	32.08
Δ	0.00	0.00	0.00	0.00	−0.36	0.00	−0.37	0.00	−3.67	−0.57	−0.50	−1.14	−6.68
TRACE	**18.86**	**17.68**	**19.77**	**19.00**	**19.06**	**19.43**	**18.82**	**19.10**	**20.56**	**18.98**	**19.50**	**19.37**	**22.16**
Δ	**−4.41**	**−0.64**	**−4.74**	**−3.17**	**−3.14**	**−6.29**	**−11.41**	**−3.73**	**−10.44**	**−2.65**	**−4.21**	**−4.33**	**−16.60**
PEMS, RMSE (all datasets and horizons)													
Base	35.00	28.85	38.26	34.99	34.81	39.22	44.86	35.56	44.85	34.12	35.73	37.61	55.18
RAFT	35.00	28.85	38.26	34.98	33.87	39.22	44.31	35.56	39.78	32.88	35.04	35.06	45.50
Δ	0.00	0.00	0.00	−0.01	−0.94	+0.01	−0.55	0.00	−5.07	−1.24	−0.69	−2.55	−9.68
SARAF	35.00	28.85	38.26	34.98	33.97	39.22	44.38	35.56	39.88	32.99	35.09	35.30	45.85
Δ	0.00	0.00	0.00	−0.01	−0.85	0.00	−0.48	0.00	−4.97	−1.13	−0.64	−2.31	−9.32
TRACE	**29.92**	**28.13**	**31.56**	**30.41**	**30.12**	**31.05**	**30.04**	**30.53**	**31.79**	**30.06**	**30.55**	**30.70**	**33.59**
Δ	**−5.08**	**−0.72**	**−6.70**	**−4.57**	**−4.70**	**−8.16**	**−14.81**	**−5.03**	**−13.06**	**−4.06**	**−5.18**	**−6.91**	**−21.59**
PEMS, MAPE (all datasets and horizons)													
Base	0.157	0.121	0.154	0.141	0.146	0.165	0.219	0.147	0.228	0.142	0.162	0.157	0.284
RAFT	0.157	0.121	0.154	0.141	0.143	0.165	0.212	0.147	0.196	0.137	0.155	0.148	0.225
Δ	0.000	0.000	0.000	0.000	−0.003	0.000	−0.007	0.000	−0.032	−0.005	−0.006	−0.009	−0.059
SARAF	0.157	0.121	0.154	0.141	0.143	0.165	0.214	0.147	0.197	0.137	0.156	0.148	0.227
Δ	0.000	0.000	0.000	0.000	−0.003	0.000	−0.005	0.000	−0.031	−0.005	−0.006	−0.008	−0.057
TRACE	**0.121**	**0.115**	**0.126**	**0.120**	**0.124**	**0.124**	**0.121**	**0.123**	**0.137**	**0.122**	**0.129**	**0.125**	**0.153**
Δ	**−0.036**	**−0.006**	**−0.028**	**−0.021**	**−0.023**	**−0.041**	**−0.098**	**−0.024**	**−0.090**	**−0.020**	**−0.033**	**−0.031**	**−0.131**
EPF, MSE (all datasets and horizons)													
Base	0.317	0.339	0.343	0.331	0.327	0.341	0.358	0.362	0.397	0.544	0.363	0.398	0.469
RAFT	0.318	0.337	0.338	0.327	0.322	0.340	0.356	0.357	0.371	0.487	0.359	0.379	0.419
Δ	0.000	−0.002	−0.005	−0.003	−0.005	−0.001	−0.001	−0.005	−0.027	−0.057	−0.004	−0.019	−0.050
SARAF	0.318	0.337	0.336	0.331	0.321	0.339	0.355	0.355	0.369	0.491	0.357	0.377	0.420
Δ	+0.001	−0.002	−0.007	+0.001	−0.006	−0.002	−0.003	−0.007	−0.029	−0.053	−0.006	−0.021	−0.049
TRACE	**0.314**	**0.332**	**0.328**	**0.316**	**0.313**	**0.335**	**0.353**	**0.347**	**0.358**	**0.434**	**0.347**	**0.361**	**0.396**
Δ	**−0.004**	**−0.007**	**−0.015**	**−0.014**	**−0.014**	**−0.006**	**−0.005**	**−0.015**	**−0.039**	**−0.110**	**−0.015**	**−0.037**	**−0.073**
EPF, MAE (all datasets and horizons)													
Base	0.269	0.294	0.297	0.286	0.286	0.287	0.310	0.312	0.344	0.412	0.305	0.328	0.364
RAFT	0.268	0.291	0.293	0.284	0.282	0.286	0.307	0.306	0.317	0.372	0.302	0.314	0.335
Δ	−0.001	−0.003	−0.004	−0.003	−0.003	−0.001	−0.002	−0.006	−0.027	−0.040	−0.003	−0.014	−0.029
SARAF	0.268	0.291	0.292	0.287	0.282	0.285	0.306	0.305	0.317	0.368	0.302	0.315	0.336
Δ	−0.001	−0.003	−0.005	0.000	−0.004	−0.002	−0.004	−0.007	−0.027	−0.043	−0.004	−0.013	−0.028
TRACE	**0.263**	**0.283**	**0.283**	**0.275**	**0.276**	**0.280**	**0.289**	**0.295**	**0.298**	**0.328**	**0.292**	**0.301**	**0.311**
Δ	**−0.005**	**−0.011**	**−0.014**	**−0.011**	**−0.010**	**−0.007**	**−0.021**	**−0.017**	**−0.046**	**−0.084**	**−0.013**	**−0.026**	**−0.053**

B DATASET AND MODEL SCOPE

Table 5: Full benchmark scope. "Cells" includes all 13 base models.

Family	Dataset	Variates	Input	Horizons	Cells	Frequency
Long-term	ETTh1	7	96	96,192,336,720	52	1 hour
	ETTh2	7	96	96,192,336,720	52	1 hour
	ETTm1	7	96	96,192,336,720	52	15 min
	ETTm2	7	96	96,192,336,720	52	15 min
	Weather	21	96	96,192,336,720	52	10 min
	Electricity	321	96	96,192,336,720	52	1 hour
	Traffic	862	96	96,192,336,720	52	1 hour
PEMS	PEMS03	358	96	6,12,24	39	5 min
	PEMS04	307	96	6,12,24	39	5 min
	PEMS07	883	96	6,12,24	39	5 min
	PEMS08	170	96	6,12,24	39	5 min
EPF	NP	1	168	24	13	1 hour
	PJM	1	168	24	13	1 hour
	BE	1	168	24	13	1 hour
	FR	1	168	24	13	1 hour
	DE	1	168	24	13	1 hour
				Total	585	

The model list is TIMEXER, TIMEMIXER, PATTN, ITRANSFORMER, TIMESNET, PATCHTST, DLINEAR, FRETS, FEDFORMER, NON-STATIONARY TRANSFORMER, LIGHTTS, INFORMER, and AUTOFORMER. Baseline training follows each paper configuration in TSLib. Shared defaults are Adam optimization, MSE training loss, at most 10 epochs, and validation-loss early stopping unless the referenced configuration overrides them. NON-STATIONARY TRANSFORMER uses a learning rate of 10^{-4} on Electricity horizons 192 and 336; all other values follow the reference configuration.

The primary matrix trains all 585 model–dataset–horizon cells under these configurations with seed 2021, PyTorch 2.8.0, and CUDA 12.9 on one `ml.p4de.24xlarge` host with eight NVIDIA A100 GPUs. Independent cells use one GPU each. A secondary execution on an `ml.g5.12xlarge` host uses 55 compatible released checkpoints and trains the remaining 530 cells on four NVIDIA A10G GPUs. It is retained only as a cross-hardware sensitivity check and is not averaged into the primary tables because the two executions are not numerically identical. The additional-system study uses AUTOGLUON-TimeSeries 1.4.0, Python 3.12, PyTorch 2.7.1, and a three-hour fit budget for each AUTOGLUON cell.

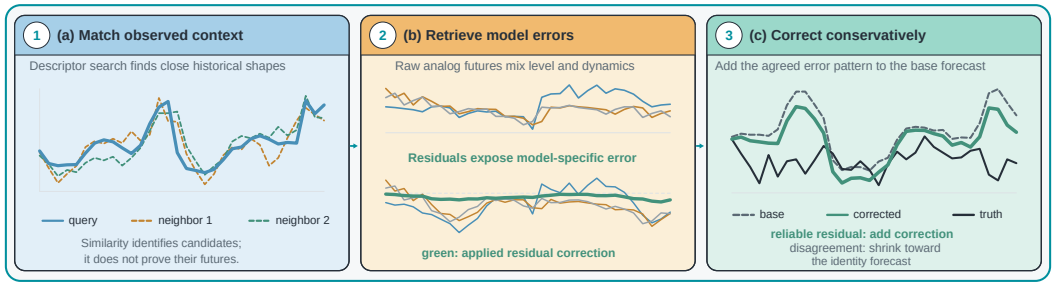

Figure 4: **Residuals isolate model-specific error on real data.** An ETTh1 MULL example from a trained TIMEMIXER. (a) Descriptor search returns historical windows with closely matching shapes. (b) Their raw futures sit at different levels and mix dynamics the base forecast already captures, whereas their residuals agree on a common error pattern. (c) Adding the shrunk residual moves the forecast toward the truth without disturbing its level. Panels (b)–(c) show a 36-step excerpt in the model's standardized space.

Algorithm 1: TRACE calibration and forecast-time correction

Input: trained forecaster f; validation stream $\{(\mathbf{x}_i, \mathbf{y}_i)\}$; candidate families and grids; metric set $\mathcal{Q}$.
Calibration. (1) For every validation origin i, evaluate $\widehat{\mathbf{y}}_i = f(\mathbf{x}_i)$ and, once $\mathbf{y}_i$ has fully elapsed, store $(\boldsymbol{\phi}_i, \widehat{\mathbf{y}}_i, \mathbf{r}_i = \mathbf{y}_i - \widehat{\mathbf{y}}_i)$ in $\mathcal{M}$. (2) Split $\mathcal{M}$ chronologically into a memory prefix and a query suffix. (3) For each candidate c, form $\widetilde{\mathbf{y}}^{(c)}$ on the query suffix using only the memory prefix and score it by $S(c)$. (4) Lock the minimizing family and parameters $c^\star$; ties go to identity.
Forecasting at origin t**.** (5) Evaluate $\widehat{\mathbf{y}}_t = f(\mathbf{x}_t)$ and build $\boldsymbol{\phi}_t$ from $\mathbf{x}_t$ and $\widehat{\mathbf{y}}_t$. (6) If $c^\star$ is retrieval, take the top-k items of $\mathcal{M}_t$ with $i + H < t$, aggregate by Equation 2 and shrink by Equation 3; otherwise evaluate the locked simpler correction. (7) Return $\widetilde{\mathbf{y}}_t = \widehat{\mathbf{y}}_t + c_t$, leaving f unchanged.

C EXACT BROAD SEARCH SPACE

The ETTh1 example in Figure 4 was screened deterministically for visualization only; the screening criteria and score weights are recorded in the released case artifact and never enter method selection.

Table 6: Validation-only candidate space for the broad selector.

Family	Candidate values
Identity	No parameters
Residual descriptor	Tail $\ell \in \{\min(48, L), L\}$; pooled steps $\min(12, \ell)$; differences on; base forecast on/off; calendar on except PEMS; at most 16 pooled channels
Residual aggregation	$(k, \tau, \lambda) \in \{(8, 0.1, 1), (16, 0.3, 4)\}$; $\alpha \in \{.05, .1, .2, .3, .5\}$
Seasonal blend	Long-term periods $\{4, 8, 12, 24, 48, 96, 168\}$, PEMS periods $\{12, 24, 48, 96\}$, and EPF periods $\{24, 48, 168\}$, each clipped at L and deduplicated; $\alpha \in \{.02, .05, .1, .2, .3\}$
Overlap blend	$(a, \gamma) \in \{(8, .75), (24, .9)\}$; $\alpha \in \{.1, .25, .5\}$
Static residual bias	$\alpha \in \{.1, .25, .5, 1\}$
Causal rolling bias	Window $W \in \{2H, 8H, 32H\}$; horizon block $b \in \{0, \min(24, H)\}$; $\alpha \in \{.1, .25, .5, 1\}$

Temporal-candidate parameters. For overlap blend, a is the maximum age in forecast origins and γ is the per-origin exponential decay. For causal rolling bias, W is the target-time lookback. Setting $b = 0$ pools all forecast leads; otherwise, contiguous groups of b leads receive separate bias estimates. The grouping changes which completed errors are averaged, not when a label becomes eligible.

Chronological validation partition. Let validation contain n forecast windows. The query suffix begins at $\max(\lfloor n/2 \rfloor, H + 8)$, adjusted to stay below n. The residual memory ends at least H origins

Table 7: Information available to each correction candidate at forecast origin t. A label is eligible only after its target timestamp has passed.

Candidate	Information used
Identity	Current model prediction only; no labels
Seasonal blend	Observed context through t; no labels
Overlap blend	Predictions issued at earlier origins for the same target timestamps; no labels
Static residual bias	Validation predictions and validation labels
Historical residual retrieval	Validation contexts, predictions, and labels
Causal rolling bias	Validation evidence and test labels with target time $u < t$
ETTm post-test retrieval	Validation evidence and completed official-test residuals; no post-test label with target time $u \geq t$
ETTh online stack	Training data for NLinear, earlier predictions of the same targets, and delayed labels with target time $u < t$

earlier, so every memory future is complete. The final test retriever is fit on the complete validation split because that split precedes test.

Selection tie breaking. Candidates minimize the worst-metric validation ratio $S(c)$ of Section 3.3. Identity wins an exact score tie. When the best score is at least 0.982, only candidates improving all validation metrics are eligible for family-prior selection. Within a family, the lowest score wins; the family order is overlap, seasonal, historical residual, static residual bias, and causal rolling bias.

Internal policy holdout. The supplementary material lists the fixed 123 development and 462 holdout cell IDs. Membership uses the first 123 completed cells ordered by `started_unix` and does not use any metric value. The boundary is nevertheless an opportunistic execution-order split rather than a stratified dataset, model, or difficulty holdout. Its scope and balance diagnostics are reported in Appendix E.

D NO-LOOKAHEAD PROTOCOL

Every forecast and residual is associated with both an origin and its target timestamps. Static validation memory is legal for test because all validation targets precede test. Within online evaluation, a contribution for target timestamp u becomes visible only for query origin $t > u$. Prefix arrays are queried at t, never $t + H$. Rolling-bias statistics write each residual component at its own target timestamp and query the strict prefix $u < t$; horizon blocks group forecast leads but cannot expose a partially realized future.

The evaluation enforces four conditions:

1. each retrieved residual's final target is earlier than the query origin;
2. online statistics and weights contain only target-time evidence $u < t$, including when forecast windows overlap;
3. the PEMS timeline includes the fresh-context gap and 12-step test stride; and
4. post-test ETT memory may contain completed validation and official-test residuals but no current or future post-test target.

Table 7 states the corresponding information set for every candidate. Historical retrieval and static bias use validation labels; seasonal and overlap corrections use no labels; causal rolling bias may incorporate test-period labels only after their target timestamps have passed. Thus the broad selector contains both static and delayed-label prequential candidates.

As a direct no-lookahead check, we replace every label at timestamps $\geq t$ and require the correction at t to remain unchanged.

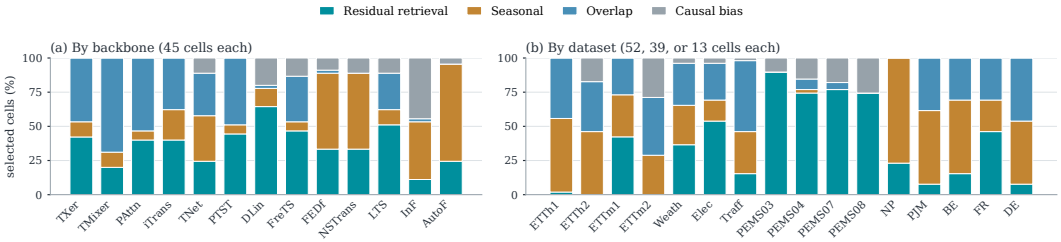

Figure 5: **Validation selects different corrections for different forecasters and regimes.** Share of the 585 primary settings in which each correction family was selected, grouped by backbone (a) and by dataset (b). Error retrieval is chosen for 29/45 DLINEAR settings but 5/45 INFORMER settings, and for 35/39 PEMS03 settings but 0/52 ETTh2 settings. A fixed retrieval-always policy would be wrong in a large fraction of deployments.

E AGGREGATE RESULTS

Table 8: Primary A100 outcomes stratified by the correction family selected on validation. This is descriptive attribution, not a portfolio-minus-retrieval ablation.

Selected family	Cells	Strict wins
Historical residual retrieval	214	207
Seasonal blend	156	151
Overlap blend	159	148
Causal rolling bias	56	56
Identity / static residual bias	0	0
Non-retrieval total	371	355

Table 9: A100 configurations for the 214 cells that select historical residual retrieval. Counts describe validation selections, not component ablations. The search couples (k, τ, λ) into two settings.

Quantity	Selected value	Cells
Correction scale α	.05 / .1 / .2 / .3 / .5	0 / 1 / 6 / 29 / 178
Aggregation (k, τ, λ)	$(8, .1, 1)$ / $(16, .3, 4)$	151 / 63
Descriptor tail length	48 / 96 / 168	57 / 148 / 9
Base-forecast features	Excluded / included	61 / 153

All 214 selected retrieval configurations use first differences, 12 pooled time steps, and a maximum of 16 channel groups; these fixed choices are not ablated. Calendar features are present in 91 configurations and absent in 123, the latter corresponding to PEMS where calendar marks are unavailable.

Selected methods agree in 569/585 cells, selected parameters in 529/585, and strict-win classifications in 581/585. A100 remains the primary result; A10G is a same-seed sensitivity check.

F FULL BENCHMARK DECOMPOSITIONS

The main tables average matched cells by dataset while accounting for the fact that TRACE corrects each backbone independently. Table 13 and Tables 14–16 expose every dataset–horizon aggregate; each row contains exactly 13 paired backbone cells. EPF has only horizon 24; its complete per-model view appears in the full cell tables below. Table 17 instead aggregates by backbone and shows strict wins together with paired relative reductions. The remaining tables in this section report every one of the 585 primary cells. Each entry gives the absolute error of the named backbone followed by the error after applying TRACE; these tables do not average over backbones or horizons.

Table 10: A100 correction-family selections by dataset. HR is historical residual retrieval, SB seasonal blend, OB overlap blend, and CB causal rolling bias. Identity and static bias have zero selections on every dataset.

Dataset	Cells	HR	SB	OB	CB	Dataset	Cells	HR	SB	OB	CB
ETTh1	52	1	28	23	0	ETTh2	52	0	24	19	9
ETTm1	52	22	16	14	0	ETTm2	52	0	15	22	15
Weather	52	19	15	16	2	Electricity	52	28	8	14	2
Traffic	52	8	16	27	1	PEMS03	39	35	0	0	4
PEMS04	39	29	1	3	6	PEMS07	39	30	0	2	7
PEMS08	39	29	0	0	10	NP	13	3	10	0	0
PJM	13	1	7	5	0	BE	13	2	7	4	0
FR	13	6	3	4	0	DE	13	1	6	6	0

Table 11: The 23 A100 cells that do not strictly improve, grouped by dataset. "Worst degradation" is the largest percentage increase among reported test metrics in that group. HR, SB, OB, and CB are defined in Table 10.

Dataset	Cells	HR	SB	OB	CB	Worst degradation
Weather	7	6	0	1	0	2.45%
DE	7	0	2	5	0	2.20%
PEMS04	3	0	0	3	0	36.50%
PEMS07	2	0	0	2	0	9.63%
ETTh2	1	0	1	0	0	1.12%
Electricity	1	1	0	0	0	0.33%
NP	1	0	1	0	0	2.64%
BE	1	0	1	0	0	1.53%
Total	23	7	5	11	0	36.50%

Table 12: A100–A10G discrepancy distribution over all 1,326 reported metric values. Relative difference is absolute difference divided by the larger absolute value. The pre-registered tolerance is $10^{-5} + 10^{-4} \max(|a|, |b|)$.

Quantity	Baseline	Corrected
Within tolerance	472/1326	469/1326
Median relative difference	0.09%	0.06%
90th percentile	2.95%	2.36%
95th percentile	5.80%	5.47%
Maximum	32.79%	28.13%

Table 13: Full long-term results by prediction length. Absolute metrics average the same 13 paired backbones; Δ is the mean cell-wise relative reduction (%).

Dataset	Horizon	Base MSE	+TRACE MSE	Δ MSE	Base MAE	+TRACE MAE	Δ MAE	Wins
ETTh1	96	0.462	**0.412**	7.8	0.454	**0.421**	6.0	13/13
ETTh1	192	0.499	**0.457**	5.7	0.478	**0.451**	4.5	13/13
ETTh1	336	0.565	**0.516**	6.4	0.513	**0.484**	4.7	13/13
ETTh1	720	0.607	**0.544**	7.3	0.553	**0.522**	4.5	13/13
ETTh2	96	0.452	**0.322**	12.8	0.437	**0.368**	9.8	13/13
ETTh2	192	0.574	**0.401**	13.4	0.499	**0.416**	10.0	12/13
ETTh2	336	0.634	**0.446**	14.9	0.540	**0.450**	10.7	13/13
ETTh2	720	0.727	**0.569**	11.9	0.584	**0.525**	7.1	13/13
ETTm1	96	0.411	**0.355**	11.1	0.415	**0.386**	6.0	13/13
ETTm1	192	0.455	**0.403**	8.8	0.437	**0.409**	5.5	13/13
ETTm1	336	0.502	**0.442**	8.7	0.463	**0.435**	4.9	13/13
ETTm1	720	0.571	**0.502**	9.4	0.501	**0.471**	4.8	13/13
ETTm2	96	0.241	**0.193**	10.1	0.310	**0.279**	6.8	13/13
ETTm2	192	0.315	**0.255**	9.7	0.358	**0.319**	7.4	13/13
ETTm2	336	0.417	**0.317**	13.4	0.417	**0.357**	9.8	13/13
ETTm2	720	0.566	**0.404**	14.4	0.492	**0.408**	10.9	13/13
Weather	96	0.223	**0.194**	8.9	0.271	**0.250**	5.6	13/13
Weather	192	0.271	**0.243**	6.8	0.309	**0.291**	4.1	9/13
Weather	336	0.348	**0.287**	8.4	0.356	**0.320**	5.9	10/13
Weather	720	0.435	**0.358**	7.6	0.410	**0.367**	5.7	13/13
Electricity	96	0.243	**0.197**	9.7	0.326	**0.294**	6.4	13/13
Electricity	192	0.204	**0.191**	5.3	0.300	**0.288**	3.7	12/13
Electricity	336	0.219	**0.205**	5.8	0.316	**0.302**	4.1	13/13
Electricity	720	0.307	**0.261**	8.3	0.379	**0.347**	6.0	13/13
Traffic	96	0.571	**0.525**	7.4	0.348	**0.327**	5.7	13/13
Traffic	192	0.579	**0.537**	6.4	0.353	**0.332**	5.5	13/13
Traffic	336	0.600	**0.564**	5.2	0.362	**0.343**	4.6	13/13
Traffic	720	0.642	**0.608**	4.5	0.382	**0.366**	3.8	13/13

Table 14: Full PEMS MAE results by prediction length. Absolute metrics average the same 13 paired backbones; Δ is the mean cell-wise relative reduction (%). Strict wins require all three PEMS metrics and therefore repeat across these tables.

Dataset	Horizon	Base MAE	+TRACE MAE	Δ MAE	Strict wins
PEMS03	6	18.504	**16.378**	10.2	13/13
PEMS03	12	20.037	**17.545**	11.6	13/13
PEMS03	24	24.055	**20.647**	13.2	13/13
PEMS04	6	23.807	**21.872**	7.4	12/13
PEMS04	12	26.688	**23.728**	9.5	12/13
PEMS04	24	31.475	**26.896**	12.2	12/13
PEMS07	6	26.781	**24.330**	8.8	12/13
PEMS07	12	29.145	**25.604**	11.3	12/13
PEMS07	24	34.510	**29.108**	14.6	13/13
PEMS08	6	20.024	**18.071**	9.6	13/13
PEMS08	12	22.363	**19.316**	12.6	13/13
PEMS08	24	27.167	**22.449**	15.9	13/13

Table 15: Full PEMS RMSE results by prediction length. Absolute metrics average the same 13 paired backbones; Δ is the mean cell-wise relative reduction (%). Strict wins require all three PEMS metrics and therefore repeat across these tables.

Dataset	Horizon	Base RMSE	+TRACE RMSE	Δ RMSE	Strict wins
PEMS03	6	28.874	**25.601**	10.2	13/13
PEMS03	12	31.319	**27.566**	11.3	13/13
PEMS03	24	37.664	**32.623**	12.4	13/13
PEMS04	6	36.313	**33.846**	6.3	12/13
PEMS04	12	40.393	**36.414**	8.7	12/13
PEMS04	24	47.375	**40.790**	12.0	12/13
PEMS07	6	40.478	**37.340**	7.5	12/13
PEMS07	12	44.220	**39.634**	9.7	12/13
PEMS07	24	51.621	**44.486**	12.9	13/13
PEMS08	6	30.493	**27.795**	8.8	13/13
PEMS08	12	33.811	**29.497**	11.9	13/13
PEMS08	24	41.074	**34.084**	15.6	13/13

Table 16: Full PEMS MAPE results by prediction length. Absolute metrics average the same 13 paired backbones; Δ is the mean cell-wise relative reduction (%). Strict wins require all three PEMS metrics and therefore repeat across these tables.

Dataset	Horizon	Base MAPE	+TRACE MAPE	Δ MAPE	Strict wins
PEMS03	6	0.187	**0.164**	10.5	13/13
PEMS03	12	0.197	**0.174**	11.1	13/13
PEMS03	24	0.238	**0.207**	12.4	13/13
PEMS04	6	0.166	**0.147**	10.2	12/13
PEMS04	12	0.189	**0.165**	10.3	12/13
PEMS04	24	0.221	**0.188**	11.8	12/13
PEMS07	6	0.122	**0.107**	11.8	12/13
PEMS07	12	0.132	**0.114**	12.8	12/13
PEMS07	24	0.165	**0.135**	16.2	13/13
PEMS08	6	0.125	**0.111**	10.2	13/13
PEMS08	12	0.141	**0.122**	12.7	13/13
PEMS08	24	0.174	**0.147**	14.5	13/13

Table 17: Backbone-level paired results. Entries are strict wins and mean cell-wise relative reductions (%); positive values favor TRACE. This view prevents weak backbones or large absolute errors from dominating the conclusion.

	Long-term (28 cells)			PEMS (12 cells)				EPF (5 cells)		
Backbone	Wins	MSE	MAE	Wins	MAE	RMSE	MAPE	Wins	MSE	MAE
TimeXer	28/28	1.1	0.9	12/12	14.1	11.9	15.9	3/5	0.3	0.4
TimeMixer	27/28	1.3	1.2	7/12	-2.4	-0.8	-5.5	4/5	3.2	2.5
PAttn	27/28	1.7	1.0	12/12	13.0	12.5	11.8	4/5	1.5	1.8
iTransformer	26/28	2.9	2.0	12/12	10.3	9.8	10.6	4/5	1.1	0.9
TimesNet	28/28	2.9	2.0	12/12	5.8	4.8	7.1	4/5	4.8	3.2
PatchTST	27/28	1.8	1.2	12/12	16.8	15.2	16.9	5/5	1.9	1.9
DLinear	27/28	10.1	7.6	12/12	24.7	23.4	27.4	5/5	4.7	4.7
FreTS	25/28	6.3	4.6	12/12	11.9	11.0	11.7	4/5	3.2	3.2
FEDformer	28/28	8.5	6.1	12/12	14.0	11.4	17.3	5/5	11.9	9.3
Non-stat. Transformer	28/28	16.4	10.0	12/12	4.9	3.9	6.5	4/5	15.0	15.3
LightTS	28/28	10.3	7.8	12/12	9.4	8.2	9.0	4/5	1.6	2.2
Informer	28/28	41.3	28.7	12/12	8.4	9.7	10.2	5/5	9.7	6.1
Autoformer	28/28	11.6	8.0	12/12	17.4	17.0	17.4	5/5	17.5	10.7

Table 18: Complete long-term results on ETTh1. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	96		192		336		720	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TimeXer	Base	0.3818	0.4029	0.4285	0.4355	0.4672	0.4494	0.4988	0.4926
	+TRACE	**0.3739**	**0.3932**	**0.4267**	**0.4342**	**0.4622**	**0.4457**	**0.4893**	**0.4871**
TimeMixer	Base	0.3784	0.3981	0.4410	0.4301	0.5003	0.4600	0.4787	0.4724
	+TRACE	**0.3782**	**0.3895**	**0.4378**	**0.4278**	**0.4959**	**0.4568**	**0.4696**	**0.4680**
PAttn	Base	0.3901	0.4048	0.4371	0.4392	0.4977	0.4703	0.5326	0.4995
	+TRACE	**0.3828**	**0.3961**	**0.4346**	**0.4374**	**0.4915**	**0.4662**	**0.5195**	**0.4927**
iTransformer	Base	0.4469	0.4439	0.5187	0.4833	0.5538	0.5018	0.5629	0.5251
	+TRACE	**0.4054**	**0.4139**	**0.4679**	**0.4506**	**0.5107**	**0.4727**	**0.5124**	**0.4930**
TimesNet	Base	0.3890	0.4120	0.4395	0.4418	0.4941	0.4709	0.5178	0.4944
	+TRACE	**0.3750**	**0.3933**	**0.4314**	**0.4258**	**0.4798**	**0.4519**	**0.5136**	**0.4928**
PatchTST	Base	0.3793	0.3996	0.4274	0.4325	0.4783	0.4643	0.5220	0.5063
	+TRACE	**0.3727**	**0.3916**	**0.4266**	**0.4318**	**0.4736**	**0.4613**	**0.5121**	**0.5016**
DLinear	Base	0.3962	0.4108	0.4450	0.4404	0.4874	0.4654	0.5126	0.5104
	+TRACE	**0.3861**	**0.3990**	**0.4441**	**0.4394**	**0.4776**	**0.4608**	**0.5114**	**0.5099**
FreTS	Base	0.3999	0.4124	0.4512	0.4426	0.5086	0.4814	0.5627	0.5398
	+TRACE	**0.3928**	**0.4063**	**0.4436**	**0.4364**	**0.5000**	**0.4748**	**0.5497**	**0.5324**
FEDFormer	Base	0.3770	0.4184	0.4200	0.4441	0.4584	0.4666	0.5021	0.5032
	+TRACE	**0.3643**	**0.3945**	**0.4133**	**0.4239**	**0.4526**	**0.4515**	**0.4974**	**0.5008**
Non-stat. Transformer	Base	0.5752	0.4933	0.4762	0.4783	0.7532	0.6289	0.8274	0.6659
	+TRACE	**0.4516**	**0.4353**	**0.4336**	**0.4411**	**0.6365**	**0.5645**	**0.6365**	**0.5716**
LightTS	Base	0.4352	0.4436	0.4936	0.4782	0.5493	0.5103	0.6117	0.5688
	+TRACE	**0.4133**	**0.4280**	**0.4653**	**0.4592**	**0.5169**	**0.4900**	**0.5748**	**0.5492**
Informer	Base	0.9590	0.7778	1.0199	0.7930	1.0310	0.7809	1.2219	0.8817
	+TRACE	**0.6328**	**0.6056**	**0.6812**	**0.6191**	**0.7092**	**0.6221**	**0.8044**	**0.6969**
Autoformer	Base	0.4993	0.4807	0.4868	0.4756	0.5648	0.5242	0.5426	0.5231
	+TRACE	**0.4262**	**0.4328**	**0.4408**	**0.4400**	**0.5024**	**0.4801**	**0.4850**	**0.4849**

Table 19: Complete long-term results on ETTh2. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	96		192		336		720	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TimeXer	Base	0.2854	0.3376	0.3627	0.3892	0.4113	0.4214	0.4485	0.4534
	+TRACE	**0.2842**	**0.3365**	**0.3606**	**0.3875**	**0.4076**	**0.4192**	**0.4481**	**0.4531**
TimeMixer	Base	0.2878	0.3395	0.3778	0.3957	0.4247	0.4298	0.4681	0.4683
	+TRACE	**0.2875**	**0.3391**	0.3820	**0.3917**	**0.4163**	**0.4246**	**0.4671**	**0.4674**
PAttn	Base	0.2988	0.3453	0.3826	0.3954	0.4232	0.4305	0.4279	0.4438
	+TRACE	**0.2941**	**0.3408**	**0.3782**	**0.3917**	**0.4182**	**0.4269**	**0.4233**	**0.4407**
iTransformer	Base	0.3004	0.3496	0.3818	0.3998	0.4236	0.4324	0.4264	0.4451
	+TRACE	**0.2997**	**0.3435**	**0.3787**	**0.3969**	**0.4202**	**0.4298**	**0.4231**	**0.4426**
TimesNet	Base	0.3370	0.3709	0.4046	0.4146	0.4520	0.4503	0.4337	0.4477
	+TRACE	**0.3140**	**0.3536**	**0.3896**	**0.4015**	**0.4340**	**0.4357**	**0.4305**	**0.4455**
PatchTST	Base	0.2921	0.3452	0.3814	0.4044	0.4185	0.4349	0.4332	0.4524
	+TRACE	**0.2890**	**0.3421**	**0.3752**	**0.4006**	**0.4103**	**0.4306**	**0.4298**	**0.4502**
DLinear	Base	0.3414	0.3953	0.4819	0.4792	0.5929	0.5422	0.8403	0.6611
	+TRACE	**0.2996**	**0.3483**	**0.3584**	**0.3902**	**0.4123**	**0.4255**	**0.6150**	**0.5612**
FreTS	Base	0.3415	0.3967	0.4396	0.4492	0.5375	0.5094	0.7956	0.6436
	+TRACE	**0.3013**	**0.3538**	**0.4045**	**0.4207**	**0.4720**	**0.4707**	**0.5933**	**0.5511**
FEDFormer	Base	0.3480	0.3906	0.4245	0.4371	0.4566	0.4676	0.4758	0.4858
	+TRACE	**0.3190**	**0.3653**	**0.3994**	**0.4133**	**0.4388**	**0.4481**	**0.4502**	**0.4638**
Non-stat. Transformer	Base	0.4594	0.4509	0.6137	0.5253	0.7591	0.5913	0.5136	0.4918
	+TRACE	**0.3688**	**0.3963**	**0.4933**	**0.4618**	**0.5975**	**0.5148**	**0.4538**	**0.4586**
LightTS	Base	0.4353	0.4730	0.5613	0.5382	0.6829	0.5985	0.9366	0.7145
	+TRACE	**0.3615**	**0.4144**	**0.4604**	**0.4708**	**0.4795**	**0.4857**	**0.6603**	**0.5912**
Informer	Base	1.7643	1.0737	2.2024	1.2022	2.1738	1.2303	2.7496	1.3864
	+TRACE	**0.4310**	**0.4645**	**0.4212**	**0.4602**	**0.4386**	**0.4705**	**1.5481**	**1.0228**
Autoformer	Base	0.3817	0.4150	0.4512	0.4530	0.4833	0.4830	0.5035	0.5025
	+TRACE	**0.3363**	**0.3848**	**0.4106**	**0.4257**	**0.4567**	**0.4620**	**0.4599**	**0.4732**

Table 20: Complete long-term results on ETTm1. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	96		192		336		720	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TimeXer	Base	0.3178	0.3563	0.3616	0.3829	0.3951	0.4067	0.4524	0.4413
	+TRACE	**0.3072**	**0.3514**	**0.3583**	**0.3808**	**0.3872**	**0.4020**	**0.4470**	**0.4401**
TimeMixer	Base	0.3184	0.3581	0.3654	0.3868	0.3894	0.4027	0.4697	0.4528
	+TRACE	**0.3040**	**0.3515**	**0.3560**	**0.3805**	**0.3812**	**0.3974**	**0.4515**	**0.4453**
PAttn	Base	0.3364	0.3648	0.3762	0.3832	0.4074	0.4036	0.4674	0.4376
	+TRACE	**0.3201**	**0.3583**	**0.3737**	**0.3808**	**0.4048**	**0.4012**	**0.4533**	**0.4357**
iTransformer	Base	0.3557	0.3813	0.3988	0.4020	0.4330	0.4242	0.5005	0.4615
	+TRACE	**0.3379**	**0.3733**	**0.3973**	**0.4007**	**0.4315**	**0.4231**	**0.4800**	**0.4582**
TimesNet	Base	0.3339	0.3754	0.4081	0.4137	0.4150	0.4223	0.5101	0.4722
	+TRACE	**0.3180**	**0.3602**	**0.3771**	**0.3878**	**0.4134**	**0.4212**	**0.5034**	**0.4704**
PatchTST	Base	0.3268	0.3664	0.3700	0.3903	0.3982	0.4084	0.4577	0.4444
	+TRACE	**0.3131**	**0.3602**	**0.3668**	**0.3883**	**0.3956**	**0.4066**	**0.4484**	**0.4412**
DLinear	Base	0.3459	0.3737	0.3818	0.3911	0.4153	0.4150	0.4730	0.4509
	+TRACE	**0.3205**	**0.3620**	**0.3807**	**0.3898**	**0.4078**	**0.4118**	**0.4557**	**0.4455**
FreTS	Base	0.3398	0.3754	0.3833	0.3984	0.4202	0.4250	0.4975	0.4742
	+TRACE	**0.3247**	**0.3663**	**0.3787**	**0.3949**	**0.4159**	**0.4218**	**0.4821**	**0.4670**
FEDformer	Base	0.5353	0.4892	0.5273	0.4908	0.6801	0.5477	0.7092	0.5623
	+TRACE	**0.4252**	**0.4418**	**0.4238**	**0.4314**	**0.5074**	**0.4858**	**0.5556**	**0.5081**
Non-stat. Transformer	Base	0.5062	0.4644	0.6041	0.4973	0.6640	0.5186	0.7208	0.5431
	+TRACE	**0.3945**	**0.4076**	**0.4652**	**0.4350**	**0.5083**	**0.4560**	**0.5741**	**0.4850**
LightTS	Base	0.3691	0.3914	0.3994	0.4056	0.4296	0.4261	0.4876	0.4623
	+TRACE	**0.3384**	**0.3767**	**0.3796**	**0.3967**	**0.4243**	**0.4217**	**0.4682**	**0.4557**
Informer	Base	0.7403	0.6205	0.7609	0.6270	0.8722	0.6878	0.9979	0.7377
	+TRACE	**0.5049**	**0.4937**	**0.5280**	**0.5034**	**0.5926**	**0.5453**	**0.6693**	**0.5802**
Autoformer	Base	0.5198	0.4821	0.5781	0.5123	0.6100	0.5325	0.6812	0.5671
	+TRACE	**0.4015**	**0.4169**	**0.4486**	**0.4437**	**0.4743**	**0.4629**	**0.5340**	**0.4949**

Table 21: Complete long-term results on ETTm2. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	96		192		336		720	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TimeXer	Base	0.1712	0.2560	0.2365	0.2987	0.2953	0.3379	0.3934	0.3942
	+TRACE	**0.1707**	**0.2557**	**0.2358**	**0.2980**	**0.2926**	**0.3359**	**0.3908**	**0.3924**
TimeMixer	Base	0.1750	0.2574	0.2360	0.2975	0.3026	0.3410	0.4009	0.4054
	+TRACE	**0.1743**	**0.2569**	**0.2350**	**0.2966**	**0.2994**	**0.3379**	**0.3958**	**0.4021**
PAttn	Base	0.1800	0.2668	0.2462	0.3079	0.3084	0.3471	0.4076	0.4031
	+TRACE	**0.1788**	**0.2657**	**0.2448**	**0.3066**	**0.3057**	**0.3451**	**0.4045**	**0.4009**
iTransformer	Base	0.1862	0.2723	0.2516	0.3119	0.3153	0.3527	0.4132	0.4060
	+TRACE	**0.1830**	**0.2658**	**0.2498**	**0.3095**	**0.3125**	**0.3499**	**0.4098**	**0.4031**
TimesNet	Base	0.1894	0.2662	0.2634	0.3119	0.3204	0.3488	0.4229	0.4069
	+TRACE	**0.1807**	**0.2623**	**0.2630**	**0.3114**	**0.3195**	**0.3479**	**0.4208**	**0.4058**
PatchTST	Base	0.1828	0.2679	0.2463	0.3058	0.3105	0.3492	0.4150	0.4108
	+TRACE	**0.1793**	**0.2654**	**0.2446**	**0.3045**	**0.3082**	**0.3478**	**0.4108**	**0.4083**
DLinear	Base	0.1934	0.2928	0.2845	0.3614	0.3849	0.4295	0.5561	0.5235
	+TRACE	**0.1771**	**0.2639**	**0.2394**	**0.3033**	**0.2840**	**0.3357**	**0.3908**	**0.3993**
FreTS	Base	0.1943	0.2846	0.2774	0.3507	0.3698	0.4042	0.5602	0.5186
	+TRACE	**0.1771**	**0.2656**	**0.2593**	**0.3280**	**0.2855**	**0.3384**	**0.3841**	**0.3997**
FEDformer	Base	0.2098	0.2933	0.2710	0.3288	0.3275	0.3641	0.4204	0.4164
	+TRACE	**0.1949**	**0.2845**	**0.2578**	**0.3213**	**0.3186**	**0.3586**	**0.4136**	**0.4051**
Non-stat. Transformer	Base	0.2233	0.3008	0.2680	0.3234	0.4862	0.4538	0.4572	0.4351
	+TRACE	**0.2029**	**0.2814**	**0.2478**	**0.3067**	**0.3899**	**0.3997**	**0.4260**	**0.4132**
LightTS	Base	0.2088	0.3076	0.3005	0.3748	0.3933	0.4347	0.5449	0.5188
	+TRACE	**0.1934**	**0.2831**	**0.2510**	**0.3164**	**0.2983**	**0.3495**	**0.3755**	**0.3993**
Informer	Base	0.7444	0.6312	0.9328	0.7436	1.2786	0.8891	1.9379	1.1306
	+TRACE	**0.2816**	**0.3771**	**0.3287**	**0.4149**	**0.3936**	**0.4399**	**0.4170**	**0.4590**
Autoformer	Base	0.2719	0.3291	0.2845	0.3429	0.3319	0.3669	0.4277	0.4223
	+TRACE	**0.2212**	**0.3031**	**0.2624**	**0.3277**	**0.3173**	**0.3573**	**0.4162**	**0.4132**

Table 22: Complete long-term results on Weather. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	96		192		336		720	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TimeXer	Base	0.1786	0.2244	0.2041	0.2475	0.2607	0.2895	0.3400	0.3404
	+TRACE	**0.1746**	**0.2236**	**0.2034**	**0.2473**	**0.2598**	**0.2892**	**0.3388**	**0.3398**
TimeMixer	Base	0.1610	0.2079	0.2064	0.2507	0.2635	0.2917	0.3438	0.3436
	+TRACE	**0.1606**	**0.2077**	**0.2057**	**0.2503**	**0.2630**	**0.2916**	**0.3423**	**0.3426**
PAttn	Base	0.1894	0.2263	0.2341	0.2641	0.2873	0.3015	0.3622	0.3495
	+TRACE	**0.1837**	**0.2257**	**0.2328**	0.2706	**0.2862**	**0.3012**	**0.3602**	**0.3488**
iTransformer	Base	0.1971	0.2366	0.2432	0.2738	0.2951	0.3104	0.3611	0.3510
	+TRACE	**0.1902**	**0.2340**	**0.2378**	0.2745	**0.2879**	0.3105	**0.3572**	**0.3490**
TimesNet	Base	0.1690	0.2195	0.2247	0.2640	0.2815	0.3036	0.3566	0.3526
	+TRACE	**0.1658**	**0.2168**	**0.2245**	**0.2639**	**0.2811**	**0.3034**	**0.3554**	**0.3519**
PatchTST	Base	0.1722	0.2135	0.2200	0.2556	0.2794	0.2979	0.3546	0.3467
	+TRACE	**0.1677**	**0.2129**	**0.2194**	0.2556	**0.2788**	**0.2978**	**0.3534**	**0.3462**
DLinear	Base	0.1962	0.2561	0.2389	0.2992	0.2811	0.3306	0.3454	0.3819
	+TRACE	**0.1901**	**0.2506**	**0.2360**	**0.2987**	**0.2785**	0.3308	**0.3353**	**0.3759**
FreTS	Base	0.1843	0.2386	0.2226	0.2735	0.2724	0.3175	0.3544	0.3868
	+TRACE	**0.1797**	**0.2355**	**0.2218**	0.2771	**0.2704**	0.3198	**0.3519**	**0.3857**
FEDFormer	Base	0.2713	0.3419	0.3217	0.3774	0.3282	0.3654	0.3938	0.3993
	+TRACE	**0.2038**	**0.2778**	**0.2603**	**0.3210**	**0.2997**	**0.3356**	**0.3817**	**0.3959**
Non-stat. Transformer	Base	0.2086	0.2512	0.2787	0.3057	0.3649	0.3584	0.3928	0.3789
	+TRACE	**0.1836**	**0.2317**	**0.2394**	**0.2778**	**0.3159**	**0.3283**	**0.3707**	**0.3636**
LightTS	Base	0.1964	0.2552	0.2419	0.2987	0.2899	0.3365	0.3497	0.3834
	+TRACE	**0.1904**	**0.2491**	**0.2392**	**0.2982**	**0.2822**	**0.3323**	**0.3384**	**0.3782**
Informer	Base	0.5204	0.5156	0.5769	0.5391	0.9701	0.7332	1.2664	0.8835
	+TRACE	**0.3321**	**0.3994**	**0.3830**	**0.4274**	**0.3254**	**0.3776**	**0.3776**	**0.3929**
Autoformer	Base	0.2582	0.3310	0.3039	0.3625	0.3481	0.3888	0.4332	0.4336
	+TRACE	**0.2056**	**0.2806**	**0.2575**	**0.3186**	**0.3021**	**0.3440**	**0.3907**	**0.3956**

Table 23: Complete long-term results on Electricity. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	96		192		336		720	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TimeXer	Base	0.1405	0.2425	0.1571	0.2556	0.1769	0.2752	0.2110	0.3061
	+TRACE	**0.1400**	**0.2379**	**0.1564**	**0.2525**	**0.1739**	**0.2693**	**0.2055**	**0.3011**
TimeMixer	Base	0.1562	0.2476	0.1701	0.2610	0.1868	0.2773	0.2283	0.3131
	+TRACE	**0.1528**	**0.2447**	**0.1679**	**0.2590**	**0.1841**	**0.2751**	**0.2251**	**0.3105**
PAttn	Base	0.1846	0.2668	0.1887	0.2717	0.2043	0.2875	0.2467	0.3216
	+TRACE	**0.1737**	**0.2620**	**0.1865**	**0.2700**	**0.2014**	**0.2854**	**0.2436**	**0.3191**
iTransformer	Base	0.1488	0.2406	0.1645	0.2560	0.1784	0.2712	0.2271	0.3114
	+TRACE	**0.1443**	**0.2364**	**0.1611**	**0.2528**	**0.1738**	**0.2670**	**0.2206**	**0.3050**
TimesNet	Base	0.1679	0.2716	0.1860	0.2883	0.2036	0.3051	0.2203	0.3183
	+TRACE	**0.1649**	**0.2676**	**0.1739**	**0.2747**	**0.1961**	**0.2948**	**0.2166**	**0.3156**
PatchTST	Base	0.1800	0.2730	0.1874	0.2797	0.2043	0.2960	0.2455	0.3283
	+TRACE	**0.1697**	**0.2622**	**0.1835**	**0.2743**	**0.1949**	**0.2856**	**0.2431**	**0.3267**
DLinear	Base	0.2104	0.3016	0.2102	0.3047	0.2231	0.3192	0.2578	0.3496
	+TRACE	**0.1897**	**0.2848**	**0.2007**	**0.2950**	**0.2058**	**0.3022**	**0.2551**	**0.3424**
FreTS	Base	0.1895	0.2768	0.1917	0.2792	0.2069	0.2966	0.2467	0.3333
	+TRACE	**0.1782**	**0.2704**	**0.1900**	0.2801	**0.1987**	**0.2897**	**0.2427**	**0.3270**
FEDFormer	Base	0.1948	0.3089	0.2022	0.3155	0.2292	0.3423	0.2620	0.3655
	+TRACE	**0.1831**	**0.2945**	**0.1905**	**0.3001**	**0.2071**	**0.3188**	**0.2420**	**0.3377**
Non-stat. Transformer	Base	0.8414	0.7579	0.1836	0.2857	0.2002	0.3033	0.8928	0.7727
	+TRACE	**0.4065**	**0.4973**	**0.1705**	**0.2684**	**0.1867**	**0.2887**	**0.4818**	**0.5314**
LightTS	Base	0.2133	0.3156	0.2218	0.3251	0.2416	0.3445	0.2770	0.3710
	+TRACE	**0.1955**	**0.2985**	**0.2038**	**0.3065**	**0.2208**	**0.3240**	**0.2560**	**0.3460**
Informer	Base	0.3290	0.4131	0.3671	0.4454	0.3523	0.4348	0.4004	0.4597
	+TRACE	**0.2811**	**0.3742**	**0.2954**	**0.3917**	**0.2967**	**0.3923**	**0.3035**	**0.3912**
Autoformer	Base	0.2006	0.3174	0.2252	0.3333	0.2445	0.3516	0.2763	0.3772
	+TRACE	**0.1796**	**0.2865**	**0.2044**	**0.3128**	**0.2250**	**0.3307**	**0.2517**	**0.3525**

Table 24: Complete long-term results on Traffic. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	96		192		336		720	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TimeXer	Base	0.4300	0.2720	0.4477	0.2812	0.4714	0.2889	0.5148	0.3069
	+TRACE	**0.4247**	**0.2679**	**0.4416**	**0.2761**	**0.4647**	**0.2843**	**0.5086**	**0.3029**
TimeMixer	Base	0.4743	0.2963	0.5029	0.3016	0.5281	0.3083	0.5543	0.3233
	+TRACE	**0.4618**	**0.2853**	**0.4941**	**0.2922**	**0.5180**	**0.2973**	**0.5446**	**0.3138**
PAttn	Base	0.5050	0.3298	0.5023	0.3234	0.5153	0.3275	0.5489	0.3459
	+TRACE	**0.4876**	**0.3169**	**0.4907**	**0.3140**	**0.5028**	**0.3180**	**0.5361**	**0.3366**
iTransformer	Base	0.3936	0.2686	0.4123	0.2770	0.4235	0.2828	0.4581	0.3002
	+TRACE	**0.3786**	**0.2599**	**0.4057**	**0.2701**	**0.4163**	**0.2760**	**0.4510**	**0.2938**
TimesNet	Base	0.5900	0.3144	0.6147	0.3243	0.6347	0.3380	0.6557	0.3467
	+TRACE	**0.5302**	**0.3041**	**0.5801**	**0.3189**	**0.6289**	**0.3338**	**0.6231**	**0.3422**
PatchTST	Base	0.4582	0.2981	0.4687	0.3008	0.4828	0.3074	0.5174	0.3261
	+TRACE	**0.4307**	**0.2827**	**0.4619**	**0.2958**	**0.4752**	**0.3024**	**0.5097**	**0.3215**
DLinear	Base	0.6965	0.4287	0.6466	0.4074	0.6532	0.4097	0.6940	0.4286
	+TRACE	**0.5378**	**0.3537**	**0.5234**	**0.3452**	**0.5719**	**0.3708**	**0.6727**	**0.4188**
FreTS	Base	0.5649	0.3679	0.5683	0.3650	0.6004	0.3737	0.6614	0.3978
	+TRACE	**0.5374**	**0.3453**	**0.5467**	**0.3449**	**0.5781**	**0.3543**	**0.6410**	**0.3808**
FEDformer	Base	0.5775	0.3624	0.6095	0.3800	0.6266	0.3901	0.6511	0.3988
	+TRACE	**0.5514**	**0.3487**	**0.5744**	**0.3591**	**0.5935**	**0.3730**	**0.6258**	**0.3862**
Non-stat. Transformer	Base	0.6307	0.3409	0.6371	0.3457	0.6565	0.3495	0.6774	0.3631
	+TRACE	**0.5687**	**0.3301**	**0.5798**	**0.3328**	**0.6164**	**0.3411**	**0.6401**	**0.3549**
LightTS	Base	0.6812	0.4218	0.6701	0.4270	0.6897	0.4360	0.7443	0.4561
	+TRACE	**0.6264**	**0.3803**	**0.6160**	**0.3854**	**0.6346**	**0.3947**	**0.6896**	**0.4169**
Informer	Base	0.7498	0.4238	0.8101	0.4606	0.8660	0.4898	1.0016	0.5617
	+TRACE	**0.6605**	**0.3888**	**0.6807**	**0.4098**	**0.7350**	**0.4374**	**0.8322**	**0.4921**
Autoformer	Base	0.6726	0.4027	0.6330	0.3969	0.6476	0.4041	0.6688	0.4133
	+TRACE	**0.6262**	**0.3818**	**0.5847**	**0.3689**	**0.6012**	**0.3821**	**0.6295**	**0.3924**

Table 25: Complete PEMS results on PEMS03. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	6			12			24		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
TimeXer	Base	16.559	26.203	0.154	18.376	28.527	0.176	21.692	33.321	0.216
	+TRACE	**14.533**	**22.930**	**0.137**	**16.004**	**25.151**	**0.153**	**18.687**	**29.249**	**0.185**
TimeMixer	Base	13.910	21.763	0.138	15.355	24.195	0.155	17.252	27.246	0.176
	+TRACE	**13.682**	**21.524**	**0.135**	**14.928**	**23.702**	**0.148**	**16.532**	**26.410**	**0.165**
PAttn	Base	15.733	24.870	0.145	18.943	30.139	0.175	24.975	40.034	0.231
	+TRACE	**14.506**	**22.904**	**0.136**	**16.694**	**26.378**	**0.158**	**20.900**	**32.963**	**0.201**
iTransformer	Base	15.031	23.749	0.140	17.389	27.595	0.163	21.992	35.361	0.213
	+TRACE	**14.193**	**22.464**	**0.132**	**15.957**	**25.289**	**0.151**	**19.354**	**31.072**	**0.188**
TimesNet	Base	16.513	26.352	0.164	17.732	28.254	0.177	20.112	32.264	0.202
	+TRACE	**15.725**	**25.025**	**0.154**	**16.969**	**27.382**	**0.168**	**18.993**	**30.957**	**0.190**
PatchTST	Base	16.619	25.762	0.163	20.246	31.971	0.180	25.350	40.365	0.227
	+TRACE	**14.641**	**22.984**	**0.142**	**16.798**	**26.499**	**0.156**	**20.601**	**32.524**	**0.194**
DLinear	Base	18.971	29.982	0.195	22.877	36.186	0.236	29.958	47.391	0.307
	+TRACE	**15.395**	**24.224**	**0.156**	**17.668**	**27.850**	**0.180**	**21.654**	**34.106**	**0.222**
FreTS	Base	15.382	24.357	0.143	17.898	28.408	0.169	22.465	35.778	0.215
	+TRACE	**14.272**	**22.523**	**0.134**	**15.962**	**25.304**	**0.153**	**19.138**	**30.402**	**0.186**
FEDformer	Base	25.876	37.318	0.294	26.304	38.518	0.278	29.555	43.522	0.301
	+TRACE	**21.596**	**32.477**	**0.241**	**21.652**	**33.021**	**0.229**	**24.338**	**36.877**	**0.253**
Non-stat. Transformer	Base	16.294	26.154	0.161	17.917	28.759	0.177	20.669	33.785	0.202
	+TRACE	**15.501**	**24.778**	**0.151**	**16.992**	**27.721**	**0.166**	**19.511**	**32.409**	**0.189**
LightTS	Base	17.656	27.076	0.180	18.792	29.051	0.189	21.942	33.751	0.218
	+TRACE	**16.505**	**25.565**	**0.169**	**17.235**	**26.777**	**0.176**	**19.705**	**30.666**	**0.199**
Informer	Base	18.424	31.355	0.186	19.623	32.938	0.194	20.855	34.355	0.208
	+TRACE	**15.962**	**25.736**	**0.157**	**17.242**	**27.509**	**0.166**	**19.482**	**32.970**	**0.191**
Autoformer	Base	33.587	50.426	0.366	29.025	42.609	0.290	35.903	52.458	0.384
	+TRACE	**26.409**	**39.679**	**0.290**	**23.989**	**35.783**	**0.252**	**29.511**	**43.494**	**0.324**

Table 26: Complete PEMS results on PEMS04. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	6			12			24		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
TimeXer	Base	22.215	33.451	0.158	24.422	36.491	0.172	28.035	42.229	0.188
	+TRACE	**19.773**	**31.039**	**0.131**	**21.278**	**33.117**	**0.141**	**24.193**	**37.410**	**0.157**
TimeMixer	Base	18.747	29.822	0.121	19.750	31.202	0.130	21.490	33.618	0.142
	+TRACE	20.598	31.501	0.138	23.184	34.212	0.172	25.496	37.059	0.194
PAttn	Base	21.504	33.940	0.137	25.486	39.844	0.161	33.920	52.384	0.219
	+TRACE	**20.033**	**32.110**	**0.126**	**22.544**	**35.754**	**0.142**	**27.762**	**42.966**	**0.181**
iTransformer	Base	20.731	32.867	0.132	23.640	37.288	0.150	29.372	45.862	0.193
	+TRACE	**19.384**	**31.277**	**0.121**	**21.307**	**34.101**	**0.133**	**25.020**	**39.300**	**0.161**
TimesNet	Base	21.198	33.376	0.139	22.399	34.951	0.149	24.460	38.015	0.165
	+TRACE	**20.691**	**32.576**	**0.133**	**21.639**	**34.162**	**0.142**	**23.186**	**36.594**	**0.154**
PatchTST	Base	23.341	35.083	0.164	26.846	40.585	0.182	34.293	52.480	0.220
	+TRACE	**20.428**	**32.074**	**0.135**	**22.683**	**35.464**	**0.148**	**27.324**	**42.093**	**0.177**
DLinear	Base	27.263	40.283	0.213	32.003	46.974	0.254	40.203	58.240	0.329
	+TRACE	**21.705**	**33.227**	**0.156**	**24.145**	**36.439**	**0.178**	**28.285**	**41.726**	**0.219**
FreTS	Base	21.054	33.114	0.138	24.229	37.670	0.159	30.438	46.446	0.207
	+TRACE	**19.620**	**31.434**	**0.127**	**21.601**	**34.279**	**0.139**	**25.284**	**39.138**	**0.169**
FEDFormer	Base	32.008	45.741	0.254	32.490	47.083	0.251	34.952	51.084	0.268
	+TRACE	**29.034**	**42.324**	**0.213**	**27.308**	**41.068**	**0.200**	**29.315**	**44.502**	**0.215**
Non-stat. Transformer	Base	20.983	33.107	0.138	21.942	34.550	0.143	23.828	37.152	0.160
	+TRACE	**20.548**	**32.465**	**0.130**	**21.264**	**33.836**	**0.138**	**22.937**	**36.120**	**0.153**
LightTS	Base	24.701	37.135	0.174	25.326	38.118	0.179	29.292	43.597	0.210
	+TRACE	**22.921**	**35.024**	**0.161**	**23.379**	**35.888**	**0.163**	**25.998**	**39.258**	**0.185**
Informer	Base	22.615	36.234	0.149	23.238	36.965	0.153	24.321	38.028	0.166
	+TRACE	**21.122**	**33.212**	**0.137**	**21.633**	**33.746**	**0.141**	**23.474**	**35.975**	**0.159**
Autoformer	Base	33.137	47.924	0.247	45.172	63.391	0.374	54.568	76.742	0.410
	+TRACE	**28.475**	**41.736**	**0.198**	**36.502**	**51.311**	**0.307**	**41.372**	**58.130**	**0.323**

Table 27: Complete PEMS results on PEMS07. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	6			12			24		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
TimeXer	Base	24.304	35.922	0.109	27.061	39.855	0.126	31.441	46.343	0.148
	+TRACE	**21.117**	**32.784**	**0.090**	**23.108**	**35.720**	**0.101**	**26.407**	**40.569**	**0.118**
TimeMixer	Base	19.752	31.426	0.083	21.294	33.866	0.090	23.887	37.663	0.104
	+TRACE	20.324	31.797	0.086	22.430	34.635	0.099	**22.836**	**36.460**	**0.098**
PAttn	Base	23.070	35.858	0.096	27.835	42.820	0.120	37.961	57.703	0.173
	+TRACE	**21.089**	**33.602**	**0.087**	**24.232**	**38.267**	**0.103**	**30.524**	**47.089**	**0.139**
iTransformer	Base	21.906	34.509	0.092	25.139	39.516	0.107	31.897	49.474	0.141
	+TRACE	**20.450**	**32.793**	**0.084**	**22.712**	**36.375**	**0.095**	**27.079**	**42.581**	**0.118**
TimesNet	Base	25.050	39.946	0.108	26.321	41.912	0.116	29.188	46.055	0.133
	+TRACE	**23.416**	**37.346**	**0.098**	**24.646**	**39.870**	**0.107**	**26.875**	**43.205**	**0.120**
PatchTST	Base	25.770	38.135	0.110	28.676	43.199	0.132	38.353	57.110	0.181
	+TRACE	**21.906**	**33.940**	**0.091**	**24.241**	**37.911**	**0.107**	**30.177**	**46.248**	**0.138**
DLinear	Base	27.854	40.792	0.136	34.096	49.306	0.175	46.074	65.728	0.250
	+TRACE	**22.562**	**34.471**	**0.103**	**26.041**	**39.075**	**0.125**	**32.371**	**47.135**	**0.167**
FreTS	Base	22.382	34.921	0.093	26.479	40.803	0.116	34.553	52.449	0.156
	+TRACE	**20.637**	**33.012**	**0.085**	**23.349**	**37.077**	**0.100**	**28.214**	**43.822**	**0.126**
FEDFormer	Base	36.665	51.492	0.188	36.897	52.679	0.184	36.887	53.140	0.186
	+TRACE	**33.124**	**46.800**	**0.162**	**31.458**	**46.439**	**0.151**	**31.577**	**46.910**	**0.154**
Non-stat. Transformer	Base	24.985	39.798	0.109	26.230	41.803	0.115	28.998	45.932	0.134
	+TRACE	**23.496**	**37.385**	**0.099**	**24.641**	**39.964**	**0.107**	**26.888**	**43.480**	**0.121**
LightTS	Base	25.362	38.053	0.115	26.393	39.420	0.120	30.613	45.129	0.145
	+TRACE	**23.650**	**36.303**	**0.105**	**24.448**	**37.438**	**0.111**	**27.333**	**41.394**	**0.128**
Informer	Base	28.279	45.443	0.122	29.154	46.374	0.126	30.464	47.648	0.135
	+TRACE	**25.251**	**39.735**	**0.105**	**25.374**	**39.274**	**0.109**	**28.311**	**43.308**	**0.127**
Autoformer	Base	42.768	59.916	0.229	43.315	63.301	0.195	48.308	66.696	0.257
	+TRACE	**39.267**	**55.458**	**0.189**	**36.169**	**53.193**	**0.166**	**39.815**	**56.115**	**0.207**

Table 28: Complete PEMS results on PEMS08. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	6			12			24		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
TimeXer	Base	18.683	28.645	0.114	20.919	31.487	0.134	25.861	39.061	0.164
	+TRACE	**15.945**	**24.634**	**0.099**	**17.516**	**26.798**	**0.113**	**21.184**	**32.267**	**0.138**
TimeMixer	Base	14.559	22.851	0.092	15.940	24.773	0.102	17.957	28.006	0.117
	+TRACE	**14.139**	**22.341**	**0.089**	**15.322**	**24.040**	**0.098**	**16.952**	**26.798**	**0.110**
PAttn	Base	17.381	27.464	0.102	20.291	31.935	0.121	27.034	42.130	0.165
	+TRACE	**15.658**	**24.680**	**0.095**	**17.625**	**27.695**	**0.109**	**21.936**	**33.965**	**0.139**
iTransformer	Base	16.364	26.167	0.099	18.688	29.602	0.115	24.182	37.991	0.151
	+TRACE	**14.962**	**23.869**	**0.092**	**16.603**	**26.280**	**0.104**	**20.339**	**31.662**	**0.130**
TimesNet	Base	18.981	29.403	0.116	20.590	31.807	0.130	23.116	35.217	0.148
	+TRACE	**17.780**	**27.920**	**0.107**	**19.193**	**29.822**	**0.118**	**20.768**	**32.310**	**0.136**
PatchTST	Base	18.967	29.108	0.113	21.668	33.094	0.134	28.246	43.562	0.173
	+TRACE	**16.092**	**24.957**	**0.099**	**17.859**	**27.672**	**0.113**	**22.221**	**34.265**	**0.143**
DLinear	Base	23.102	34.846	0.144	26.421	39.072	0.171	33.929	49.463	0.226
	+TRACE	**17.672**	**26.887**	**0.110**	**19.469**	**29.175**	**0.125**	**23.583**	**34.702**	**0.156**
FreTS	Base	16.453	26.153	0.099	18.799	29.610	0.117	23.813	37.008	0.151
	+TRACE	**15.062**	**23.889**	**0.093**	**16.545**	**26.163**	**0.105**	**19.711**	**30.720**	**0.127**
FEDformer	Base	26.277	37.762	0.183	26.718	38.576	0.182	29.564	43.005	0.203
	+TRACE	**24.537**	**36.594**	**0.152**	**24.066**	**35.410**	**0.156**	**23.868**	**35.615**	**0.164**
Non-stat. Transformer	Base	17.464	27.001	0.108	18.808	28.916	0.120	21.376	33.074	0.138
	+TRACE	**16.874**	**26.407**	**0.103**	**18.044**	**28.022**	**0.112**	**19.762**	**31.002**	**0.130**
LightTS	Base	19.279	29.497	0.119	21.080	31.857	0.137	24.081	36.107	0.156
	+TRACE	**17.299**	**26.469**	**0.111**	**18.492**	**28.164**	**0.122**	**20.521**	**30.996**	**0.135**
Informer	Base	21.383	32.541	0.138	22.577	34.086	0.147	23.819	35.744	0.156
	+TRACE	**19.577**	**30.630**	**0.119**	**20.624**	**31.303**	**0.130**	**22.901**	**34.009**	**0.148**
Autoformer	Base	31.425	44.967	0.195	38.215	54.727	0.226	50.192	73.587	0.317
	+TRACE	**29.328**	**42.059**	**0.180**	**29.748**	**42.912**	**0.182**	**38.095**	**54.786**	**0.249**

Table 29: Complete EPF results on NP. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	MSE	MAE
TimeXer	Base	0.2434	0.2714
	+TRACE	0.2498	0.2733
TimeMixer	Base	0.2642	0.2899
	+TRACE	**0.2612**	**0.2841**
PAttn	Base	0.2774	0.3018
	+TRACE	**0.2700**	**0.2926**
iTransformer	Base	0.2599	0.2875
	+TRACE	**0.2580**	**0.2807**
TimesNet	Base	0.2432	0.2848
	+TRACE	**0.2403**	**0.2751**
PatchTST	Base	0.2814	0.2991
	+TRACE	**0.2781**	**0.2940**
DLinear	Base	0.2966	0.3115
	+TRACE	**0.2919**	**0.3091**
FreTS	Base	0.3105	0.3215
	+TRACE	**0.3049**	**0.3184**
FEDformer	Base	0.3284	0.3641
	+TRACE	**0.2973**	**0.3266**
Non-stat. Transformer	Base	0.5862	0.4916
	+TRACE	**0.4833**	**0.4267**
LightTS	Base	0.2991	0.3178
	+TRACE	**0.2919**	**0.3082**
Informer	Base	0.3037	0.3486
	+TRACE	**0.2747**	**0.3180**
Autoformer	Base	0.3823	0.3809
	+TRACE	**0.3284**	**0.3402**

Table 30: Complete EPF results on PJM. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	MSE	MAE
TimeXer	Base	0.0968	0.1944
	+TRACE	**0.0952**	**0.1930**
TimeMixer	Base	0.1176	0.2302
	+TRACE	**0.1095**	**0.2174**
PAttn	Base	0.1114	0.2160
	+TRACE	**0.1089**	**0.2111**
iTransformer	Base	0.0960	0.1958
	+TRACE	**0.0925**	**0.1928**
TimesNet	Base	0.1239	0.2285
	+TRACE	**0.0998**	**0.2037**
PatchTST	Base	0.1077	0.2129
	+TRACE	**0.1031**	**0.2065**
DLinear	Base	0.1042	0.2110
	+TRACE	**0.0954**	**0.1995**
FreTS	Base	0.1218	0.2305
	+TRACE	**0.1121**	**0.2156**
FEDformer	Base	0.1312	0.2371
	+TRACE	**0.1079**	**0.2110**
Non-stat. Transformer	Base	0.2582	0.3609
	+TRACE	**0.1726**	**0.2737**
LightTS	Base	0.1092	0.2119
	+TRACE	**0.1074**	**0.2075**
Informer	Base	0.1287	0.2128
	+TRACE	**0.1001**	**0.1953**
Autoformer	Base	0.1613	0.2662
	+TRACE	**0.1216**	**0.2283**

Table 31: Complete EPF results on BE. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	MSE	MAE
TimeXer	Base	0.3794	0.2423
	+TRACE	**0.3716**	**0.2414**
TimeMixer	Base	0.4146	0.2758
	+TRACE	**0.4135**	**0.2746**
PAttn	Base	0.3998	0.2661
	+TRACE	**0.3959**	**0.2653**
iTransformer	Base	0.4042	0.2767
	+TRACE	0.4103	0.2777
TimesNet	Base	0.3899	0.2696
	+TRACE	**0.3881**	**0.2684**
PatchTST	Base	0.4041	0.2610
	+TRACE	**0.3944**	**0.2592**
DLinear	Base	0.4591	0.3147
	+TRACE	**0.4351**	**0.3084**
FreTS	Base	0.4370	0.3106
	+TRACE	**0.4240**	**0.2988**
FEDformer	Base	0.4739	0.3324
	+TRACE	**0.4157**	**0.3044**
Non-stat. Transformer	Base	0.6884	0.4620
	+TRACE	**0.6070**	**0.3804**
LightTS	Base	0.4436	0.2967
	+TRACE	**0.4268**	**0.2864**
Informer	Base	0.4841	0.3247
	+TRACE	**0.4417**	**0.3130**
Autoformer	Base	0.5049	0.3219
	+TRACE	**0.4039**	**0.2862**

Table 32: Complete EPF results on FR. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	MSE	MAE
TimeXer	Base	0.3960	0.2119
	+TRACE	**0.3926**	**0.2069**
TimeMixer	Base	0.4445	0.2447
	+TRACE	**0.4047**	**0.2332**
PAttn	Base	0.4692	0.2602
	+TRACE	**0.4604**	**0.2510**
iTransformer	Base	0.4203	0.2279
	+TRACE	**0.4170**	**0.2271**
TimesNet	Base	0.4233	0.2295
	+TRACE	**0.4113**	**0.2259**
PatchTST	Base	0.4135	0.2236
	+TRACE	**0.4087**	**0.2156**
DLinear	Base	0.4196	0.2537
	+TRACE	**0.3943**	**0.2225**
FreTS	Base	0.4030	0.2356
	+TRACE	**0.3872**	**0.2243**
FEDformer	Base	0.4563	0.2762
	+TRACE	**0.3969**	**0.2460**
Non-stat. Transformer	Base	0.6214	0.3883
	+TRACE	**0.5330**	**0.2953**
LightTS	Base	0.4734	0.2581
	+TRACE	**0.4691**	**0.2506**
Informer	Base	0.4785	0.2680
	+TRACE	**0.4645**	**0.2507**
Autoformer	Base	0.6188	0.3017
	+TRACE	**0.4783**	**0.2608**

Table 33: Complete EPF results on DE. Base and +TRACE are reported separately for every backbone and horizon; bold marks a lower corrected error. No result in this table is averaged over backbones or horizons.

Backbone	System	MSE	MAE
TimeXer	Base	0.4581	0.4224
	+TRACE	0.4610	0.4258
TimeMixer	Base	0.4690	0.4370
	+TRACE	0.4736	0.4371
PAttn	Base	0.4676	0.4348
	+TRACE	0.4693	0.4353
iTransformer	Base	0.4660	0.4353
	+TRACE	**0.4580**	**0.4323**
TimesNet	Base	0.4747	0.4331
	+TRACE	0.4752	0.4332
PatchTST	Base	0.5099	0.4521
	+TRACE	**0.5071**	**0.4507**
DLinear	Base	0.5081	0.4567
	+TRACE	**0.4970**	**0.4433**
FreTS	Base	0.5109	0.4517
	+TRACE	0.5149	0.4524
FEDformer	Base	0.6164	0.5268
	+TRACE	**0.5731**	**0.4955**
Non-stat. Transformer	Base	0.5171	0.4599
	+TRACE	0.5258	0.4700
LightTS	Base	0.4891	0.4421
	+TRACE	0.4931	0.4433
Informer	Base	0.5849	0.4930
	+TRACE	**0.5568**	**0.4764**
Autoformer	Base	0.5967	0.5069
	+TRACE	**0.5606**	**0.4871**

G ADDITIONAL FORECASTING SYSTEMS

Table 35 gives a compact view of every dataset–system cell. The entry is the smallest reduction among the metrics reported for that cell, so positive values correspond exactly to the strict criterion used in the main text. Each static ensemble is one forecasting system, and each AUTOGLUON or CHRONOS-BOLT variant is one trained forecaster. The accompanying machine-readable results retain the absolute metrics.

Table 34: **Correction of additional ensemble, AutoML, and foundation-model systems.** Each system is evaluated once per dataset at horizon 96 (long-term) or 24 (PEMS and EPF). "Strict" requires every applicable metric to improve. Entries Δ are mean paired error reductions in percent (higher is better). MSE uses 12 cells per system except AutoGluon (10); MAE uses 16 except AutoGluon (14); RMSE and MAPE use four.

System	Eval.	Strict	Rate	ΔMSE	ΔMAE	ΔRMSE	ΔMAPE
FORWARD SELECTION	16	13	81.2%	0.87	1.56	3.12	5.35
PORTFOLIO ENSEMBLE	16	12	75.0%	1.17	1.57	3.06	5.62
ZEROSHOT ENSEMBLE	16	13	81.2%	1.40	2.68	5.44	8.01
AUTOGLUON HQ	14	13	92.9%	13.10	14.87	27.32	28.88
CHRONOS-BOLT FT	16	14	87.5%	8.85	8.61	24.42	21.62
CHRONOS-BOLT ZS	16	16	100.0%	10.74	9.36	26.97	25.50
All systems	94	81	86.2%	5.82	6.26	15.05	15.83

Table 35: **Dataset-level robustness for additional systems.** Each entry is the smallest paired error reduction across the metrics reported for that cell; it is positive exactly when the cell is a strict win. Higher is better.

Family	Dataset	FORWARD SELECTION	PORTFOLIO ENSEMBLE	ZEROSHOT ENSEMBLE	AUTOGLUON HQ	CHRONOS-BOLT FT	CHRONOS-BOLT ZS
Long-term	ETTh1	0.03	-0.20	0.89	6.69	2.25	1.74
	ETTh2	-0.12	-0.24	-0.35	2.79	2.56	1.41
	ETTm1	1.15	1.35	1.85	13.45	13.86	11.51
	ETTm2	0.72	0.64	0.91	7.41	5.37	7.44
	weather	0.55	-0.14	-0.03	5.75	9.34	10.75
	electricity	-0.04	0.71	1.41	OOT	1.94	1.01
	traffic	1.79	2.71	4.79	OOT	7.36	2.94
PEMS	PEMS03	2.55	2.51	4.52	26.91	19.49	24.56
	PEMS04	1.79	1.82	3.98	24.81	19.47	24.94
	PEMS07	2.91	2.64	5.13	27.03	23.55	26.12
	PEMS08	5.24	5.28	8.13	28.39	23.10	25.13
EPF	NP	-0.78	-0.69	-1.01	17.44	0.85	0.48
	PJM	0.36	0.43	0.42	20.13	0.77	0.61
	BE	0.38	0.90	0.47	-6.01	-0.22	0.33
	FR	0.04	0.21	0.33	4.71	0.03	0.66
	DE	0.11	0.55	0.20	9.11	-1.04	4.05

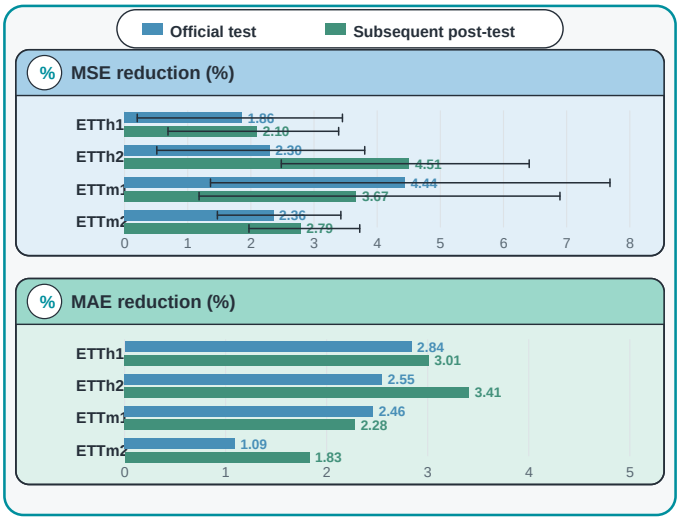

Figure 6: Pooled MSE and MAE reductions over matched TIMEMIXER forecasts across three seeds. The post-test timeline is chronologically subsequent to the standard split and is excluded from model and correction selection. Horizontal lines on MSE bars are 95% paired moving-block-bootstrap intervals. Higher is better.

H FULL ETT PREQUENTIAL RESULTS

Table 36: Pooled MSE reductions and 95% paired moving-block-bootstrap intervals. Intervals are candidate-minus-baseline MSE, so values below zero favor TRACE.

	Official test		Subsequent post-test	
Dataset	Reduction	95% interval	Reduction	95% interval
ETTh1	1.86%	$[-0.013169, -0.000768]$	2.10%	$[-0.017964, -0.003646]$
ETTh2	2.30%	$[-0.011055, -0.001485]$	4.51%	$[-0.011830, -0.004583]$
ETTm1	4.44%	$[-0.024903, -0.004411]$	3.67%	$[-0.034576, -0.005918]$
ETTm2	2.36%	$[-0.006030, -0.002587]$	2.79%	$[-0.004747, -0.002511]$

Table 37: Per-seed MSE reductions (%) for the ETT prequential study. All corresponding MAE values also improve.

	Official test			Subsequent post-test		
Dataset	2021	2022	2023	2021	2022	2023
ETTh1	1.59	1.49	2.48	2.39	2.08	1.84
ETTh2	4.31	1.18	1.37	6.29	3.69	3.49
ETTm1	2.72	8.35	2.06	2.12	5.50	3.39
ETTm2	2.36	2.54	2.17	2.62	2.72	3.02

The intervals use 5,000 moving-block bootstrap resamples of length 96 after averaging aligned seed losses at each origin. The later timeline begins on 2018-02-21 for ETTh and ends on 2018-06-25 for ETTm. ETTh uses the fixed stack specified in the experiment protocol; ETTm retrieval adds an official-test residual only after its target interval has elapsed.