ECHO: CORRECTING TIME SERIES FOUNDATION MODEL FORECASTS WITH HISTORICAL ANALOGUES AND ERRORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time Series Foundation Models (TSFMs) provide general-purpose forecasting capabilities, yet their forecasts can remain misaligned with the temporal characteristics of a target domain. This motivates lightweight forecast correction even when only black-box access to the TSFM is available. Our key observation is that two forms of historical information provide complementary correction signals: historical analogues reveal how similar contexts evolved, while past forecast errors reveal a TSFM's characteristic deviations in the target domain. Based on this observation, we propose ECHO, a black-box forecast correction framework that learns from both sources. ECHO uses AnalogE to assess retrieved historical analogues together with the current context and base forecast, learning transferable corrections through cross-domain pretraining. In parallel, RecentE learns target-specific residual corrections from past forecast errors and refines them as causally available feedback accumulates. ECHO dynamically combines the two corrections according to their observed forecasting performance. Across 5 representative TSFMs and 9 real-world datasets, ECHO reduces MSE in all 45 backbone–dataset settings, yielding an average relative reduction of 9.26% in macro-average MSE. Its gains also extend to long-horizon forecasting.

1 INTRODUCTION

Time series forecasting is widely used in real-world applications, including energy management (Hong & Fan, 2016), traffic management (Vlahogianni et al., 2014), financial analysis (Fischer & Krauss, 2018), and industrial monitoring (Malhotra et al., 2015). With advances in deep learning, time series forecasting has evolved from classical statistical approaches (Winters, 1960; Box et al., 2015) to deep learning methods (Zhou et al., 2021; Zeng et al., 2023). Conventional deep learning models are typically trained on specific datasets to capture domain-specific temporal patterns (Wu et al., 2021; Zhang et al., 2026). Inspired by large-scale pretraining in natural language processing (Devlin et al., 2019; Brown et al., 2020), Time Series Foundation Models (TSFMs), such as TimesFM (Das et al., 2023), Chronos (Ansari et al., 2024), and Moirai (Woo et al., 2024), are pretrained on multi-domain time series to support general-purpose forecasting. However, time series from different domains exhibit substantial differences in temporal patterns and distributional properties (Godahewa et al., 2021). Pretrained TSFMs may not fully accommodate this heterogeneity, limiting their forecasting performance in specific domains (Aksu et al., 2024). These limitations highlight the need for efficient target-domain adaptation.

Leveraging information available in the target domain is an important avenue for adapting TSFMs (Das et al., 2024). Existing approaches leverage cross-variate dependencies (Ekambaram et al., 2024), exogenous covariates (Arango et al., 2025), or textual context (Chen et al., 2026) to enrich the information available for forecasting. Among the available information sources, historical observations and past model forecasts can be accumulated during routine forecasting, providing a direct basis for target-domain adaptation. Prior work exploits this history by retrieving relevant historical cases to provide references for how the current series may evolve (Tire et al., 2024; Zhang et al., 2025; Ning et al., 2026), or by using past forecasts and their observed outcomes to improve subsequent predictions through forecast combination or learned corrections (Lee et al., 2025; Liang et al., 2026). However, these history-based methods typically focus on one form of historical information.

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.

Historical analogues provide references for how similar contexts evolved, while past forecast errors reveal how a particular backbone deviates from observed outcomes in the target domain. Their complementary roles motivate us to explore how to learn effective corrections from these two distinct forms of historical evidence and coordinate their contributions to the current forecast.

Building on this perspective, we propose ECHO, a framework for correcting TSFM forecasts using historical analogues and past forecast errors. To account for their distinct properties, ECHO employs two correction modules. Historical Analog Evidence (AnalogE) uses historical analogues, comprising past contexts similar to the current context and their observed futures. By assessing these analogues together with the current context and base forecast, it constructs a historical reference and determines how strongly to adjust the base forecast toward it. These correction strategies are learned through cross-domain pretraining. Adaptive Recent Evidence (RecentE) learns residual patterns from past forecast errors, conditioned on the current context and base forecast, to provide corrections tailored to the target domain. It is initialized using target-domain training data and periodically refined as newly matured forecasting feedback becomes available. To coordinate the two corrections, Temporal Reliability Aggregation (TempRA) estimates their reliability from historical forecasting performance and adjusts their contributions to the final forecast. ECHO keeps the backbone frozen and accesses only its inputs and forecast outputs, enabling black-box target-domain adaptation.

Our main contributions are summarized as follows:

- Complementary historical evidence for forecast correction. We jointly exploit historical analogues and past forecast errors for TSFM adaptation, bringing together references for temporal evolution and evidence of model-specific errors in the target domain to guide forecast correction.
- An evidence-driven correction framework. We propose ECHO, which learns transferable corrections from historical analogues through cross-domain pretraining (AnalogE) and target-specific residual corrections from past forecast errors (RecentE). It coordinates the two corrections using reliability estimates derived from historical forecasting performance (TempRA), enabling target-domain adaptation with a black-box TSFM.
- Extensive empirical evaluation. Across 5 representative TSFMs and 9 real-world datasets, ECHO reduces MSE in all 45 backbone–dataset settings, with an average relative reduction of 9.26% in macro-average MSE. These improvements also extend to long-horizon forecasting.

2 Related Work

2.1 Time Series Foundation Models

Time Series Foundation Models (TSFMs) aim to learn transferable temporal knowledge through large-scale pretraining (Liu et al., 2024d; Goswami et al., 2024). Existing research explores both adapting pretrained language models for time series forecasting and developing native time series foundation models. Language-model adaptation approaches include partial finetuning (Zhou et al., 2023), numerical encoding (Gruver et al., 2023), domain instructions and cross-modal alignment (Liu et al., 2024b), and input reprogramming (Jin et al., 2024). Native TSFMs adopt diverse architectures and representations: Chronos (Ansari et al., 2024) uses discretized autoregressive modeling, TimesFM (Das et al., 2023) and Timer (Liu et al., 2024d) adopt patch-based representations, and Moirai (Woo et al., 2024) supports unified modeling across frequencies and variables. Subsequent developments include multivariate and covariate-informed forecasting in Chronos-2 (Ansari et al., 2025), flow-matching-based probabilistic forecasting in Sundial (Liu et al., 2025b), lightweight architectures in TTM (Ekambaram et al., 2024), sparse mixture-of-experts modeling in Moirai-MoE (Liu et al., 2024c), and long-context modeling in Timer-S1 (Liu et al., 2026). However, differences in temporal patterns and distributions across domains can limit the performance of frozen TSFMs in specific target environments, motivating target-domain adaptation (Aksu et al., 2024).

2.2 Retrieval-Augmented Time Series Forecasting

Retrieval-augmented time series forecasting retrieves relevant historical sequences to provide additional context or reference information for the current forecast (Ning et al., 2026; Tire et al., 2024). Existing studies explore different ways to incorporate retrieved information: ReTime (Jing et al.,

2022) investigates retrieval-based forecasting, RATD (Liu et al., 2024a) integrates retrieved information into diffusion-based forecasting, TimeRAG (Yang et al., 2025) augments language-model forecasting contexts with retrieved sequences, and RAFT (Han et al., 2025) aggregates retrieved historical futures across multiple temporal scales and combines them with the current input through a linear predictor. Recent studies further integrate retrieval mechanisms with TSFMs. TimeRAF introduces learnable retrieval and channel prompting (Zhang et al., 2025), while TS-RAG combines retrieved representations with internal TSFM representations (Ning et al., 2026). These integration mechanisms require interaction with internal backbone representations, limiting their direct applicability to black-box TSFMs. Naive RAF (Tire et al., 2024) supports black-box forecasting by augmenting the input context with retrieved examples, but its gains depend on the backbone's ability to exploit this information. Learning to use historical analogues to correct forecasts without accessing backbone internals therefore merits further study.

2.3 Forecast Adaptation and Correction

Beyond historical sequences, past forecasts and their observed outcomes provide another source of information for adaptation. Existing adaptation methods differ in their access requirements to the backbone. SOLID (Chen et al., 2024) finetunes the prediction layer using contextually similar samples, while DSOF (Lau et al., 2025) updates forecasting models through fast and slow online learning streams. TAFAS (Kim et al., 2025) and Ada-X (Liang et al., 2026) keep the backbone frozen, but optimizing their input-side modules still requires backpropagation through the backbone. Black-box approaches instead improve forecasts through external combination or correction without accessing backbone internals. ELF (Lee et al., 2025) combines a lightweight online forecaster with foundation-model predictions; Ada-Y (Liang et al., 2026), the output-side variant of δ-Adapter, learns corrections from base forecasts; and ORCA (Dai et al., 2026) learns residual corrections conditioned on both the input context and the base forecast.

Taken together, these studies demonstrate the value of historical analogues and forecasting feedback for adaptation, but generally study them separately. How to learn corrections tailored to the distinct properties of these two forms of evidence and coordinate their contributions in black-box TSFMs warrants further investigation.

3 Methodology

Figure 1 presents an overview of ECHO. AnalogE and RecentE learn corrections from historical analogues and past forecast errors, respectively, while TempRA coordinates the two corrections based on their historical forecasting performance.

3.1 Problem Formulation

We consider target-domain forecasting with a black-box TSFM. For time series with D variates, forecast correction is performed independently for each variate. For clarity, we present the formulation for a single variate and omit the variate index. At each forecast origin t, observations up to and including time t are available. The input $X_t \in \mathbb{R}^L$ contains the latest L observations, and the target $Y_t \in \mathbb{R}^H$ covers the next H time steps, $t+1, \ldots, t+H$. Let $Y_t^{\text{base}} \in \mathbb{R}^H$ denote the base forecast produced by the TSFM for this variate.

Let $\mathcal{H}_t$ denote the historical evidence available for this variate at forecast origin t. An external correction framework G uses this evidence to refine the base forecast:

$$\hat{Y}_t = G\big(X_t, Y_t^{\text{base}}; \mathcal{H}_t\big)\,. \tag{1}$$

3.2 Historical Analog Evidence

Historical Analog Evidence (AnalogE) corrects base forecasts using retrieved historical analogues. It constructs a historical reference and estimates a gate that controls how strongly the base forecast is adjusted toward this reference.

Given X_t, AnalogE retrieves K context–future pairs (X_i, Y_i) with fully observed futures from a fixed training-set memory, along with their retrieval distances d_i. Retrieval details are provided in

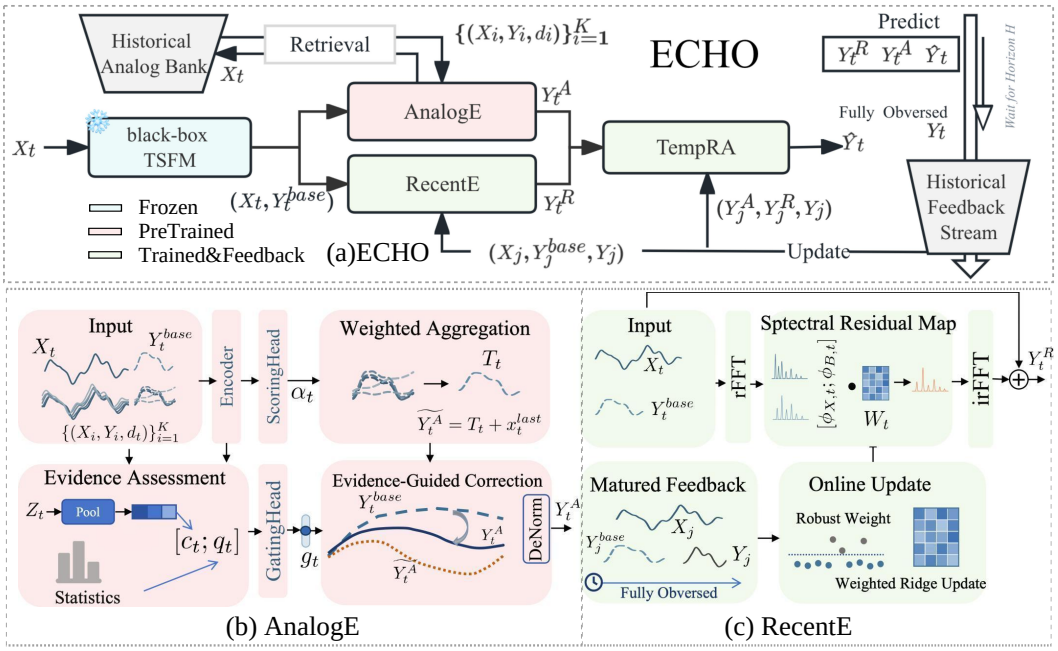


Figure 1: Overview of ECHO. (a) Overall framework. (b) AnalogE constructs a weighted historical template and estimates a gate to control the correction of the base forecast. (c) RecentE predicts spectral residuals and updates its residual map using fully observed feedback.

Appendix A.1. AnalogE operates on window-normalized sequences, and corrected forecasts are mapped back to the dataset-standardized scale. We subtract the last observations x_i^{last} and x_t^{last} of the corresponding contexts from each historical future Y_i and the base forecast Y_t^{base}, respectively, obtaining the historical transitions $\Delta_i \in \mathbb{R}^H$ and the base forecast transition $p_t^{\text{base}} \in \mathbb{R}^H$.

Historical Template Construction. Under non-stationarity, similar contexts may lead to different futures, making retrieval distances alone insufficient to assess the usefulness of historical candidates (Zhou et al., 2026). AnalogE therefore reassesses retrieved context–future pairs jointly with the current context and base forecast.

An encoder integrates this information and models relationships among candidates:

$$Z_t = \text{Encoder}\left(X_t, p_t^{\text{base}}, \{(X_i, \Delta_i)\}_{i=1}^K, \mathbf{d}_t\right), \tag{2}$$

where $\mathbf{d}_t = (d_1, \ldots, d_K)$ and $Z_t \in \mathbb{R}^{K \times d_z}$ contains the candidate representations, with $Z_{t,i}$ denoting the d_z-dimensional representation of candidate i.

The resulting representations Z_t and retrieval distances $\mathbf{d}_t$ are passed to a trainable scoring head, which refines distance-based scores through bounded adjustments and internally applies softmax across the K candidates. Its output weights are used to construct the historical template:

$$\boldsymbol{\alpha}_t = \text{ScoringHead}(Z_t, \mathbf{d}_t), \qquad T_t = \sum_{i=1}^{K} \alpha_{t,i} \Delta_i, \tag{3}$$

where $T_t \in \mathbb{R}^H$ aggregates historical transitions according to their estimated usefulness for the current forecast.

Evidence Gate Estimation. Disagreement between the historical reference and the base forecast does not necessarily warrant a substantial correction. AnalogE pools representations across candidates to obtain a candidate summary representation $c_t \in \mathbb{R}^{d_z}$ and concatenates it with explicit statistics $q_t \in \mathbb{R}^{d_q}$ to form the gating input e_t. These statistics characterize historical support quality, correction evidence ambiguity, and foundation–experience disagreement. Their definitions are provided in Appendix A.2.

The gating head estimates correction strength from e_t:

$$g_t = \mathrm{GatingHead}(e_t) = g_{\max}\sigma\left(\mathrm{MLP}(e_t)\right), \tag{4}$$

where σ is the sigmoid function and $g_{\max}$ bounds the gate g_t.

Evidence-Guided Correction. AnalogE aligns the historical transition template T_t with the latest observation x_t^{last}. The gate g_t controls how far the base forecast moves toward this aligned reference, yielding the corrected forecast Y_t^A:

$$Y_t^A = Y_t^{\mathrm{base}} + g_t\left(x_t^{\mathrm{last}} + T_t - Y_t^{\mathrm{base}}\right). \tag{5}$$

Through cross-domain pretraining, AnalogE learns to construct historical templates and estimate correction gates that jointly guide forecast correction.

3.3 Adaptive Recent Evidence

Adaptive Recent Evidence (RecentE) learns residual patterns in the target environment to correct base forecasts and refines its corrections as new forecasting feedback becomes available. For a historical instance j with a fully observed future, the residual target is $r_j = Y_j - Y_j^{\mathrm{base}}$.

Spectral Residual Adaptation. RecentE estimates residuals conditioned on the current context X_t and base forecast Y_t^{base}. We center X_t by its mean and Y_t^{base} by the latest context observation, transform both into the frequency domain, and concatenate their spectra into $\phi_t \in \mathbb{C}^{n_L+n_H}$. Here, $n_L = \lfloor L/2 \rfloor + 1$ and $n_H = \lfloor H/2 \rfloor + 1$ are the corresponding numbers of frequency coefficients.

A complex linear map $W_t \in \mathbb{C}^{(n_L+n_H)\times n_H}$ predicts the residual spectrum:

$$\hat{z}_t = \phi_t W_t. \tag{6}$$

The predicted spectrum is transformed back to a length-H residual sequence and added to the base forecast:

$$Y_t^R = Y_t^{\mathrm{base}} + \mathrm{irFFT}(\hat{z}_t, n = H). \tag{7}$$

Residual Map Fitting and Refinement. RecentE initializes the residual map on training instances and periodically updates it using selected mature records from the feedback stream $\mathcal{M}_t$ defined in Section 3.4. The map is fitted to residual target spectra through weighted ridge regression, alternating between estimating robust sample weights from fitting errors and refitting the map with these weights fixed. This procedure limits the influence of large fitting errors. The fitting objective and online update details are provided in Appendix A.3.

3.4 Temporal Reliability Aggregation

Temporal Reliability Aggregation (TempRA) assesses the reliability of AnalogE and RecentE from historical forecasting performance and adjusts their contributions to the final forecast.

At forecast origin t, the available historical feedback forms the following record set:

$$\mathcal{M}_t = \left\{ \left(X_j, Y_j^{\mathrm{base}}, Y_j^A, Y_j^R, Y_j\right) \mid j + H \le t \right\}. \tag{8}$$

The condition $j + H \le t$ ensures that each record's target horizon is fully observed. From this feedback stream, RecentE draws historical inputs, base forecasts, and observed outcomes $(X_j, Y_j^{\mathrm{base}}, Y_j)$ for residual learning, while TempRA draws the two corrected forecasts and their observed outcomes (Y_j^A, Y_j^R, Y_j) for reliability estimation.

TempRA constructs three temporal views from this feedback stream. The long view summarizes accumulated forecasting performance, while the short view uses a recent subset of sparsely selected feedback records. The recent view uses more densely sampled feedback to capture local performance differences. Within each view, TempRA computes prediction errors on the selected records and uses these errors to assign AnalogE a weight $w_t^v \in [0, 1]$, where $v \in \{L, S, R\}$. The corresponding RecentE weight is $1 - w_t^v$. View-specific record selection and weight estimation are detailed in Appendix A.4.

We combine the three weights and aggregate the two forecasts:

$$w_t = \lambda_L w_t^L + \lambda_S w_t^S + \lambda_R w_t^R, \qquad \hat{Y}_t = w_t Y_t^A + (1 - w_t) Y_t^R, \tag{9}$$

where $\lambda_v \ge 0$ and $\sum_v \lambda_v = 1$.

Table 1: Cross-backbone forecasting MSE on 9 datasets with context length $L = 512$ and forecast horizon $H = 64$. Ori denotes the original forecasts of each frozen TSFM, and ECHO denotes its corrected forecasts. The better result within each pair is bolded. Overall reports the relative change in macro-average MSE over the 9 datasets.

Models	**TimesFM 2.5**		**Chronos-2**		**Chronos-Bolt**		**Moirai 2.0**		**Sundial**	
MSE	Ori	ECHO	Ori	ECHO	Ori	ECHO	Ori	ECHO	Ori	ECHO
ETTh1	.355	**.346**	.377	**.348**	.362	**.343**	.367	**.340**	.337	**.336**
ETTh2	.257	**.242**	.266	**.246**	.252	**.240**	.266	**.241**	.242	**.233**
ETTm1	.280	**.263**	.309	**.274**	.311	**.271**	.322	**.273**	.268	**.257**
ETTm2	.148	**.136**	.149	**.138**	.149	**.139**	.155	**.138**	.149	**.135**
Exchange	.059	**.056**	.060	**.058**	.067	**.064**	.060	**.058**	.082	**.067**
Weather	.135	**.124**	.143	**.128**	.152	**.134**	.148	**.127**	.136	**.123**
Electricity	.120	**.116**	.116	**.113**	.113	**.111**	.117	**.114**	.107	**.106**
PEMS08	.097	**.089**	.232	**.179**	.289	**.190**	.112	**.098**	.094	**.088**
Solar	.208	**.180**	.147	**.145**	.469	**.359**	.181	**.159**	.210	**.189**
Overall	–	**-6.60%**	–	**-9.36%**	–	**-14.43%**	–	**-10.38%**	–	**-5.55%**

4 Experimental Setup

Datasets and Data Usage. We evaluate ECHO on nine time series datasets: ETTh1, ETTh2, ETTm1, and ETTm2 (Zhou et al., 2021); Exchange Rate, Electricity, and Solar-Energy (Lai et al., 2018); Weather (Wu et al., 2021); and PEMS08 (Song et al., 2020). These datasets cover electricity, exchange rates, weather, solar power generation, and traffic flow, with diverse sampling frequencies and temporal dynamics. Following TS-RAG (Ning et al., 2026), we pretrain AnalogE using queries sampled from the Chronos pretraining corpus (Ansari et al., 2024) and keep it frozen on all downstream datasets.

Forecasting Backbones. We evaluate ECHO with five TSFMs: TimesFM 2.5 (Das et al., 2023), Chronos-2 (Ansari et al., 2025), Chronos-Bolt (Ansari et al., 2024), Moirai 2.0 (Liu et al., 2025a) and Sundial (Liu et al., 2025b). All backbones remain frozen throughout evaluation.

Evaluation Protocol and Implementation. Unless otherwise specified, experiments use TimesFM 2.5 as the frozen backbone, with a context length of $L = 512$, a forecast horizon of $H = 64$, and $K = 20$ retrieved cases. Online updates use only feedback whose complete target horizon is observed by the current forecast origin. We use Mean Squared Error (MSE) as the primary evaluation metric. Mean Absolute Error (MAE) results are reported in Appendix C. Additional experimental details are provided in Appendix B.

5 Main Results and Model Analysis

We evaluate ECHO's forecasting performance, examine the effectiveness of its components and learning designs, and assess its computational efficiency.

5.1 Main Results

Cross-Backbone Forecasting. To evaluate the generality of ECHO as a forecast correction framework, we compare the original and ECHO-corrected forecasts of 5 frozen TSFMs on 9 datasets. As shown in Table 1, ECHO achieves lower MSE in all 45 dataset–backbone configurations, with relative changes in macro-average MSE ranging from -14.43% to -5.55%. These results support the applicability of ECHO across TSFMs with different architectures and forecasting behaviors.

Long-Horizon Forecasting. We next examine whether ECHO's improvements extend to longer forecast horizons by comparing original and ECHO-corrected forecasts on 9 datasets with $H \in \{96, 192, 336, 720\}$. For horizons beyond AnalogE's fixed training length, we apply retrieval and correction in a rolling manner without retraining AnalogE. As shown in Table 2, ECHO achieves lower MSE in most dataset–horizon configurations. The relative change in macro-average MSE

Table 2: Long-horizon MSE of TimesFM 2.5 with $L = 512$ and $H \in \{96, 192, 336, 720\}$. ETTh Avg. and ETTm Avg. average the corresponding ETT datasets. Lower is better; pairwise best results are bolded. Overall reports the relative change in macro-average MSE over the nine original datasets.

Dataset	96		192		336		720	
	Ori	ECHO	Ori	ECHO	Ori	ECHO	Ori	ECHO
ETTh Avg.	.337	**.320**	.391	**.367**	.413	**.390**	.412	**.399**
ETTm Avg.	.238	**.222**	.293	**.273**	.340	**.322**	.422	**.400**
Exchange	.091	**.085**	.194	**.181**	.360	**.326**	1.046	**.944**
Weather	.159	**.144**	.209	**.189**	.275	**.246**	.362	**.332**

Dataset *cont.*	96		192		336		720	
	Ori	ECHO	Ori	ECHO	Ori	ECHO	Ori	ECHO
PEMS08	.109	**.098**	.143	**.126**	.177	**.156**	.249	**.199**
Electricity	.131	**.127**	.148	**.146**	**.165**	.166	**.205**	.209
Solar	.222	**.193**	.255	**.231**	.292	**.230**	.315	**.235**
Overall	**-6.91%**		**-7.11%**		**-8.10%**		**-8.50%**	

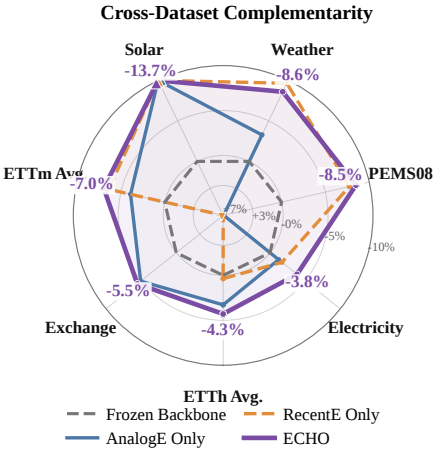


Figure 2: Relative MSE changes (%) of AnalogE-only, RecentE-only, and ECHO against TimesFM 2.5 across 9 datasets, averaging ETT pairs. Larger reductions appear farther outward. Extreme values are clipped for visualization and marked with triangles.

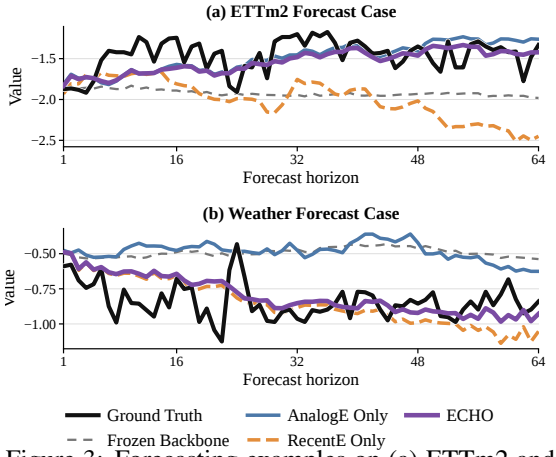


Figure 3: Forecasting examples on (a) ETTm2 and (b) Weather. AnalogE-only better captures the sequence level in (a), while RecentE-only better captures the downward shift in (b). ECHO combines the complementary corrections from both evidence sources.

decreases from -6.91% at $H = 96$ to -8.50% at $H = 720$, supporting the effectiveness of ECHO over longer forecast horizons.

5.2 Complementarity of Evidence Sources

Historical and recent evidence offer complementary strengths, motivating their joint use for robust forecast correction. To assess this, we compare three configurations: AnalogE-only, which corrects the base forecast using only historical evidence from AnalogE; RecentE-only, which corrects the base forecast using only forecast-error evidence from RecentE; and full ECHO, which combines both sources.

Figure 2 summarizes relative MSE changes against the frozen backbone across nine datasets. Larger reductions appear farther outward. AnalogE-only and RecentE-only exhibit different strengths across datasets, while ECHO achieves more consistent improvements, supporting the benefit of using both evidence sources jointly (Appendix C.3).

Figure 3 illustrates the complementary strengths of the two evidence sources. In Figure 3(a), AnalogE-only better captures the overall sequence level, whereas RecentE-only increasingly underestimates the target values. In Figure 3(b), RecentE-only better captures the downward shift, whereas AnalogE-only overestimates the sequence level.

Using identical branch forecasts, TempRA also achieves lower macro-average MSE than validation-selected fixed fusion (Appendix C.6).

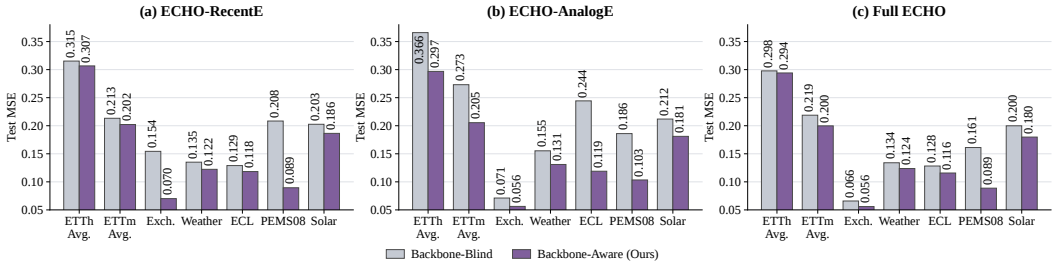


Figure 4: Role of the base forecast. Backbone-Blind predicts directly without the base forecast. Panels (a) and (b) compare the recent and historical analogue branches. Panel (c) compares ECHO. ECL denotes Electricity. Lower MSE indicates better forecasting performance.

5.3 Role of the Base Forecast

We examine whether evidence is more effective for correcting the backbone forecast than for independent prediction. We construct Backbone-Aware and Backbone-Blind variants of AnalogE, RecentE, and full ECHO. For AnalogE, the aware variant corrects the backbone forecast using retrieved cases, while the blind variant predicts directly from them. For RecentE, the aware variant predicts residual corrections, while the blind variant predicts target values from the context. For full ECHO, both variants use TempRA to aggregate the outputs of AnalogE and RecentE: corrected forecasts in the aware variant and independent predictions in the blind variant. As shown in Figure 4, the Backbone-Aware variants outperform their Backbone-Blind counterparts for both individual branches and full ECHO.

5.4 Evidence Construction

We evaluate AnalogE and RecentE against alternative methods for using the corresponding historical evidence. Full ECHO is also included as a reference. For historical analog evidence, we compare AnalogE with a black-box variant of TS-RAG (Ning et al., 2026) and Naive RAF (Tire et al., 2024). Since the original TS-RAG fuses retrieved representations with internal TSFM representations before the output projection, we implement a variant compatible with black-box access (BB-TS-RAG). For RAF, we use Naive RAF with the backbone kept frozen. Table 3(a) shows that AnalogE outperforms both retrieval baselines in 3 of the 4 dataset groups, supporting its effectiveness in constructing corrections from historical analogues.

For adaptive recent evidence, we compare RecentE with ELF (Lee et al., 2025) and Ada-Y (Liang et al., 2026). ELF retains its original forecaster and dynamic ensemble weighting. We additionally evaluate ELF+Y^{base} and Ada-Y+X, which only add the base forecast and context to their respective predictor inputs, leaving the remaining components unchanged. These comparisons examine correction designs with matched predictor inputs. Table 3(b) shows that RecentE achieves lower MSE than all four alternatives on ETTm and Weather, including those with the same predictor inputs. These results support the effectiveness of RecentE's design beyond input access alone. Detailed baseline configurations and additional results are provided in Appendices B.4, C.4, and C.5.

5.5 AnalogE Pretraining and RecentE Online Updates

We find that pretraining AnalogE on a multi-domain dataset enables it to better use retrieved historical cases for forecast correction than training only on target-domain data. To assess this benefit, we compare the same AnalogE architecture under these two training settings. As shown in Figure 5(a), cross-domain pretraining improves AnalogE on all evaluated datasets.

For RecentE, we examine whether online updates with matured forecasting feedback help track error patterns specific to the foundation model and target environment. We compare continued online updates with a frozen residual map, both starting from the same target-domain initialization. Figure 5(b) shows that online updates yield small but consistent gains, supporting the benefit of refining residual corrections as new feedback becomes available.

Table 3: MSE comparison of (a) historical and (b) recent evidence methods with TimesFM 2.5. X and Y^{base} indicate predictor inputs. ECHO is included as a full-framework reference. ETTh and ETTm average their respective dataset pairs; Exch. denotes Exchange Rate. Lower is better.

(a) Historical Analog Evidence

Method	ETTh	ETTm	Exch.	Weather
Backbone	.306	.214	.059	.135
BB-TS-RAG	.299	.212	.074	.139
Naive RAF	.299	.213	.058	.130
AnalogE	.297	.205	.056	.131
ECHO	.294	.200	.056	.124

(b) Adaptive Recent Evidence

Method	X	Y^{base}	ETTh	ETTm	Exch.	Weather
ELF	✓	–	.295	.211	.067	.133
ELF + Y^{base}	✓	✓	.296	.212	.067	.133
Ada-Y	–	✓	.307	.214	.059	.134
Ada-Y+X	✓	✓	.308	.214	.057	.133
RecentE	✓	✓	.307	.202	.070	.122
ECHO	✓	✓	.294	.200	.056	.124

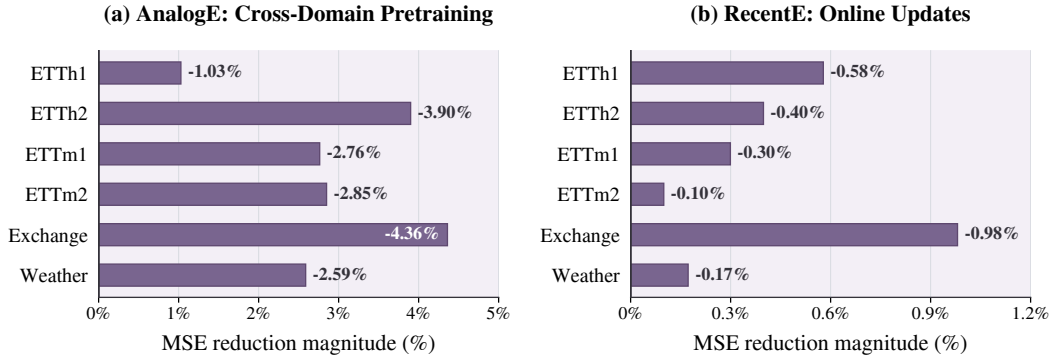


Figure 5: Effects of AnalogE Pretraining and RecentE Online Updates. (a) Relative MSE changes from cross-domain pretraining of AnalogE compared with training only on the target domain. (b) Relative MSE changes from online updates of RecentE compared with keeping the residual map frozen. Bar lengths represent relative reduction magnitudes.

5.6 Efficiency Analysis

We evaluate ECHO's parameter requirements and computational cost. AnalogE uses approximately 42K parameters shared across series, while RecentE maintains a linear residual map for each series. On ETTh1, ECHO takes 31.66 ms per inference query, compared with 26.01 ms for the frozen backbone, corresponding to an additional 5.65 ms. Each online update takes 486.69 ms, with this cost amortized over the predictions between successive updates. These results quantify the additional computational cost of forecast correction.

6 Conclusion

We presented ECHO, a lightweight framework for correcting black-box TSFM forecasts using historical analogues and past forecast errors. These sources provide complementary signals: how similar contexts evolved and how the backbone deviates from observed outcomes in the target domain. ECHO learns transferable analogue-based corrections through cross-domain pretraining and target-specific residual corrections through online updates, dynamically combining them according to observed forecasting performance. Across 5 TSFMs and 9 real-world datasets, ECHO reduces MSE in all 45 backbone–dataset settings, with an average relative reduction of 9.26% in macro-average MSE. Its gains also extend to long-horizon forecasting.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search, writing portions of the code, and language polishing of parts of the manuscript. All AI-assisted content was reviewed and approved by the authors. This included verifying references and their support for the associated claims, checking code correctness, and ensuring that the polished text accurately conveyed the intended meaning. The authors take full responsibility for the final content of this work, including text and code produced with the assistance of generative AI tools.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

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*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Sebastian Pineda Arango, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Senetaire, Caner Turkmen, Oleksandr Shchur, Danielle C Maddix, et al. Chronosx: Adapting pretrained time series models with exogenous variables. *arXiv preprint arXiv:2503.12107*, 2025.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jialin Chen, Ziyu Zhao, Gaukhar Nurbek, Aosong Feng, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. Trace: Grounding time series in context for multimodal embedding and retrieval. *Advances in Neural Information Processing Systems*, 38:2440–2472, 2026.

Mouxiang Chen, Lefei Shen, Han Fu, Zhuo Li, Jianling Sun, and Chenghao Liu. Calibration of time-series forecasting: Detecting and adapting context-driven distribution shift. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 341–352, 2024.

Xilin Dai, Yiding Liu, Hongjie Xia, Yifan Hu, Zewei Dong, Jiang-Ming Yang, and Qiang Xu. Learning the context of errors: Black-box online adaptation of time series foundation models. *arXiv preprint arXiv:2606.14222*, 2026.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Abhimanyu Das, Matthew Faw, Rajat Sen, and Yichen Zhou. In-context fine-tuning for time-series foundation models. *arXiv preprint arXiv:2410.24087*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H Nguyen, Wesley M Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *Advances in Neural Information Processing Systems*, 37:74147–74181, 2024.

Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, 270(2):654–669, 2018.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in neural information processing systems*, 36:19622–19635, 2023.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2505.04163*, 2025.

Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International conference on learning representations*, volume 2024, pp. 23857–23880, 2024.

Baoyu Jing, Si Zhang, Yada Zhu, Bin Peng, Kaiyu Guan, Andrew Margenot, and Hanghang Tong. Retrieval based time series forecasting. *arXiv preprint arXiv:2209.13525*, 2022.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

Ying-yee Ava Lau, Zhiwen Shao, and Dit-Yan Yeung. Fast and slow streams for online time series forecasting without information leakage. In *International Conference on Learning Representations*, volume 2025, pp. 28736–28762, 2025.

Thomas L Lee, William Toner, Rajkarn Singh, Artjom Joosen, and Martin Asenov. Lightweight online adaption for time series foundation model forecasts. *arXiv preprint arXiv:2502.12920*, 2025.

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. *arXiv preprint arXiv:2601.20280*, 2026.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. *Advances in Neural Information Processing Systems*, 37:2766–2786, 2024a.

Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference 2024*, pp. 4095–4106, 2024b.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024c.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*, 2024d.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025b.

Yong Liu, Xingjian Su, Shiyu Wang, Haoran Zhang, Haixuan Liu, Yuxuan Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Timer-s1: A billion-scale time series foundation model with serial scaling. *arXiv preprint arXiv:2603.04791*, 2026.

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, Puneet Agarwal, et al. Long short term memory networks for anomaly detection in time series. In *Esann*, volume 2015, 2015.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James 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 forecaster. *Advances in Neural Information Processing Systems*, 38:163170–163199, 2026.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 914–921, 2020.

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

Eleni I Vlahogianni, Matthew G Karlaftis, and John C Golias. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43: 3–19, 2014.

Peter R Winters. Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324–342, 1960.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Silin Yang, Dong Wang, Haoqi Zheng, and Ruochun Jin. Timerag: Boosting llm time series forecasting via retrieval-augmented generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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

Boya Zhang, Shuaijie Yin, Huiwen Zhu, and Xing He. Freqcycle: A multi-scale time-frequency analysis method for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28159–28167, 2026.

Huanyu Zhang, Chang Xu, Yi-Fan Zhang, Zhang Zhang, Liang Wang, and Jiang Bian. Timeraf: Retrieval-augmented foundation model for zero-shot time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5654–5665, 2025.

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 V. 2*, pp. 6875–6886, 2026.

Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.

APPENDIX

CONTENTS

A Methodological Details **15**

A.1 Historical Case Retrieval 15

A.2 Explicit Evidence Statistics 15

A.3 Causal Online Residual Updates 17

A.4 View-Specific Record Selection and Weight Estimation 17

B Experimental Setup and Implementation Details **19**

B.1 Datasets and Data Usage 19

B.2 ECHO Implementation and Training 20

B.3 Long-Horizon Forecasting Protocol 20

B.4 Baseline Configurations 21

B.5 Experimental Environment and Runtime Measurement 21

C Additional Experimental Results **21**

C.1 Cross-Backbone Evaluation 21

C.2 Long-Horizon Forecasting 22

C.3 Complementarity of Historical and Recent Evidence 23

C.4 Construction of Historical Analog Evidence 23

C.5 Construction of Adaptive Recent Evidence 24

C.6 Dynamic Aggregation versus Fixed Fusion 24

C.7 Role of Temporal Views in TempRA 24

C.8 Robustness to Random Seeds 25

A METHODOLOGICAL DETAILS

This section provides additional methodological details of ECHO, covering historical case retrieval, explicit evidence statistics, causal residual updates, and temporal reliability aggregation.

A.1 HISTORICAL CASE RETRIEVAL

Given the current context X_t, we retrieve context–future pairs (X_i, Y_i) from the historical memory of the same variate.

We use a frozen encoder f_{ret} to map the query and historical contexts into a shared embedding space, and measure their similarity using squared Euclidean distance:

$$d_i = \|f_{\mathrm{ret}}(X_t) - f_{\mathrm{ret}}(X_i)\|_2^2 . \tag{10}$$

We select the K nearest valid candidates as historical evidence for AnalogE. Historical futures Y_i are excluded from retrieval keys and are used only after retrieval to construct the correction reference. Invalid candidates are masked out; if no valid candidates are available, AnalogE returns the base forecast.

A.2 EXPLICIT EVIDENCE STATISTICS

The evidence gate uses a 19-dimensional statistics vector q_t to characterize historical support quality, correction evidence ambiguity, and foundation–experience disagreement. Table 4 summarizes these groups. All statistics use only the current context, retrieved historical cases, and base forecast, without access to current target values.

Table 4: Explicit evidence statistics used by AnalogE. The three groups form the 19-dimensional vector q_t.

Category	**Dim.**	**Statistics**
Historical support quality	9	Log-transformed minimum, weighted mean, weighted standard deviation, and nearest-neighbor margin of retrieval distances; weighted context RMSE, correlation, and last-value mismatch; bounded context-volatility mismatch and signed log context-volatility ratio.
Correction evidence ambiguity	6	Normalized effective support; normalized top-4 template discrepancy; absolute and relative transition dispersion; directional agreement; and mean pairwise transition cosine.
Foundation–experience disagreement	4	Mean absolute difference, symmetric relative RMS difference, correlation, and maximum absolute difference between the historical transition template and the base forecast transition.

Below, we clarify statistics whose computation is not fully specified by their names. We omit the forecast-origin index t for readability. The index h denotes a forecast step, and subscripts h indicate the corresponding vector components. We use $\epsilon = 10^{-6}$ for numerical stability, $\mathrm{clip}_{[a,b]}$ to denote clipping to $[a, b]$, and $\mathbf{1}[\cdot]$ to denote the indicator function.

Distance Weights and Reference Template. Let $\mathcal{C}$ denote the valid candidate set, d_i the retrieval distance of candidate i, and Δ_i its historical transition vector. The fixed distance weights and diagnostic template are

$$\beta_i = \frac{\exp(-d_i)}{\sum_{j\in\mathcal{C}} \exp(-d_j)}, \qquad T^{\mathrm{dist}} = \sum_{i\in\mathcal{C}} \beta_i \Delta_i. \tag{11}$$

These weights use temperature 1 and are distinct from the learned scoring-head weights $\alpha_{t,i}$. The explicit statistics are computed using the diagnostic template T^{dist}, rather than the learned template T_t.

The four distance statistics are the minimum distance, the β_i-weighted mean and standard deviation, and the nearest-neighbor margin. The margin is the difference between the second-smallest and smallest valid distances. Each of these statistics is transformed by $\log(1 + \cdot)$. Context RMSE, correlation, and absolute last-value mismatch are computed for each query–candidate pair and then averaged using β_i.

Effective Support and Template Discrepancy. The normalized effective support is

$$s_{\text{eff}} = \text{clip}_{[0,1]} \left(\frac{(\sum_{i \in \mathcal{C}} \beta_i^2)^{-1} - 1}{K - 1} \right). \tag{12}$$

The normalization uses the configured number of candidate slots K, even when fewer candidates are valid.

For a transition vector $u \in \mathbb{R}^H$, define its stabilized root mean square (RMS) as

$$R(u) = \sqrt{\frac{1}{H} \sum_{h=1}^{H} u_h^2 + \epsilon}, \tag{13}$$

where H is the forecast horizon. Let $T^{(4)}$ be the template formed from the four nearest candidate slots, with distance weights renormalized over their valid candidates. The template discrepancy is

$$s_{\text{top4}} = \frac{R(T^{(4)} - T^{\text{dist}})}{\max\big(R(T^{(4)}) + R(T^{\text{dist}}), \epsilon\big)}. \tag{14}$$

Smaller values indicate greater consistency between the top-4 template and the full retrieved template.

Transition Dispersion and Directional Agreement. Absolute and relative transition dispersion are

$$s_{\text{abs}} = \sqrt{\frac{1}{H} \sum_{h=1}^{H} \sum_{i \in \mathcal{C}} \beta_i (\Delta_{i,h} - T_h^{\text{dist}})^2 + \epsilon}, \qquad s_{\text{rel}} = \frac{s_{\text{abs}}}{\max\big(\sum_{i \in \mathcal{C}} \beta_i R(\Delta_i), \epsilon\big)}. \tag{15}$$

Directional agreement measures the majority-sign fraction at each forecast step, averaged over the horizon:

$$s_{\text{dir}} = \frac{1}{H} \sum_{h=1}^{H} \max \left(\frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \mathbf{1}[\Delta_{i,h} > 0], \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \mathbf{1}[\Delta_{i,h} < 0] \right). \tag{16}$$

Here, $|\mathcal{C}|$ denotes the number of valid candidates. Zero transitions contribute to neither sign. Directional agreement and mean pairwise transition cosine are unweighted over valid candidates. The latter averages cosine similarity over distinct valid candidate pairs and is set to zero when no such pair exists.

Context-Volatility Statistics. Volatility is the population standard deviation of first differences. Let σ_X denote the query-context volatility, σ_i the volatility of historical context X_i, and $\bar{\sigma} = \sum_{i \in \mathcal{C}} \beta_i \sigma_i$ the weighted historical-context volatility. The bounded mismatch and signed log ratio are

$$s_{\text{vol}} = \frac{|\bar{\sigma} - \sigma_X|}{\max(\bar{\sigma} + \sigma_X, \epsilon)}, \qquad s_{\text{logvol}} = \text{clip}_{[-8,8]} \left(\log \frac{\bar{\sigma} + \epsilon}{\sigma_X + \epsilon} \right). \tag{17}$$

The first statistic measures unsigned mismatch, while the second preserves its direction. Both compare historical and query contexts, rather than future forecasts.

Relative Template–Backbone Disagreement. Let p^{base} denote the base forecast transition defined in the main text, with the forecast-origin index omitted. The symmetric relative RMS difference is

$$s_{\text{dis}} = \frac{R(T^{\text{dist}} - p^{\text{base}})}{\max(R(T^{\text{dist}}) + R(p^{\text{base}}), \epsilon)}. \tag{18}$$

The remaining disagreement statistics are the mean absolute difference, centered correlation, and maximum absolute difference between these two transition vectors.

A.3 Causal Online Residual Updates

RecentE and TempRA share the causal stream of matured feedback $\mathcal{M}_t$, but apply different record selection rules. For RecentE, feedback records are selected at intervals of s forecast origins. A selected record from origin j becomes eligible for fitting only when its complete target horizon has been observed, i.e., $j + H \leq t$. The residual map is refitted after every u newly selected, fully matured feedback records. Fitting uses the initial training records and all selected mature feedback available by time t, whose union is denoted by $\mathcal{D}_t$. Selected records are retained for subsequent updates.

Robust sample weights are computed from the fitting errors of the current residual map. For each record $i \in \mathcal{D}_t$, let Y_i denote the observed future, Y_i^{base} the frozen backbone's forecast, and $\hat{r}_i$ the residual estimate produced by the current map. The fitting error is

$$e_i = (Y_i - Y_i^{\text{base}}) - \hat{r}_i. \tag{19}$$

We partition the forecast horizon into $G = \lceil H/b \rceil$ contiguous groups, each containing at most b time steps. Let $\mathcal{I}_g$ denote the time-step indices in group g. We compute the root mean squared error within each group and average the resulting Huber weights to obtain a sample weight:

$$q_{i,g} = \sqrt{\frac{1}{|\mathcal{I}_g|} \sum_{h \in \mathcal{I}_g} e_{i,h}^2}, \qquad a_i = \frac{1}{G} \sum_{g=1}^{G} \min\left(1, \frac{\delta}{q_{i,g} + \epsilon}\right). \tag{20}$$

Here, $e_{i,h}$ is the fitting error at forecast step h, $q_{i,g}$ is the root mean squared error in group g, $\delta > 0$ is the Huber threshold, and $\epsilon > 0$ is a small constant for numerical stability. Grouping is used only to compute the scalar sample weight a_i; separate residual maps are not fitted for individual groups.

With the sample weights fixed, the residual map is updated through weighted ridge regression:

$$W_t = \arg\min_{W} \sum_{i \in \mathcal{D}_t} a_i \left\| \phi_i W - z_i \right\|_2^2 + \lambda \|W\|_F^2. \tag{21}$$

Here, ϕ_i contains the spectral input features derived from the context and base forecast, and z_i is the target spectrum of the residual $Y_i - Y_i^{\text{base}}$. The complex linear map W is optimized to obtain the updated map W_t. The norms $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the Euclidean and Frobenius norms, respectively, and $\lambda > 0$ is the ridge regularization coefficient. Each weight a_i scales the entire spectral fitting loss of its corresponding sample, limiting the influence of large fitting errors.

Each refitting procedure starts with an ordinary ridge regression solution, followed by alternating sample-weight estimation and weighted refitting until the stopping criterion is met or the maximum number of iterations is reached.

A.4 View-Specific Record Selection and Weight Estimation

TempRA constructs long, short, and recent views from the causal stream of matured feedback $\mathcal{M}_t$ to estimate the reliability of the two evidence branches. We index these views by $v \in \{L, S, R\}$, respectively. Each record contains the AnalogE forecast Y_i^A and RecentE forecast Y_i^R issued at the original forecast origin, together with the subsequently observed outcome Y_i. Reliability is evaluated using these issued forecasts; historical predictions are not recomputed with updated modules.

Feedback Record Selection. Let $\mathcal{S}_t \subseteq \mathcal{M}_t$ denote the sparsely selected mature records, with consecutive selected forecast origins separated by at least H time steps. The record sets for the three views are

$$\mathcal{V}_t^L = \mathcal{S}_t, \qquad \mathcal{V}_t^S = \text{Last}_{n_S}(\mathcal{S}_t), \qquad \mathcal{V}_t^R = \text{Last}_{n_R}(\mathcal{M}_t), \tag{22}$$

where Last_m returns the latest available records, up to m, ordered by forecast origin. The parameters n_S and n_R specify the maximum numbers of records in the short and recent views, respectively.

The long view uses all sparsely selected records to capture cumulative performance. The short view uses the latest n_S records from this set. The recent view directly uses the latest n_R records from the dense feedback stream without the sparse selection interval, capturing the most recent performance differences.

Long- and Short-View Weights. For branch $k \in \{A, R\}$, the forecast MSE on record i is

$$\ell_i^k = \frac{1}{H} \left\| Y_i^k - Y_i \right\|_2^2. \tag{23}$$

We normalize this loss by the error of a persistence forecast:

$$\ell_i^{\text{pers}} = \frac{1}{H} \left\| x_i^{\text{last}} \mathbf{1}_H - Y_i \right\|_2^2, \qquad \bar{\ell}_i^k = \frac{\ell_i^k}{\ell_i^{\text{pers}} + \epsilon}, \tag{24}$$

where the persistence forecast uses the last context observation x_i^{last} at every forecast step, $\mathbf{1}_H$ is a length-H vector of ones, and $\epsilon > 0$ ensures numerical stability.

For view $v \in \{L, S\}$, we accumulate the normalized losses over its record set and obtain the AnalogE weight through exponential weighting:

$$L_{t,v}^k = \sum_{i \in \mathcal{V}_t^v} \bar{\ell}_i^k, \qquad w_t^v = \frac{\exp(-\eta L_{t,v}^A)}{\exp(-\eta L_{t,v}^A) + \exp(-\eta L_{t,v}^R)}. \tag{25}$$

Here, $\eta > 0$ controls the sensitivity of the weights to differences in accumulated loss. The corresponding RecentE weight is $1 - w_t^v$. When the record set is empty, the two branches receive equal weights.

Recent-View Weight. The recent view accounts for both the magnitude and joint behavior of the two branches' errors. The historical forecast errors are

$$e_{i,h}^A = Y_{i,h}^A - Y_{i,h}, \qquad e_{i,h}^R = Y_{i,h}^R - Y_{i,h}, \tag{26}$$

where h denotes a forecast step. We estimate the error second moments and cross-moment using the recent feedback records:

$$\begin{aligned} m_t^{AA} &= \frac{1}{N_t} \sum_{i \in \mathcal{V}_t^R} \sum_{h=1}^{H} (e_{i,h}^A)^2, \\ m_t^{RR} &= \frac{1}{N_t} \sum_{i \in \mathcal{V}_t^R} \sum_{h=1}^{H} (e_{i,h}^R)^2, \\ m_t^{AR} &= \frac{1}{N_t} \sum_{i \in \mathcal{V}_t^R} \sum_{h=1}^{H} e_{i,h}^A e_{i,h}^R, \end{aligned} \tag{27}$$

where $N_t = |\mathcal{V}_t^R| H$ is the number of error terms used for estimation. These statistics are computed without mean centering and therefore represent second moments and a cross-moment.

For an AnalogE weight w, the empirical combination error is

$$J_t(w) = w^2 m_t^{AA} + (1-w)^2 m_t^{RR} + 2w(1-w) m_t^{AR}. \tag{28}$$

The recent-view weight assigned to AnalogE minimizes this error:

$$w_t^R = \arg \min_{w \in [0,1]} J_t(w). \tag{29}$$

When $m_t^{AA} + m_t^{RR} - 2m_t^{AR} > 0$, the solution is

$$w_t^R = \text{clip}_{[0,1]} \left(\frac{m_t^{RR} - m_t^{AR}}{m_t^{AA} + m_t^{RR} - 2m_t^{AR}} \right), \tag{30}$$

where $\text{clip}_{[0,1]}$ restricts the weight to $[0, 1]$. Here, the superscript R in w_t^R denotes the recent view, whereas R in the forecast errors and moments denotes the RecentE branch. The corresponding RecentE weight is $1 - w_t^R$.

When the recent view contains fewer than $n_{\min}$ mature records, where $n_{\min}$ is the minimum required record count, we skip moment estimation and assign equal weights to the two branches. If the denominator is zero, the combination error is independent of the weight, and equal weighting is also an optimal solution.

Multi-View Aggregation. The final AnalogE weight combines the three view-specific weights:

$$w_t = \lambda_L w_t^L + \lambda_S w_t^S + \lambda_R w_t^R, \tag{31}$$

where the view-combination coefficients satisfy

$$\lambda_v \geq 0, \qquad \sum_{v \in \{L,S,R\}} \lambda_v = 1. \tag{32}$$

The corresponding RecentE weight is $1 - w_t$, and the final prediction is

$$\hat{Y}_t = w_t Y_t^A + (1 - w_t) Y_t^R. \tag{33}$$

B Experimental Setup and Implementation Details

B.1 Datasets and Data Usage

We evaluate ECHO on time series datasets covering electricity transformer measurements, exchange rates, electricity consumption, solar power generation, weather, and traffic flow. We additionally evaluate on PEMS04 in the supplementary cross-backbone experiments.

ETT. ETTh1, ETTh2, ETTm1, and ETTm2 (Zhou et al., 2021) contain oil temperature and load measurements from two electricity transformers. ETTh1 and ETTh2 are sampled hourly, while ETTm1 and ETTm2 are sampled at 15-minute intervals.

Exchange Rate. Exchange Rate (Lai et al., 2018) contains daily exchange rates of eight countries, providing a financial forecasting setting with irregular fluctuations and changing temporal patterns.

Electricity and Solar-Energy. Electricity (Lai et al., 2018) records hourly electricity consumption from 321 clients. Solar-Energy (Lai et al., 2018) records solar power generation from 137 photovoltaic plants at 10-minute intervals. These datasets represent energy demand and renewable generation, respectively.

Weather. Weather (Wu et al., 2021) contains 21 meteorological variables recorded at 10-minute intervals, including temperature, humidity, and atmospheric pressure. It covers variables with diverse physical meanings and temporal dynamics.

PEMS. PEMS08 and PEMS04 (Song et al., 2020) contain traffic flow measurements from 170 and 307 sensors, respectively, at 5-minute intervals. They capture recurring traffic patterns and short-term fluctuations in traffic flow.

Data Splits. We split each dataset chronologically into training, validation, and test sets, using a ratio of 6:2:2 for the ETT datasets and 7:1:2 for the remaining datasets.

Data Usage. We pretrain AnalogE using queries sampled from the Chronos pretraining corpus (Ansari et al., 2024) and keep it frozen on all downstream datasets. For each downstream dataset, the training split is used to construct a fixed retrieval memory of context–future pairs and initialize the RecentE residual map. The retrieval memory remains unchanged throughout validation and testing.

The validation split provides causal online warm-up for RecentE and TempRA. The test split inherits their states at the end of validation and is evaluated chronologically. Online updates use only feedback whose complete target horizons have been observed by the current forecast origin. Validation and test feedback is neither added to the retrieval memory nor used to update AnalogE.

Normalization and Evaluation Scale. Each variate is standardized using statistics estimated from the training split only. Query and candidate windows are then independently normalized using their own context statistics, while the base forecast uses the query statistics. Context standard deviations are floored at 10^{-3}, and normalized values are clipped to $[-20, 20]$. Corrected forecasts are mapped back to the dataset-standardized scale before computing MSE and MAE. Macro averages are unweighted means of dataset-level metrics.

B.2 ECHO IMPLEMENTATION AND TRAINING

Default Configuration. Unless otherwise specified, experiments use frozen TimesFM 2.5 with context length $L = 512$, forecast horizon $H = 64$, and $K = 20$ retrieved cases. Forecast correction is performed independently for each variate. AnalogE remains frozen during downstream evaluation, while RecentE and TempRA use only causally available, fully matured feedback.

AnalogE Architecture and Retrieval. AnalogE uses separate candidate-set and relation encoders, each with hidden dimension 32 and one four-head self-attention layer. The candidate-set encoder uses masked mean pooling to produce a set summary, while the relation encoder supplies candidate representations to the scoring head. The learned score adjustment is bounded by $s_{\max} = 0.5$, with distance temperature $\tau = 1$. The gate combines the set summary with evidence statistics through an MLP with hidden dimension 64, and its sigmoid output is scaled by $g_{\max} = 0.5$. In implementation, the 19 explicit statistics are supplemented with auxiliary local features. The 64-step horizon is divided into eight contiguous groups. Candidate weights and correction gates are computed separately for each group, with network parameters shared across groups. Group indices are omitted in the main-text formulation for clarity. AnalogE has approximately 42K parameters, excluding the frozen backbone and retrieval encoder. Retrieval uses the 768-dimensional last-token representation of the frozen `amazon/chronos-t5-base` encoder and selects $K = 20$ cases by squared Euclidean distance. Downstream memories contain only complete training-split cases whose futures are fully observed by the current forecast origin.

AnalogE Pretraining. For each backbone, we pretrain a separate AnalogE checkpoint from scratch on 50,000 cross-domain examples from the Chronos pretraining corpus and reuse it across all downstream datasets without parameter updates. Training uses AdamW with learning rate 10^{-4}, weight decay 10^{-4}, batch size 32, and three epochs, with cosine learning-rate scheduling and gradient-norm clipping at 0.5. The objective combines forecast MSE in the window-normalized space with a gate-supervision MSE weighted by 0.1. For each forecast group, the gate target is the least-squares coefficient for correcting the base forecast toward the historical reference, clipped to $[0, g_{\max}]$. This target is used only during offline training. The best checkpoint is selected by normalized forecast MSE on the upstream validation set.

RecentE. The residual map is initialized on the target-domain training split. The standard record-selection interval is $s = 64$ forecast origins, and refitting is triggered by $u = 64$ newly selected, fully matured feedback records, as defined in Appendix A.3. Robust fitting uses groups of at most eight forecast steps, a Huber threshold of $\delta = 0.5$, and ridge regularization $\lambda = 10^{-3}$. Each refitting procedure uses at most five reweighting iterations with an early-stopping tolerance of 10^{-7}.

TempRA. The short and recent views use up to 8 sparse records and 32 dense records, respectively. The long and short views use exponential weighting with $\eta = 1$. The view-combination coefficients are selected on the validation set and held fixed during testing. Full record-selection and weight-estimation rules are given in Appendix A.4.

B.3 LONG-HORIZON FORECASTING PROTOCOL

Long-horizon experiments use frozen TimesFM 2.5 with context length $L = 512$ and horizons $H \in \{96, 192, 336, 720\}$. We treat the model as a black-box forecaster. Since the interface used in our experiments returns at most 128 steps per call, longer forecasts are generated recursively by appending predicted values to the input and retaining the latest 512 steps as context.

AnalogE follows a separate 64-step rollout without retraining. At each step, it retrieves historical cases using its current rollout context and corrects the resulting forecast. The corrected segment is appended to its own context, again retaining the latest 512 steps. Only the required steps of the final segment are retained. RecentE uses a residual map defined for the complete horizon H, taking the original observed context and the complete base forecast as inputs.

For a forecast issued at origin t, all rollout segments use only observations and matured feedback available at t. Internal rollout positions do not advance the feedback-availability clock. Historical

forecasts become eligible for online updates only after their complete H-step targets have been observed.

We refer to this long-horizon configuration as ECHO-Long. Original and corrected forecasts are evaluated on the same test queries and targets. Complete results are reported in Appendix C.2.

B.4 Baseline Configurations

All baselines follow the same data splits, evaluation queries, normalization, and causal feedback availability as ECHO. Below, we describe the adaptations made to each method.

BB-TS-RAG. Original TS-RAG integrates retrieved information with internal backbone representations before the output projection. To support black-box access, we move this integration to the output side. Retrieved futures are pooled using distance-based weights, and a lightweight gating network uses the current context, base forecast, and retrieval reference to produce a scalar weight for their convex combination. The retrieved cases, pretraining data, and training budget are matched to AnalogE.

Naive RAF. We adapt the originally Chronos-based Naive RAF input augmentation to frozen TimesFM, using the same retrieved cases as AnalogE. We retain its procedure of averaging retrieved segments, normalizing and shifting the resulting segment for alignment, and prepending it to the current context. No trainable adapter is introduced.

ELF and ELF+Y^{base}. ELF retains its original frequency-domain forecaster and slow, fast, and merge weighters. It uses a frequency-retention ratio of 0.9, Woodbury updates with interval $M = 200$, L_2 regularization $\lambda = 20$, a fast-weighting window of $B = 5$, and five block updates for warm-up. The forecaster directly predicts future values, which are dynamically combined with the frozen backbone forecast. Evaluation follows ECHO's data splits, normalization, test queries, and fully matured feedback constraints. ELF+Y^{base} only adds the base forecast to the forecaster inputs; all remaining components are unchanged.

Ada-Y and Ada-Y+X. Ada-Y retains its output adapter, which generates additive corrections from the base forecast, with the optional router disabled. Ada-Y+X changes only the input to the concatenation of the current context and base forecast. Its remaining architecture, initialization, and online update settings are identical to Ada-Y. Both variants retain a gradient update for each fully matured feedback record.

B.5 Experimental Environment and Runtime Measurement

Experiments use one NVIDIA H200 GPU and 16 allocated CPU cores. Accuracy evaluation uses a batch size of 32. Runtime measurements use a batch size of 1, averaging over 100 queries after 10 warm-up queries.

C Additional Experimental Results

This section provides complete forecasting results and additional analyses of ECHO, covering cross-backbone performance, long-horizon forecasting, evidence construction and aggregation, and robustness to random seeds.

C.1 Cross-Backbone Evaluation

We examine whether ECHO improves forecasting across different frozen backbones and target datasets. We evaluate TimesFM 2.5, Chronos-2, Chronos-Bolt, Moirai 2.0, and Sundial, comparing each backbone's original forecasts with their correction by ECHO. Within each comparison, the backbone, data splits, test queries, context length $L = 512$, and forecasting horizon $H = 64$ remain unchanged. ECHO accesses only the backbone's numerical inputs and forecast outputs, and online updates use only fully observed feedback.

Table 5: Complete cross-backbone forecasting results on the nine main datasets and the additional PEMS04 dataset. Ori denotes the frozen backbone, and +ECHO denotes its correction by ECHO. Lower values are better. The better result within each Ori/+ECHO pair is bolded, including ties at the displayed precision.

Dataset	TimesFM 2.5				Chronos-2				Chronos-Bolt				Moirai 2.0				Sundial			
	Ori		+ECHO		Ori		+ECHO		Ori		+ECHO		Ori		+ECHO		Ori		+ECHO	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	.355	**.369**	**.346**	.371	.377	**.366**	**.348**	**.366**	.362	**.365**	**.343**	**.365**	.367	.367	**.340**	**.365**	.337	**.369**	**.336**	.370
ETTh2	.257	.315	**.242**	**.310**	.266	**.301**	**.246**	.302	.252	**.299**	**.240**	.302	.266	.306	**.241**	**.305**	.242	.308	**.233**	**.306**
ETTm1	.280	.316	**.263**	**.314**	.309	.310	**.274**	**.303**	.311	.318	**.271**	**.312**	.322	.321	**.273**	**.314**	.268	**.309**	**.257**	.310
ETTm2	.148	.232	**.136**	**.225**	.149	.219	**.138**	**.218**	.149	.224	**.139**	**.222**	.155	.228	**.138**	**.223**	.149	.234	**.135**	**.224**
Exchange	.059	.168	**.056**	**.164**	.060	.167	**.058**	**.166**	.067	.178	**.064**	**.174**	.060	.169	**.058**	**.167**	.082	.196	**.067**	**.176**
Weather	.135	.171	**.124**	**.164**	.143	.162	**.128**	**.160**	.152	.182	**.134**	**.174**	.148	.172	**.127**	**.164**	.136	.179	**.123**	**.166**
Electricity	.120	.212	**.116**	**.209**	.116	**.200**	**.113**	**.200**	.113	**.200**	**.111**	**.200**	.117	.205	**.114**	**.203**	.107	.202	**.106**	**.201**
PEMS04	.112	.212	**.104**	**.207**	.219	.285	**.167**	**.259**	.255	.317	**.171**	**.267**	.127	.223	**.112**	**.215**	.112	.208	**.105**	**.205**
PEMS08	.097	.196	**.089**	**.191**	.232	.295	**.179**	**.268**	.289	.338	**.190**	**.278**	.112	.212	**.098**	**.203**	.094	.193	**.088**	**.189**
Solar	.208	**.217**	**.180**	**.217**	.147	**.162**	**.145**	.172	.469	.337	**.359**	**.297**	.181	**.197**	**.159**	.202	.210	.225	**.189**	**.224**

Table 6: Complete long-horizon forecasting results of TimesFM 2.5 on nine datasets using black-box forecasting. The context length is $L = 512$, with forecasting horizons $H \in \{96, 192, 336, 720\}$. Ori denotes the original TimesFM 2.5 forecasts, and +ECHO denotes their correction using ECHO-Long. MSE and MAE are reported, with lower values indicating better performance. The better result within each Ori/+ECHO pair is bolded. The last row reports the relative change in the unweighted macro average over the nine datasets. Aggregates and relative changes are computed from unrounded values.

Dataset	$H = 96$				$H = 192$				$H = 336$				$H = 720$			
	Ori		+ECHO		Ori		+ECHO		Ori		+ECHO		Ori		+ECHO	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	.3753	.3826	**.3637**	**.3825**	.4149	.4071	**.3939**	**.4035**	.4344	.4183	**.4103**	**.4140**	.4191	**.4269**	**.4062**	.4313
ETTh2	.2989	.3429	**.2763**	**.3339**	.3681	.3876	**.3408**	**.3779**	.3907	.4101	**.3691**	**.4018**	.4043	.4297	**.3924**	**.4261**
ETTm1	.3011	.3299	**.2842**	**.3298**	.3457	**.3596**	**.3271**	.3602	.3773	**.3826**	**.3662**	.3899	.4405	**.4185**	**.4316**	.4323
ETTm2	.1741	.2519	**.1607**	**.2437**	.2398	.2979	**.2187**	**.2855**	.3018	.3388	**.2780**	**.3264**	.4039	.3989	**.3682**	**.3846**
Exchange	.0914	.2103	**.0854**	**.2048**	.1943	.3144	**.1812**	**.3028**	.3597	.4363	**.3258**	**.4132**	1.0458	.7496	**.9443**	**.7445**
Weather	.1590	.1990	**.1437**	**.1910**	.2094	.2465	**.1889**	**.2379**	.2748	.2915	**.2465**	**.2830**	.3620	.3477	**.3322**	**.3431**
Electricity	.1308	**.2217**	**.1272**	.2220	.1478	**.2379**	**.1462**	.2430	**.1655**	**.2560**	.1657	.2647	**.2053**	**.2911**	.2090	.3049
PEMS08	.1091	.2040	**.0984**	**.1983**	.1432	.2244	**.1256**	**.2157**	.1772	.2465	**.1563**	**.2350**	.2492	.3082	**.1991**	**.2763**
Solar	.2217	**.2255**	**.1932**	.2384	.2552	**.2493**	**.2311**	.2696	.2915	**.2754**	**.2303**	.2865	.3150	**.2977**	**.2352**	.3020
Macro Avg.	.2068	.2631	**.1925**	**.2605**	.2576	.3027	**.2393**	**.2996**	.3081	.3395	**.2831**	**.3350**	.4272	.4076	**.3909**	**.4050**
Avg. Change	–	–	**-6.91%**	**-0.99%**	–	–	**-7.11%**	**-1.05%**	–	–	**-8.10%**	**-1.33%**	–	–	**-8.50%**	**-0.63%**

Table 5 reports complete MSE and MAE results on the nine main datasets and the additional PEMS04 dataset. ECHO reduces MSE across all evaluated backbone–dataset combinations. On PEMS04, the relative MSE reductions range from 5.81% to 32.74% across the five backbones. MAE also improves in many settings, although small increases occur in some cases. These results support the applicability of ECHO across different backbones and domains, with particularly consistent improvements in the primary MSE metric.

C.2 Long-Horizon Forecasting

We examine whether ECHO's forecasting improvements extend to longer prediction horizons. Using frozen TimesFM 2.5 with black-box Rolling-128 forecasting, we compare the original forecasts with those corrected by ECHO-Long at $H \in \{96, 192, 336, 720\}$. The context length remains fixed at $L = 512$. For each horizon, both configurations use the same test queries and target values, while online adaptation uses only feedback whose complete target horizons have been observed.

Table 6 reports the results on nine datasets. ECHO-Long reduces macro-average MSE by 6.91%, 7.11%, 8.10%, and 8.50% at the four horizons, respectively. MSE improves in 34 of the 36 dataset–horizon combinations.

Table 7: Ablation study of ECHO on TimesFM 2.5 across nine datasets with context length 512 and forecasting horizon 64. AnalogE Only and RecentE Only retain Historical Analog Evidence and Adaptive Recent Evidence, respectively. The corresponding radar visualization uses the same MSE results, with ETTh1/ETTh2 and ETTm1/ETTm2 averaged as ETTh Avg. and ETTm Avg. Each entry reports MSE/MAE; lower values are better. Best and second-best results are shown in **bold** and underlined, respectively, according to unrounded values.

	ETTh1		ETTh2		ETTm1		ETTm2		Exchange		Weather		Electricity		PEMS08		Solar	
Method	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Backbone	.355	.369	.257	.315	.280	.316	.148	.232	.059	.168	.135	.171	.120	.212	.097	.196	.208	**.217**
AnalogE Only	.347	**.369**	.246	.313	.268	.318	.143	.231	.056	.165	.131	.172	.119	.212	.103	.209	.181	.217
RecentE Only	.364	.389	.249	.319	.270	.319	**.135**	.225	.070	.188	**.122**	.165	.118	.214	.089	.193	.186	.246
ECHO	**.346**	.371	**.242**	**.310**	**.263**	**.314**	.136	**.225**	**.056**	**.164**	.124	**.164**	**.116**	**.209**	**.089**	**.191**	**.180**	.217

Table 8: Comparison of historical evidence integration methods on a frozen TimesFM 2.5 backbone. Naive RAF uses a frozen backbone. ECHO AnalogE denotes our Historical Analog Evidence module. Each result is reported as MSE/MAE; lower values are better. Best results are bolded according to unrounded values.

Method	**ETTh1**	**ETTh2**	**ETTm1**	**ETTm2**	**Exchange**	**Weather**	**Electricity**	**PEMS08**	**Solar**
TimesFM 2.5	.355 / .369	.257 / .315	.280 / .316	.148 / .232	.059 / .168	.135 / .171	.120 / .212	**.097 / .196**	.208 / .217
BB-TS-RAG	.351 / .369	.247 / .310	.272 / **.315**	.152 / .235	.074 / .183	.139 / .180	.127 / .214	.101 / .202	.196 / .221
Naive RAF	.350 / .372	.248 / **.306**	.280 / .319	.146 / **.230**	.058 / .166	**.130 / .167**	.127 / .221	.137 / .238	.202 / **.208**
ECHO AnalogE	**.347 / .369**	**.246** / .313	**.268** / .318	**.143** / .231	**.056 / .165**	.131 / .172	**.119 / .212**	.103 / .209	**.181** / .217

C.3 Complementarity of Historical and Recent Evidence

We examine whether jointly using historical analogues and recent forecast errors provides more consistent corrections than relying on either evidence source alone. Using frozen TimesFM 2.5, we compare four configurations: the uncorrected backbone, AnalogE-only, RecentE-only, and full ECHO. AnalogE-only corrects the base forecast using historical analogues, whereas RecentE-only applies the learned residual correction. Full ECHO combines the two corrected forecasts through TempRA. The standalone branches retain their respective configurations from full ECHO, and all variants use the same base forecasts and test queries.

Table 7 provides the complete results underlying Figure 2. The stronger standalone branch varies across datasets. AnalogE-only achieves lower MSE than RecentE-only on ETTh1, ETTh2, ETTm1, Exchange, and Solar, while RecentE-only performs better on ETTm2, Weather, Electricity, and PEMS08. Full ECHO achieves the lowest MSE on seven of the nine datasets and improves upon the backbone on all nine. RecentE-only remains better on ETTm2 and Weather. These findings support the complementarity of the two sources: their joint use yields more consistent improvements, although aggregation does not always outperform the stronger individual branch.

C.4 Construction of Historical Analog Evidence

We examine the effectiveness of AnalogE in using retrieved historical cases for forecast correction. We compare standalone AnalogE with BB-TS-RAG and Naive RAF on frozen TimesFM 2.5. BB-TS-RAG is a black-box-compatible adaptation of TS-RAG, whose original formulation integrates retrieved representations with internal TSFM representations. Naive RAF keeps the backbone frozen. RecentE and TempRA are excluded so that the comparison focuses on historical evidence construction.

Table 8 reports the complete results on nine datasets. AnalogE achieves the lowest MSE on seven datasets. Naive RAF performs better on Weather, while the uncorrected backbone remains best on PEMS08. These results support AnalogE's effectiveness in constructing corrections from historical analogues. The PEMS08 results also show that historical evidence alone can be insufficient, consistent with the benefit of incorporating recent forecast-error evidence.

Table 9: MSE/MAE comparison of recent-evidence methods on six datasets with TimesFM 2.5. X and Y^{base} indicate predictor inputs. ELF retains its original forecaster and ensemble weighting. All methods are evaluated without AnalogE or TempRA. Lower is better.

Family	Variant	X	Y^{base}	ETTh1	ETTh2	ETTm1	ETTm2	Exchange	Weather
ELF	–	✓	–	.342/.367	.247/.314	.276/.315	.146/.232	.067/.183	.133/.171
ELF	+Y^{base}	✓	✓	.342/.367	.250/.316	.276/.316	.147/.232	.067/.184	.133/.171
Ada-Y	–	–	✓	.357/.370	.257/.319	.279/.316	.149/.233	.059/.171	.134/.170
Ada-Y	+X	✓	✓	.357/.371	.258/.319	.280/.317	.148/.232	.057/.167	.133/.171
RecentE	–	✓	✓	.364/.389	.249/.319	.270/.319	.135/.225	.070/.188	.122/.165

C.5 Construction of Adaptive Recent Evidence

We compare RecentE with ELF and Ada-Y, including variants with matched predictor inputs. ELF retains its original forecaster and ensemble weighting; ELF+Y^{base} additionally provides the base forecast to its forecaster. Ada-Y+X additionally provides the context to Ada-Y. Each input-augmented variant otherwise retains its baseline settings. RecentE uses both inputs to estimate residual corrections. All methods use frozen TimesFM 2.5 and are evaluated without AnalogE or TempRA.

Table 9 shows that RecentE achieves the lowest MSE on ETTm1, ETTm2, and Weather, including against alternatives with matched predictor inputs. ELF variants achieve the lowest MSE on ETTh1 and ETTh2, while Ada-Y+X performs best on Exchange. These results support RecentE's design beyond input access alone, although its advantage varies across datasets and metrics.

C.6 Dynamic Aggregation versus Fixed Fusion

We examine whether dynamically adjusting the contributions of AnalogE and RecentE improves upon fixed fusion. Using TimesFM 2.5 and Chronos-2, we compare TempRA with a fixed fusion baseline whose weights are selected on the validation split and held constant during testing. The two methods combine exactly the same AnalogE and RecentE forecasts on the same test queries. Thus, the comparison isolates the aggregation rule: fixed fusion retains validation-selected weights, whereas TempRA updates the relative contributions using causally available forecasting feedback.

Table 10 shows that TempRA reduces macro-average MSE by 0.81% and 0.84% for TimesFM 2.5 and Chronos-2, respectively. The corresponding MAE reductions are 1.48% and 2.78%. The largest improvements occur on Exchange. However, fixed fusion achieves lower MSE in some settings, including ETTm2 and Weather for both backbones. These results support a modest average benefit from dynamic aggregation, while showing that fixed weights remain competitive on individual datasets.

C.7 Role of Temporal Views in TempRA

We examine whether combining multiple temporal views improves the reliability estimates used by TempRA. The long, short, and recent views use the same historical feedback stream but differ in temporal coverage and feedback density, with the recent view using dense feedback. We compare single-view configurations, which retain only one view; leave-one-out configurations, which omit one view; and full TempRA, which uses all three. The AnalogE and RecentE forecasts remain fixed, so that the comparison focuses on the temporal information used to determine aggregation weights.

Figure 6 presents the results on ETTm1, Exchange, and Weather. The relative advantages of individual views vary across datasets, and no single view consistently performs best. Full TempRA achieves the best or near-best results across the evaluated datasets. These findings support using multiple temporal views to combine reliability information from different portions of the forecasting history.

Table 10: Comparison of TempRA with validation-selected fixed fusion. Macro reports the average over the six datasets. Relative Change reports the change in macro-average MSE/MAE from fixed fusion to ECHO. Lower values are better; the best results within each backbone are bolded, including ties at the displayed precision. Aggregates and relative changes are computed from unrounded values.

Dataset	**TimesFM 2.5**				**Chronos-2**			
	Fixed Fusion		ECHO		Fixed Fusion		ECHO	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	.352	.380	**.346**	**.371**	.355	.380	**.348**	**.366**
ETTh2	.243	.312	**.242**	**.310**	**.241**	.305	.246	**.302**
ETTm1	.264	.315	**.263**	**.314**	**.266**	.311	.274	**.303**
ETTm2	**.133**	**.223**	.136	.225	**.135**	**.218**	.138	**.218**
Exchange	.064	.179	**.056**	**.164**	.084	.183	**.058**	**.166**
Weather	**.121**	**.163**	.124	.164	**.122**	.161	.128	**.160**
Macro	.196	.262	**.195**	**.258**	.200	.260	**.199**	**.253**
Relative Change	–	–	**-0.81%**	**-1.48%**	–	–	**-0.84%**	**-2.78%**

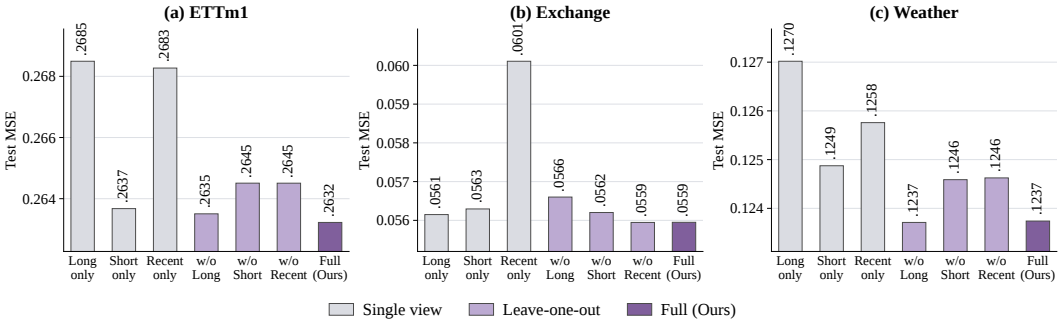


Figure 6: Ablation of temporal views in TempRA. Panels (a)–(c) report test MSE on ETTm1, Exchange, and Weather, respectively, comparing single-view, leave-one-out, and full TempRA configurations. Long, Short, and Recent denote the long, short, and recent views, with Recent using dense feedback. Each panel uses a different vertical axis range; lower MSE is better.

C.8 Robustness to Random Seeds

We examine whether ECHO's improvements are sensitive to the random seed used for AnalogE pretraining. We evaluate three seeds, 42, 2024, and 2025, across five backbones and six datasets: ETTh1, ETTh2, ETTm1, ETTm2, Exchange, and Weather. For each backbone and seed, we independently pretrain AnalogE from scratch on 50,000 cross-domain examples for three epochs and freeze the resulting checkpoint during downstream evaluation. The data splits, fixed training-set retrieval memory, retrieval size $K = 20$, and RecentE and TempRA settings remain unchanged. Each run follows the same chronological validation warm-up and test evaluation procedure.

Tables 11 and 12 report the mean and sample standard deviation over the three seeds for each dataset–backbone combination. MSE standard deviations remain below 1.4% of the corresponding means, except for Sundial on Exchange. MAE standard deviations remain below 1.1% of the corresponding means. ECHO also reduces MSE relative to the frozen backbone in all 90 dataset–backbone–seed configurations. These results indicate low variability and consistent MSE improvements across the evaluated seeds.

Table 11: MSE of ECHO on six datasets across five forecasting backbones. Each entry reports the mean ± sample standard deviation over three seeds (42, 2024, and 2025), with AnalogE independently pretrained for each seed. Lower values are better.

Dataset	TimesFM 2.5	Chronos-2	Chronos-Bolt	Moirai 2.0	Sundial
ETTh1	0.34601 ± 0.00080	0.34853 ± 0.00040	0.34358 ± 0.00092	0.33998 ± 0.00108	0.33314 ± 0.00225
ETTh2	0.24253 ± 0.00085	0.24634 ± 0.00084	0.23981 ± 0.00045	0.24126 ± 0.00053	0.23371 ± 0.00055
ETTm1	0.26401 ± 0.00076	0.27446 ± 0.00074	0.27301 ± 0.00356	0.27541 ± 0.00181	0.25828 ± 0.00152
ETTm2	0.13644 ± 0.00034	0.13789 ± 0.00029	0.13941 ± 0.00034	0.13801 ± 0.00015	0.13519 ± 0.00031
Exchange	0.05627 ± 0.00054	0.05847 ± 0.00012	0.06306 ± 0.00047	0.05785 ± 0.00024	0.06588 ± 0.00126
Weather	0.12380 ± 0.00028	0.12841 ± 0.00035	0.13400 ± 0.00046	0.12779 ± 0.00080	0.12290 ± 0.00015

Table 12: MAE of ECHO on six datasets across five forecasting backbones. Each entry reports the mean ± sample standard deviation over three seeds (42, 2024, and 2025), with AnalogE independently pretrained for each seed. Lower values are better.

Dataset	TimesFM 2.5	Chronos-2	Chronos-Bolt	Moirai 2.0	Sundial
ETTh1	0.37077 ± 0.00052	0.36652 ± 0.00013	0.36462 ± 0.00006	0.36536 ± 0.00028	0.36795 ± 0.00146
ETTh2	0.30948 ± 0.00132	0.30042 ± 0.00115	0.30017 ± 0.00149	0.30414 ± 0.00113	0.30505 ± 0.00065
ETTm1	0.31487 ± 0.00119	0.30361 ± 0.00054	0.31177 ± 0.00046	0.31358 ± 0.00003	0.30899 ± 0.00109
ETTm2	0.22421 ± 0.00049	0.21674 ± 0.00091	0.22051 ± 0.00120	0.22241 ± 0.00024	0.22374 ± 0.00054
Exchange	0.16422 ± 0.00075	0.16562 ± 0.00021	0.17289 ± 0.00067	0.16712 ± 0.00050	0.17544 ± 0.00183
Weather	0.16390 ± 0.00028	0.15913 ± 0.00085	0.17201 ± 0.00145	0.16418 ± 0.00026	0.16605 ± 0.00018