When Does RevIN Hurt? A Statistical Analysis of Reversible Normalization

Anonymous authors
Paper under double-blind review

Abstract

Reversible Instance Normalization (RevIN) has become widely adopted in time series forecasting. However, we show that mean-only normalization can match or outperform full RevIN across a substantial fraction of configurations. To understand this counterintuitive phenomenon, we develop a statistical framework that characterizes the effect of normalization through two competing factors, a structural floor induced by information loss and a transfer distance arising from shifts in the learned mapping between training and deployment. This formulation reveals an inherent tradeoff: normalization can reduce transfer distance by discarding distributional information, but the same information loss introduces a structural floor. Based on this framework, we derive a closed-form characterization of the performance gap between mean normalization and raw data, yielding an exact criterion for mean removal. For variance scaling, whose nonlinear effect precludes an exact characterization, we introduce proxy statistics that capture the corresponding structural floor and transfer distance, and validate their causal roles through controlled experiments on real datasets. Together, these results provide a unified theoretical characterization of RevIN and criteria for determining when its two components are effective.

1 Introduction

Time series forecasting (TSF) is a fundamental problem with broad impact in real-world applications. In recent years, deep learning has driven remarkable progress in TSF, with representative models spanning architectures such as Transformers (Liu et al., 2025; Wang et al., 2025a), MLPs (Lin et al., 2025; Tang & Zhang, 2025), and CNNs (Wang et al., 2026; Li et al., 2025).

Despite the diversity of architectures, their surrounding processing pipelines tend to converge. Nearly all competitive models adopt RevIN, a reversible normalization layer that removes the mean and variance of each input window before forecasting and restores them afterward. This widespread practice rests on an implicit assumption that mean removal and variance scaling are jointly beneficial. However, normalization is inherently a lossy operation, and we cannot simply assume that its benefits will always outweigh the information it discards. In fact, it is reasonable to expect that, in some scenarios, retaining part of the original information may be preferable to the fully bundled scheme, for example by applying mean normalization alone. To test this hypothesis, we conduct a systematic study across a range of architectures and benchmarks (Figure 1). Surprisingly, we find that mean normalization alone can match or even outperform full RevIN across a substantial fraction of the configurations. This observation is consistent with recent studies reporting similar empirical trends (Han et al., 2024b; Fu & Yang, 2025), suggesting that this phenomenon is not an isolated finding.

Yet, despite this growing evidence, we still lack a principled understanding of why mean normalization sometimes suffices and when variance scaling is beneficial. In particular, there is no well-established mechanism for characterizing when additional variance normalization helps rather than hurts, nor a principled criterion for deciding whether it should be applied. To fill this gap, we develop a statistical framework based on deployment risk to characterize the effectiveness of the two components of RevIN, mean removal and variance scaling. Under this framework, the test risk of each normalization variant can be decomposed into two components: a structural floor induced by the normalization operation itself and a transfer distance arising from shifts in the learned mapping between training and deployment. This decomposition reveals normalization as an inherent tradeoff.

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.

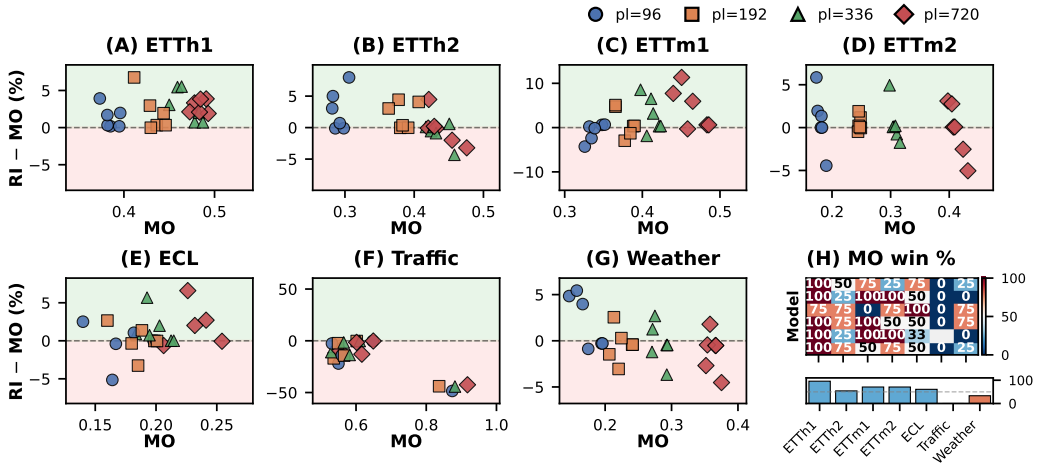

Figure 1: **Comparison between Mean-only (MO) and Full RevIN (RI).** (A)–(G) Relative MSE gap $(\mathrm{RI}-\mathrm{MO})/\mathrm{MO}$ versus MO error across seven datasets. The gaps vary in both directions within most datasets, indicating that RI and MO can each outperform the other depending on the model configuration. ETTh1 and Traffic represent two extremes, with MO consistently outperforming RI on the former and RI consistently outperforming MO on the latter. (H) MO win rate per dataset-model pair (top) and per dataset (bottom). See Table 6 for details.

By discarding certain information, a normalization operation eliminates the corresponding source of transfer distance, but inevitably introduces a structural floor. For mean normalization, removing the input mean forces the model to predict in a mean-centered space, while the future mean is determined by the input mean. Although this discards mean information and introduces a structural floor, it also eliminates the transfer distance that would arise in the raw space from learning a mean mapping that may shift between training and deployment. Variance scaling follows the same principle.

Building on this theory, we further develop a quantitative analysis of the gaps between different normalization schemes. For the gap between the raw and mean-only schemes, the linearity of mean removal allows us to derive a closed-form expression. This gap can therefore be accurately estimated using only a model trained on the raw data. In contrast, the nonlinear nature of variance scaling prevents an exact characterization of the gap between mean-only normalization and full RevIN. We therefore introduce proxy statistics that capture the structural floor and transfer distance, and conduct controlled experiments on real datasets by reparameterizing them to vary these quantities. The results show that the gap between mean-only normalization and full RevIN changes as these quantities are varied, providing evidence for their causal roles. These results not only explain the performance gaps between normalization schemes, but also provide an empirical basis for selecting an appropriate normalization strategy.

Our contributions can be summarized as follows.

1. We conduct a systematic study showing that mean-only normalization can outperform full RevIN across a substantial fraction of architecture-benchmark configurations, challenging the prevailing practice of jointly removing the mean and scaling the variance.
2. We propose a deployment risk framework that decomposes the effect of normalization into a structural floor and a transfer distance. Based on this framework, we derive a closed-form characterization of the performance gap between mean-only normalization and raw data, revealing the conditions under which mean removal is beneficial.
3. We further analyze the role of variance scaling through proxy statistics that capture the structural floor and transfer distance. While an exact characterization is unavailable due to the nonlinear nature of variance scaling, controlled experiments on real datasets demonstrate the causal role of these factors and provide an empirical basis for assessing when variance scaling is likely to be effective.

2 Related work

Normalization in Time Series Forecasting. Reversible Instance Normalization (RevIN) Kim et al. (2022) has become a de facto standard in time series forecasting. However, its widespread adoption has also prompted closer examination of its underlying assumptions. Dish-TS Fan et al. (2023) emphasizes that the statistics observed in the input window may not necessarily characterize the future distribution. SAN Liu et al. (2023) further argues that instance-level normalization can be overly restrictive when statistical properties evolve within an instance, while FAN Ye et al. (2024) and DDN Dai et al. (2024) model evolving non-stationary components beyond conventional temporal statistics. These studies collectively suggest that the statistical components removed by normalization are not necessarily nuisance factors, and their removal is not always beneficial.

Selective and Robust Normalization. Recent work has made this issue more explicit by investigating which statistics should be normalized and under what conditions. SIN Han et al. (2024b) formulates normalization as a selective process and identifies statistics according to their predictive relevance. Fu et al. Fu & Yang (2025) further reveal severe failure modes of variance-based normalization under extreme outliers, while STAIR Yu et al. (2026) shows that standard RevIN can impose an overly strong normalization prior on simple forecasting backbones and proposes a relaxed α-RevIN to improve modeling flexibility. Despite these advances, existing approaches primarily address this issue by designing new normalization mechanisms, selection criteria, or robust estimators, yet lack a fundamental understanding of why normalization succeeds or fails.

Analytical Perspectives on Time Series Forecasters. Beyond novel architecture design, a growing line of work studies the intrinsic properties and behavioral regularities of time series forecasters. Han et al. (2024a) revisits the channel-independent strategy and characterizes its capacity–robustness trade-off; Ahmed et al. (2025) systematically examines the effects of different normalization schemes in time series foundation models. Other studies characterize fundamental limits of forecasting accuracy by relating the minimum achievable error to window-wise pattern complexity Wang et al. (2025b). These works shift the focus from designing new architectures to understanding why forecasters behave. Our work follows this analysis-oriented perspective.

3 Problem setup

We define a lookback window $x \in \mathbb{R}^L$ and a forecast horizon $y \in \mathbb{R}^H$. Let $\bar{x} := \frac{1}{L}\mathbf{1}_L^\top x$ be the window mean, $\tilde{x} := x - \bar{x}\mathbf{1}_L$ the centered window, and $\hat{\sigma} := \|\tilde{x}\|_2/\sqrt{L}$ the window scale. For any forecast $\hat{y} \in \mathbb{R}^H$, the per-window squared error splits into two orthogonal components:

$$\frac{1}{H}\|\hat{y} - y\|^2 = \underbrace{(\hat{\bar{y}} - \bar{y})^2}_{\text{level channel}} + \underbrace{\tfrac{1}{H}\big\|(\hat{y} - \hat{\bar{y}}\mathbf{1}_H) - (y - \bar{y}\mathbf{1}_H)\big\|^2}_{\text{shape channel}} \tag{1}$$

We attribute the performance of any normalization scheme to its level and shape channels respectively. To systematically study normalization, we define two binary switches: $A \in \{0, 1\}$ for mean subtraction and $B \in \{0, 1\}$ for variance scaling. We evaluate four schemes based on (A, B): **Raw** (0,0) feeds x directly to a predictor g; **Mean** (1,0) removes the mean ($\tilde{x}$) and adds $\bar{x}$ to the output; **Var** (0,1) divides by $\hat{\sigma}$; and **Full** (1,1, standard RevIN) applies both operations. Let $\mathcal{L}^{(a,b)}$ denote the test MSE under switch configuration (a, b). We quantify the effect of each operation via pairwise gaps, defined inline as $g_{A|B=b} := \mathcal{L}^{(0,b)} - \mathcal{L}^{(1,b)}$ and $g_{B|A=a} := \mathcal{L}^{(a,0)} - \mathcal{L}^{(a,1)}$. For instance, $g_{A|B=0}$ measures the exact benefit of mean subtraction without variance scaling. All mathematical proofs and additional experimental details are provided in the Appendix.

4 A two-distribution risk decomposition

Theorem 4.1 (Two-distribution master identity)**.** *Assume that* $\mathbb{E}_{\mathrm{te}}\big[\sigma_T^2\langle m_{\mathrm{te}} - \bar{g}^*_{\mathrm{te}}, \bar{g}^*_{\mathrm{te}} - g\rangle\big] = 0$ *for all* $g \in \mathcal{G}$*. Then, for any predictor* $\hat{g}$ *trained on* P_{tr}*, its test risk decomposes as:*

$$\mathcal{L}_{\mathrm{te}}(\hat{g}) \;=\; \underbrace{R^*_{\mathrm{te}}}_{\textit{Bayes noise}} \;+\; \underbrace{\Delta_T(\mathrm{te})}_{\textit{structural floor}} \;+\; \underbrace{\Theta_T(\mathrm{tr} \to \mathrm{te})}_{\textit{transfer distance}} \tag{2}$$

where $R^*_{\mathrm{te}} = \frac{1}{H}\mathbb{E}_{\mathrm{te}}[\sigma_T^2\|z - m_{\mathrm{te}}\|^2]$ *is the irreducible noise floor.*

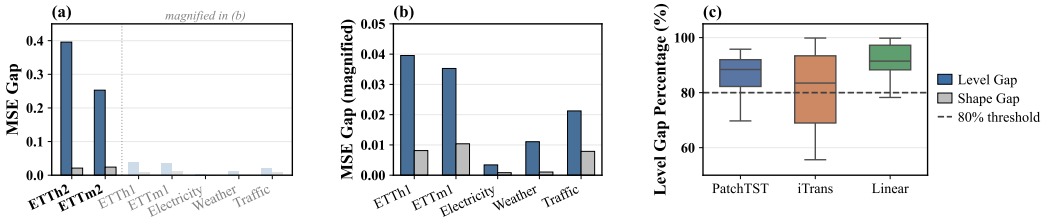

Figure 2: **Decomposition of Prediction Error Gap ($g_{A|B=0}$).** The results demonstrate that the performance gap is predominantly driven by the *Level* component rather than the *Shape* component across various datasets and models. **Left:** Comparison of mean absolute gaps. **Right:** Distribution of the Level Gap percentage across different models.

We evaluate the deployment risk of a scheme T under a test distribution Q. Scheme T defines an envelope $\nu_T = (\mu_T, \sigma_T)$, yielding a normalized input $\varphi_T(x) = (x - \mu_T)/\sigma_T$ and standardized target $z = (y - \mu_T)/\sigma_T$. The predictor g operates on φ_T, and the original-space MSE corresponds to a variance-weighted risk in the normalized space: $\mathcal{L}_Q(g) := \frac{1}{H}\mathbb{E}_Q\|\hat{y} - y\|^2 = \frac{1}{H}\mathbb{E}_Q[\sigma_T^2\|z - g\|^2]$.

To isolate the sources of error, we define three key components for a hypothesis class $\mathcal{G}$: (1) the **within-class oracle** $\bar{g}_Q^* := \arg\min_{g\in\mathcal{G}} \mathbb{E}_Q[\sigma_T^2\|m_Q - g\|^2]$, where $m_Q := \mathbb{E}_Q[z \mid \nu_T, \varphi_T]$ is the Bayes-optimal normalized predictor; (2) the **structural floor** $\Delta_T(Q) := \frac{1}{H}\mathbb{E}_Q[\sigma_T^2\|m_Q - \bar{g}_Q^*\|^2] \geq 0$, the irreducible error due to the choice of normalization and hypothesis class; and (3) the **transfer distance** $\Theta_T(\mathrm{tr} \to \mathrm{te}) := \frac{1}{H}\mathbb{E}_{\mathrm{te}}[\sigma_T^2\|\bar{g}_{\mathrm{te}}^* - \hat{g}\|^2] \geq 0$, quantifying how far the trained predictor $\hat{g}$ deviates from the test oracle. The deployment risk admits an exact decomposition (Theorem 4.1).

Notably, for general deep model classes, the cross-term is non-zero, and the orthogonality condition above serves merely as an assumption to facilitate analysis, while the level-channel identities presented below (Theorem 5.1 and Theorem 5.2) are unconditional population least-squares identities that do not rely on this orthogonality condition.

5 Comparing Mean Removal and Raw for the Level Channel

The Mean scheme strips level information from the input window. When the level and shape channels are approximately independent, the model without access to level information is restricted to making predictions purely within the shape space, hardcoding its level response to the identity mapping. In contrast, the Raw scheme grants the model complete flexibility by feeding both level and shape information into the predictor, allowing it to freely learn the mapping between them from the training data. This design divergence directly dictates the distinct prices each scheme pays in the deployment-risk decomposition established in the previous section: $\mathcal{L}(\hat{g}) = R^* + \Delta_T + \Theta_T$. By rigidly assuming the level mapping to be the identity, the Mean scheme pays a structural floor Δ_T, because this assumption rarely holds exactly in reality. The Raw scheme, conversely, freely fits the level mapping on the training set, thereby incurring a transfer distance Θ_T. Determining which scheme is ultimately superior is essentially a trade-off between the structural bias and the transfer penalty. The core contribution of this section is to quantify this trade-off. We will demonstrate that by training only a single Raw model, we can precisely infer the cost of the Mean scheme through a set of closed-form solutions.

5.1 Level moments and an exact decomposition

The two schemes differ only in how the level is estimated, pinned versus learned, so the gap between them should concentrate in the level component of Equation (1). Figure 2 confirms this across datasets and models, and the derivation below therefore restricts to the level discrepancy.

Level-channel quantities. For each split $s \in \{\mathrm{tr}, \mathrm{te}\}$, define $\ell_s := \mathbb{E}_s[\bar{x}]$,$u := \hat{\bar{y}}$, $V_s := \mathrm{Var}_s(\bar{x})$, $\beta_s^* := \mathrm{Cov}_s(\bar{y}, \bar{x})/V_s$, and $\gamma_s := \mathbb{E}_s[\bar{y}] - \mathbb{E}_s[\bar{x}]$; equivalently, consider the population level regression $\bar{y} = \alpha_s + \beta_s^*\,\bar{x} + \varepsilon_s$ with $\mathbb{E}_s[\varepsilon_s] = 0$ and $\mathrm{Cov}_s(\bar{x}, \varepsilon_s) = 0$. For simplicity, we consider only first-order correlations. Here β_s^* is the population regression slope of the horizon mean on the window mean,

the optimal level translation slope under split s, and γ_s is the horizon level offset, typically nonzero on trending series.

Theorem 5.1 (Exact level decomposition)**.** *Fix a split s and a predictor with level output u, and write the population least-squares regression of u on $\bar{x}$ as $u = \mathbb{E}_s[u] + c_s\,(\bar{x} - \ell_s) + \eta$, so that $\mathbb{E}_s[\eta] = 0$ and $\mathrm{Cov}_s(\bar{x}, \eta) = 0$. Then*

$$\mathbb{E}_s\big[(u-\bar{y})^2\big] = \underbrace{b_s^2}_{\text{level bias}} + \underbrace{(c_s - \beta_s^*)^2\, V_s}_{\text{slope mismatch}} + \underbrace{\mathbb{E}_s\big[(\eta - \varepsilon_s)^2\big]}_{\text{residual energy}} \tag{3}$$

with $b_s := \mathbb{E}_s[u - \bar{y}]$. The three terms are orthogonal by construction of the two regressions, and every quantity on the right is computable from windows, predictions, and deployment levels.

We call c_s the predictor's effective level slope under split s, and the left side of Equation (3) is the predictor's level cost under split s. The rest of this section specializes the level cost to the two schemes' deployment costs $C_{\mathrm{mean}}(\mathrm{te})$ and $C_{\mathrm{raw}}(\mathrm{te})$. Theorem 5.2 then writes their difference in closed form.

5.2 The deployment level costs of the two schemes

Mean-only pays the structural floor. Mean-only forecasts $\hat{y} = \bar{x}\mathbf{1}_H + g(\tilde{x})$, restricting its level output to the identity map $u = \bar{x}$. In the notation of Theorem 5.1, this means its effective slope is locked at $c = 1$, its residual is identically $\eta = 0$ on every split, and its deployment level bias is $b_{\mathrm{mean}} = \mathbb{E}_{\mathrm{te}}[\bar{x} - \bar{y}] = -\gamma_{\mathrm{te}}$. Substituting these into the decomposition yields a closed-form deployment level cost:

$$C_{\mathrm{mean}}(\mathrm{te}) \;=\; \gamma_{\mathrm{te}}^2 \;+\; \Lambda_\mu(\mathrm{te}), \qquad \Lambda_\mu(\mathrm{te}) \;:=\; V_{\mathrm{te}}\,(1-\beta_{\mathrm{te}}^*)^2 \;\geq\; 0 \tag{4}$$

This floor decomposes into two components: the offset floor γ_{te}^2, incurred by forcing a zero intercept when the data exhibits a global shift γ_{te}; and the slope floor $\Lambda_\mu(\mathrm{te})$, the price of forcing a unit slope when the optimal slope is β_{te}^*.

Raw pays the transfer distance. Raw's level output is learned freely on P_{tr}, so its effective slope c, residual η, and bias b_{raw} are not closed-form a priori. However, every quantity it contributes to Equation (3) is measurable post-hoc from the trained model's deployment predictions: c and η come from regressing predictions on window means, while b_{raw} additionally uses the deployment levels. Raw's deployment level cost is therefore the directly measurable quantity:

$$C_{\mathrm{raw}}(\mathrm{te}) := b_{\mathrm{raw}}^2 + (c - \beta_{\mathrm{te}}^*)^2 V_{\mathrm{te}} + \mathbb{E}_{\mathrm{te}}\big[\eta^2\big] - 2\mathbb{E}_{\mathrm{te}}\big[\eta\varepsilon_{\mathrm{te}}\big] \tag{5}$$

Level gap of raw over mean-only. With both costs expressed on the same scale, a fixed structural floor for Mean versus a measurable transfer distance for Raw, the comparison reduces to a simple subtraction. Differencing the two costs gives the exact deployment gap:

Theorem 5.2 (Level gap of raw over mean-only)**.** *Let the trained raw predictor have deployment level bias b_{raw}, effective slope c, and regression residual η, as in Theorem 5.1. The deployment level gap of raw over mean-only is the difference of their respective costs:*

$$\mathrm{gap}_{\mathrm{level}} = C_{\mathrm{raw}} - C_{\mathrm{mean}} = \underbrace{b_{\mathrm{raw}}^2 - \gamma_{\mathrm{te}}^2}_{\textit{bias term}} + \underbrace{(c-\beta_{\mathrm{te}}^*)^2 V_{\mathrm{te}} - \Lambda_\mu(\mathrm{te})}_{\textit{slope term}} + \underbrace{\mathbb{E}_{\mathrm{te}}\big[\eta^2\big] - 2\mathbb{E}_{\mathrm{te}}\big[\eta\varepsilon_{\mathrm{te}}\big]}_{\textit{non-affine remainder}} \tag{6}$$

with the common noise floor $\mathbb{E}_{\mathrm{te}}[\varepsilon_{\mathrm{te}}^2]$ cancelling in the difference.

6 Analyzing variance division in the scale channel

This section studies the effect of variance division. We begin by comparing the Mean-only and Full schemes. Full further divides out the scale $\hat{\sigma}$ on top of Mean-only, forcing the model to learn a scale-invariant shape representation to guard against scale drift at test time. Following the same analytical framework, whether Full prevails again hinges on the trade-off between its structural cost and its transfer benefit. However, unlike the level channel, variance division is a nonlinear operation, and thus no exact gap identity can be derived. To address this, we resort to proxy measures and establish causality through synthetic control experiments.

Specifically, the structural cost of Full relative to Mean-only consists of two terms, each priced by a directly estimable proxy. The first term is the amplitude-information loss Λ_σ, priced by the regression slope ρ_s, which measures how well window amplitude predicts future deviation amplitude.

The second term is the scale-estimation noise, priced by the innovation kurtosis $\hat{\kappa}_4$, which controls the relative estimation error of the sample standard deviation $\hat{\sigma}$. Meanwhile, the amplitude drift δ_s and δ_ρ serves as a proxy for the transfer benefit, quantifying the shift in the scale distribution between training and test sets. When both ρ_s and $(\hat{\kappa}_4 - 1)/(4L)$ are not too large, the structural cost is kept sufficiently low, and Full prevails as long as δ_s, δ_ρ indicate the presence of drift.

Moreover, we note that using variance normalization alone is harmful in most cases, because this operation introduces additional hazards. Without centering, the normalized feature $x/\hat{\sigma}$ has a DC-component projection of $\sqrt{L}\,\bar{x}/\hat{\sigma}$. When a window exhibits near-zero fluctuation ($\hat{\sigma} \to 0$) with nonzero mean ($\bar{x} \neq 0$), this ratio diverges, producing an extreme Cauchy-type heavy tail.

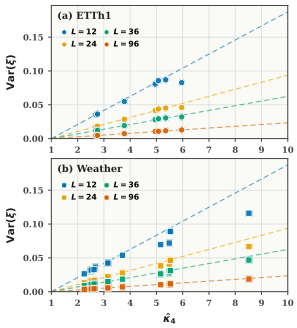

Figure 3: Measured $\mathrm{Var}(\xi)$ against $\hat{\kappa}_4$ on the two datasets for different values of L; dashed lines indicate the theoretical approximation.

Scale-channel quantities. We define several diagnostic quantities: $\hat{\sigma}_i := \|\tilde{x}_i\|_2/\sqrt{L}, a_i := \|y_i - \bar{x}_i \mathbf{1}_H\|_2/\sqrt{H}, r_s := \arg\min_r \sum_{i \in s} (a_i - r\,\hat{\sigma}_i)^2, \hat{\kappa}_4 := \mathbb{E}[\varepsilon^4]/\big(\mathbb{E}[\varepsilon^2]\big)^2, \rho_s := r_{\mathrm{tr}} + r_{\mathrm{te}}, \delta_s := \mathbb{E}_{\mathrm{te}}[\log\hat{\sigma}] - \mathbb{E}_{\mathrm{tr}}[\log\hat{\sigma}], \delta_\rho := r_{\mathrm{te}} - r_{\mathrm{tr}}$. Here $\hat{\sigma}_i$ is the scale of window i and a_i is the amplitude of the future deviation. The coefficient r_s is the least-squares slope of a_i on $\hat{\sigma}_i$ under split s. We use $\rho_s := r_{\mathrm{tr}} + r_{\mathrm{te}}$ as the structural amplitude-coupling proxy, since this cost arises during training. The quantity $\hat{\kappa}_4$ is the standardized fourth moment of the innovations ε, and it controls the relative estimation noise of $\hat{\sigma}$. Finally, δ_s and δ_ρ denote the shifts in log-scale and regression slope between the test and training sets, respectively.

6.1 Population floors in the nonlinear regime

Proposition 6.1 (Floor gap under a scale model)**.** *Consider the scale model $x = \sigma\,\varphi$, where $\sigma > 0$ is the window scale, φ is the normalized shape, and $\sigma \perp \varphi$, and write the future deviation as $y_{\mathrm{dev}} := y - \bar{x}\mathbf{1}_H = \rho(\sigma)\,\sigma\,\varphi_{\mathrm{fut}} + \varepsilon$ with $\mathbb{E}[\varepsilon \mid \sigma, \varphi] = 0$, and assume shape persistence: φ_{fut} is a deterministic function of φ. Assume the scale estimate satisfies $\hat{\sigma} = \sigma(1+\xi)$ with $\mathbb{E}[\xi] = 0$, $\mathrm{Var}(\xi) = \frac{\kappa_4 - 1}{4L} + O(L^{-2})$, and ξ independent of $(\sigma, \varphi, \varepsilon)$. Let $h_0 \in \arg\min_h \mathbb{E}\big\|\rho(\sigma)\,\sigma\,\varphi_{\mathrm{fut}} - \sigma\,h(\varphi)\big\|^2$ be the zero-noise optimal full readout, and define the* amplitude-information loss *as its attained minimum, $\Lambda_\sigma := \mathbb{E}\big\|\rho(\sigma)\,\sigma\,\varphi_{\mathrm{fut}} - \sigma\,h_0(\varphi)\big\|^2 \geq 0$. Then the optimal risks satisfy*

$$\mathcal{L}_*^{\mathrm{full}}(Q) - \mathcal{L}_*^{\mathrm{mean}}(Q) = \Lambda_\sigma + \frac{\kappa_4 - 1}{4L}\,\mathbb{E}\big\|\sigma\,h_0(\varphi)\big\|^2 + O(L^{-2}) \tag{7}$$

Since $\sigma\,h_0(\varphi)$ is the σ^2-weighted projection of the signal $\rho(\sigma)\,\sigma\,\varphi_{\mathrm{fut}}$ onto the full hypothesis class, $\mathbb{E}\|\sigma\,h_0(\varphi)\|^2 \leq \mathbb{E}\|\rho(\sigma)\,\sigma\,\varphi_{\mathrm{fut}}\|^2 \leq \mathbb{E}\|y_{\mathrm{dev}}\|^2$. In particular, if the persistence coefficient is constant, $\rho(\sigma) = \rho$, then $\Lambda_\sigma = 0$ and the estimation-noise term becomes $\frac{\kappa_4 - 1}{4L}\,\mathbb{E}\|\rho\,\sigma\,\varphi_{\mathrm{fut}}\|^2$.

Here we assume that the relative error ξ in $\hat{\sigma} = \sigma(1+\xi)$ is independent of the true scale σ, shape φ, and noise ε, thereby decoupling scale estimation uncertainty from the signal itself. Although such cross-dependence may arise in practice, modeling it explicitly would entangle ξ with the prediction target and substantially complicate the analysis. This independence can be viewed as an analytically convenient approximation that becomes increasingly reasonable as L grows. The empirical results on the effect of $\hat{\kappa}_4$ on $\mathrm{Var}(\xi)$ are presented in Figure 3, showing consistency with the theoretical prediction when $\hat{\kappa}_4$ is not too large.

Amplitude-information loss and scale-estimation noise. The proposition reveals two intrinsic costs of variance normalization. The first is the amplitude-information loss Λ_σ: normalization removes the window scale, preventing the model from conditioning its prediction on scale, and any information about the future carried by the amplitude is consequently discarded. Therefore, we employ $r_{\mathrm{te}} + r_{\mathrm{tr}}$ as a metric, since the structural bias arises from the entire dataset as a consequence of

the operation. Here, r measures the least-squares regression slope of the future deviation amplitude a_i on the window scale $\hat{\sigma}_i$. A positive value indicates information loss. The second cost is scale-estimation noise: multiplying the prediction back by the noisy envelope $\hat{\sigma}$ introduces a relative error of order $(\kappa_4 - 1)/(4L)$.

Quantifying scale drift and amplitude-coupling drift. We characterize the distribution shift between the training and deployment distributions through two distinct proxy statistics: $\delta_s := \mathbb{E}_{\text{te}}[\log \hat{\sigma}] - \mathbb{E}_{\text{tr}}[\log \hat{\sigma}], \delta_\rho := r_{\text{te}} - r_{\text{tr}}$. The first quantity, δ_s, captures the shift in the marginal scale distribution between training and test sets. The second quantity, δ_ρ, captures changes in the structural relationship between the window scale $\hat{\sigma}$ and the future deviation amplitude a. Thus, even when the overall scale level remains stable, i.e., $\delta_s \approx 0$, a nonzero δ_ρ can indicate a change in the scale-to-amplitude relationship. Together, these two quantities provide complementary proxies for distribution shift in the scale dynamics.

6.2 Why variance division requires mean subtraction

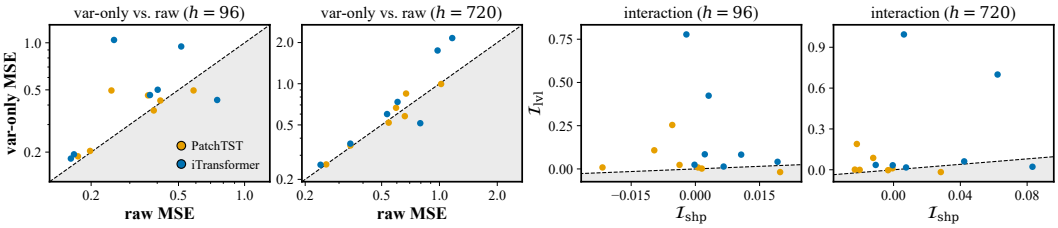

Figure 4: Empirical evaluation of variance scaling and its interaction with mean subtraction based on PatchTST and iTransformer. **Left**: Comparison of test MSE between the Var-only and Raw. **Right**: Channel decomposition of the interaction term $\mathcal{I}$, with the shape component $\mathcal{I}_{\text{shp}}$ on the horizontal axis and the level component $\mathcal{I}_{\text{lvl}}$ on the vertical axis. Scatter points lying above the gray region indicate that Var performs worse and that the level component dominates $\mathcal{I}$, respectively.

Both Full and Mean methods subtract the mean. In principle, this framework remains applicable to Raw and Var as well. However, we note that using variance division alone is harmful in most cases. This section explains the reason: without prior mean removal, division by the standard deviation folds level information into the normalized feature and introduces an additional statistical pathology. Specifically, any window $x \in \mathbb{R}^L$ admits the orthogonal decomposition $x = \bar{x}\mathbf{1}_L + \tilde{x},\ x/\hat{\sigma} = (\bar{x}/\hat{\sigma})\mathbf{1}_L + \tilde{x}/\hat{\sigma}$, showing that variance division acts on both components simultaneously. The mean of the normalized window is simply $\overline{(x/\hat{\sigma})} = (1/L)\mathbf{1}_L^\top (x/\hat{\sigma}) = \bar{x}/\hat{\sigma}$. At this point, the mean and standard deviation become entangled: the standard deviation directly modulates the magnitude of the mean input. Such normalization is ill-posed and confuses the model. Moreover, when $\bar{x} \neq 0$ and $\hat{\sigma} \to 0$, this ratio diverges, producing extremely heavy tails and resulting in a pathological distribution, as illustrated in Figure 5.

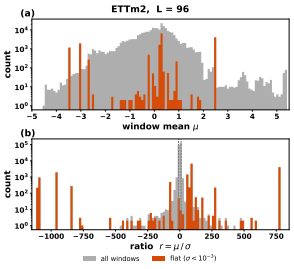

Figure 5: Histograms of $\bar{x}$ and $\bar{x}/\hat{\sigma}$ on ETTm2.

We explicitly mark the location of flat windows ($\hat{\sigma} < 10^{-3}$). While the raw mean $\bar{x}$ ranges approximately from -5 to 5, the range of $\bar{x}/\hat{\sigma}$ spans roughly $[-1000, 750]$, causing the overall distributional mean to explode.

To verify the proposed coupling effect, we define the interaction term $\mathcal{I}$ using the marginal gaps g. The interaction can be computed equivalently in two forms: $\mathcal{I} = g_{A|B=1} - g_{A|B=0} = g_{B|A=1} - g_{B|A=0}$. The left-hand side measures how the benefit of mean subtraction changes when variance scaling is turned on; the right-hand side measures how the benefit of variance scaling changes when mean subtraction is turned on. A non-zero $\mathcal{I}$ indicates that the two operations are not additive: the effect of one depends on whether the other is applied. We decompose this total interaction into level and shape components, $\mathcal{I} = \mathcal{I}_{\text{lvl}} + \mathcal{I}_{\text{shp}}$. Since our preceding analysis suggests that the coupling resides in the level channel, theory predicts that $\mathcal{I}_{\text{lvl}}$ should dominate. As shown in Figure 4, the majority

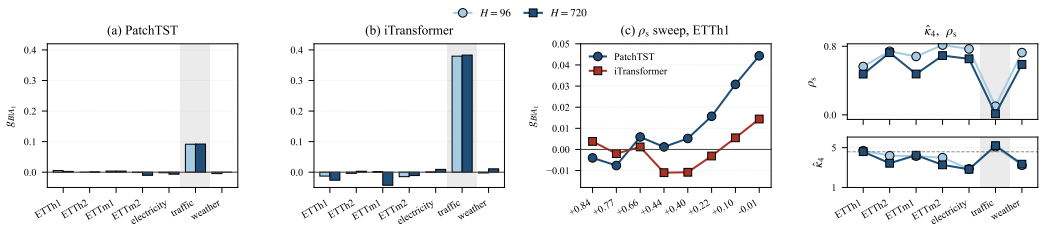

Figure 6: **Empirical Validation of the Variance Division Gate.** The deployment gap $g_{B|A=1}$ and the amplitude-information proxy ρ_s for PatchTST and iTransformer are presented across various datasets and prediction horizons. Additionally, the variation curve of $g_{B|A=1}$ with respect to ρ_s is illustrated ($L = 720$).

of points lie above the diagonal, confirming our conjecture: variance scaling alone introduces a level-channel penalty that is neutralized only when mean subtraction is applied first.

7 Experiment

This section presents the experimental results. Due to space constraints, we focus on two representative models, iTransformer and PatchTST, using an input window size of $L = 96$ and forecast horizons $H \in 96, 720$. The evaluation consists of two parts, corresponding to the performance gaps between Raw and Mean, and between Mean and Full. The Raw-versus-Mean gap admits a closed-form characterization, whereas only proxy quantities are available for the Mean-versus-Full gap. We therefore rely primarily on synthetic control experiments for the latter, using reparameterization on ETTh1 to systematically control the relevant variables. Because our theoretical quantities are defined on the deployment distribution, computing them requires access to the test set and its associated statistics; this use is for evaluation only and does not involve model fitting or hyperparameter selection. We omit the affine layer in RevIN, as it has been shown to be ineffective Berthelier et al. (2026) and would introduce additional learnable parameters that could confound the comparison.

The appendix provides detailed information on the setup and datasets (Appendix A) and the synthetic control methodology (Appendix B). It also extends the evaluation to TimesNet Wu et al. (2023) and TSMixer Chen et al. (2023) across a wider set of datasets (Appendix F).

7.1 Empirical Validation of the Exact Level Gap

We empirically validate the exact level-gap identity derived in Theorem 5.2. The deployment risk of the Mean-only scheme can be accurately predicted using only statistics extracted from a single trained Raw model. We quantify the discrepancy between the theoretically estimated gap and the empirically measured gap for the configuration $g_{A|B=0}$. As shown in Figure 7, the estimated values agree closely with the actual measurements. This high predictive accuracy extends beyond the

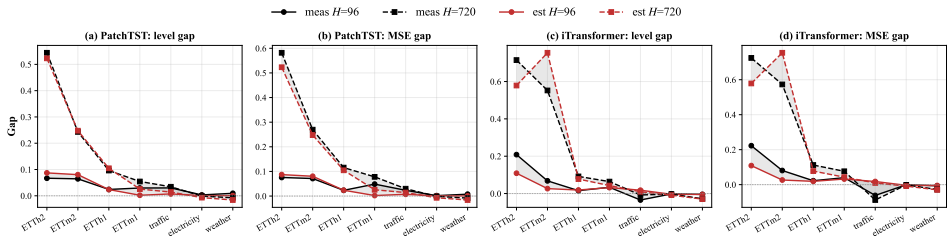

Figure 7: **Empirical validation of the exact level gap identity**. We compare the theoretically estimated MSE (red) against the actual empirical measurements (black) for both PatchTST (left) and iTransformer (right) across multiple datasets with varying prediction horizons.

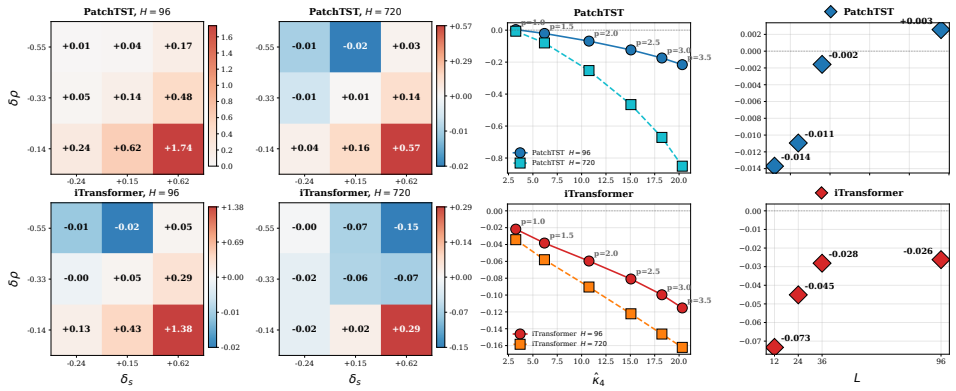

Figure 8: **Synthetic controlled experiments. Left:** Variation of the deployment gap $g_{B|A=1}$ with respect to δ_s and δ_ρ. **Middle:** Gap as a function of the innovation kurtosis κ_4; point labels indicate the synthesis parameter p. **Right:** Gap as a function of the lookback length L at fixed $H = 720$.

isolated level component to the overall MSE, with both exhibiting consistent trends across different datasets and forecast horizons. More detailed data for each component are provided in Table 4.

7.2 Causal Validation of the Variance Normalization Gating Condition

We first employ synthetic control to verify causality by manipulating the amplitude information proxy ρ_s. The results indicate that, the gap ($g_{B|A=1}$) increases monotonically as ρ_s decreases. This provides direct causal evidence that a lower ρ_s structurally favors the Full strategy by eliminating the penalty associated with amplitude-induced variance (Figure 6). Subsequently, we sequentially validate the response of the gating mechanism to the remaining theoretical parameters (Figure 8). Regarding scale drift and variation, experiments confirm that the gap widens with increasing δ_s and δ_ρ, suggesting that the benefits of the Full strategy are indeed amplified when non-stationary scale dynamics become more pronounced. As for the innovation kurtosis $\hat{\kappa}_4$, the results show that heavier noise tails actually diminish the advantage of the Full scheme's robust estimation. Finally, experiments on the lookback length L validate the correctness of the $(\kappa_4 - 1)/4L$ formula, revealing a clear $1/L$ decay trend in the gap.

Meanwhile, we observe an asymmetry in the effects of the two drift quantities: neither δ_s nor δ_ρ produces a comparably large negative gap when taking substantially negative values. This is consistent with expectations, as these quantities measure the magnitude of removable distributional shift; negative values do not imply enhanced generalization. We also observe that, in practice, δ_s and δ_ρ are predominantly negative or only marginally positive on real-world datasets (Table 3). Consequently, we may focus solely on structural bias and propose the following empirical rule: Full RevIN is most advantageous when both ρ_s and $\hat{\kappa}_4$ remain relatively small, as illustrated in Figure 6. For instance, on the Traffic dataset, extremely small ρ_s corresponds to a positive gap. Nevertheless, this observation remains purely empirical. Current evidence supporting Full RevIN is limited to this single dataset and suffers from insufficient sample size, thus requiring further investigations.

8 Conclusion

In this work, we challenge the default use of full RevIN and propose a risk decomposition framework to explain its effectiveness. Our analysis reveals a trade-off between the structural floor introduced by normalization and the transfer distance it mitigates. However, the choice between Mean and Full RevIN remains unresolved. Since the clear advantage of Full is observed only on Traffic, our current decision rule remains empirical. A natural direction for future work is to derive a closed-form Mean-versus-Full gap, analogous to our Raw-versus-Mean analysis, to better understand the role of variance normalization.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, figure and table generation, code implementation, and literature searches for related work. We did not use generative AI tools to write the manuscript or perform mathematical derivations, and data synthesis is not applicable to this work. All AI-assisted outputs were carefully reviewed by the authors: polished text was manually verified to ensure fidelity to our original ideas; AI-generated code was rigorously tested to confirm its correctness; retrieved references were checked against primary sources; and generated figures and tables were validated for accuracy. We take full responsibility for the final content of this work, including the text, claims, code, and artifacts produced with the assistance of generative AI.

REFERENCES

Ihab Ahmed, Denis Krompaß, Cheng Feng, and Volker Tresp. A comparative study on how data normalization affects zero-shot generalization in time series foundation models, 2025.

Gaspard Berthelier, Tahar Nabil, Etienne Le Naour, Richard Niamke, Samir Perlaza, and Giovanni Neglia. On the role of reversible instance normalization, 2026. URL https://arxiv.org/abs/2603.11869.

Si-An Chen, Chun-Liang Li, Sercan Ö. Arik, Nathanael C. Yoder, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023.

Tao Dai, Beiliang Wu, Peiyuan Liu, Naiqi Li, Yuerong Xue, Shu-Tao Xia, and Zexuan Zhu. DDN: Dual-domain dynamic normalization for non-stationary time series forecasting. In *Advances in Neural Information Processing Systems*, 2024.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan K. Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with TiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023.

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

Fanzhe Fu and Yang Yang. Noise or signal? deconstructing contradictions and an adaptive remedy for reversible normalization in time series forecasting, 2025. URL https://arxiv.org/abs/2510.04667.

Lu Han, Han-Jia Ye, and De-Chuan Zhan. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7129–7142, 2024a.

Lu Han, Han-Jia Ye, and De-Chuan Zhan. SIN: Selective and interpretable normalization for long-term time series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17437–17453. PMLR, 2024b. URL https://proceedings.mlr.press/v235/han24e.html.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

Chenghan Li, Mingchen Li, and Ruisheng Diao. TVNet: A novel time series analysis method based on dynamic convolution and 3d-variation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=MZDdTzN6Cy.

Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. Revisiting long-term time series forecasting: An investigation on linear mapping, 2023. URL https://arxiv.org/abs/2305.10721.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 37797–37814. PMLR, 2025. URL https://proceedings.mlr.press/v267/lin25e.html.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. SCINet: Time series modeling and forecasting with sample convolution and interaction. In *Advances in Neural Information Processing Systems*, 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 *International Conference on Learning Representations*, 2024.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-context transformers for unified time series forecasting. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=KMCJXjlDDr.

Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. In *Advances in Neural Information Processing Systems*, 2023.

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

Peiwang Tang and Weitai Zhang. Unlocking the power of patch: Patch-based mlp for long-term time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12640–12648. AAAI Press, 2025. doi: 10.1609/aaai.v39i12.33378.

Huiqiang Wang, Jieming Shi, and Li Qing. FACT: Fine-grained across-variable convolution for multivariate time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=j3gNYqrHtl.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. TimeMixer++: A General Time Series Pattern Machine for Universal Predictive Analysis. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=1CLzLXSFNn.

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

Yuxuan Wang, Haixu Wu, Yuezhou Ma, Yuchen Fang, Ziyi Zhang, Yong Liu, Shiyu Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Exploring accuracy law for deep time series forecasters: An empirical study, 2025b.

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

Weiwei Ye, Songgaojun Deng, Qiaosha Zou, and Ning Gui. Frequency adaptive normalization for non-stationary time series forecasting. In *Advances in Neural Information Processing Systems*, 2024.

Zhenan Yu, Guangxin Jiang, and Jin Yang. Three-stage learning unlocks strong performance in simple models for long-term time series forecasting, 2026. URL https://arxiv.org/abs/2605.13678.

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

A Experimental setup

Datasets and evaluation protocol. We evaluate on nine public benchmarks: ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Traffic, Weather, Exchange, and Solar-Energy. The main-text experiments use the first seven datasets, while the latter two are included in the appendix. All experiments follow the Time-Series-Library (TSLib) (Wang et al., 2024), including its data loaders and train/validation/test splits. Dataset statistics and split details are summarized in Table 1.

We study multivariate forecasting with a fixed lookback window of $L = 96$ and prediction horizons $H \in 96, 192, 336, 720$. All models are trained with Adam, an initial learning rate of 10^{-4}, and the `type3` decay schedule. Training runs for at most 100 epochs with early stopping (patience = 3) based on validation loss, and the best checkpoint is used for testing. The batch size is 32, except for TiDE and Crossformer on Electricity and Traffic, where it is reduced to 4 for memory efficiency. We report test-set MSE and MAE on the original scale, after reversing the corresponding normalization.

Baseline Details. The main-text study covers six backbones spanning the major architecture families: SCINet Liu et al. (2022), PatchTST Nie et al. (2023), iTransformer Liu et al. (2024), TiDE Das et al. (2023), Crossformer Zhang & Yan (2023), and RLinear Li et al. (2023). Here, RLinear refers to a simple linear model combined with RevIN and serves as a linear baseline in our study. The extended experiments in this appendix additionally include TimesNet Wu et al. (2023) and TSMixer Chen et al. (2023). All models use the TSLib implementations with a unified hyperparameter configuration where applicable: $d_{\text{model}} = 128$, $d_{\text{ff}} = 512$, $n_{\text{heads}} = 8$, $e_{\text{layers}} = 2$, dropout $= 0.1$, and GELU activation.

Table 1: Dataset detailed descriptions. The dataset size is organized in (Train, Validation, Test).

Tasks	Dataset	Dim	Series Length	Dataset Size	Frequency	Information
Long-term Forecasting	ETTm1	7	{96, 192, 336, 720}	(34465, 11521, 11521)	15min	Temperature
	ETTm2	7	{96, 192, 336, 720}	(34465, 11521, 11521)	15min	Temperature
	ETTh1	7	{96, 192, 336, 720}	(8545, 2881, 2881)	15 min	Temperature
	ETTh2	7	{96, 192, 336, 720}	(8545, 2881, 2881)	15 min	Temperature
	Electricity	321	{96, 192, 336, 720}	(18317, 2633, 5261)	Hourly	Electricity
	Traffic	862	{96, 192, 336, 720}	(12185, 1757, 3509)	Hourly	Transportation
	Exchange	8	{96, 192, 336, 720}	(5120, 665, 1422)	Daily	Weather
	Weather	21	{96, 192, 336, 720}	(36792, 5271, 10540)	10 min	Weather
	Solar-Energy	137	{96, 192, 336, 720}	(36601, 5161, 10417)	10min	Electricity

B Synthetic Control Construction

This section describes the construction of the synthetic control datasets used in the main text. Our goal is to independently control the key factors that determine the performance gap between Mean and Full, allowing each diagnostic quantity to vary while leaving other factors as unchanged as possible. Specifically, we control the quantities r_s, ρ_s, δ_s, δ_ρ, and $\hat{\kappa}_4$ defined in the main text, using corresponding reparameterization procedures to induce approximately orthogonal variations.

Here, $\rho_s = r_{\text{tr}} + r_{\text{te}}$ characterizes the overall coupling strength between scale and future amplitude and is a structural factor contributing to the structural floor of Full. Changing ρ_s alters the scale-to-amplitude mapping that the model needs to learn. Therefore, to isolate its structural effect, we reparameterize both the training and test sets, varying r_{tr} and r_{te} in the same direction while keeping $\delta_\rho = r_{\text{te}} - r_{\text{tr}}$ fixed, thereby avoiding additional amplitude-coupling drift. In contrast, δ_ρ characterizes the coupling shift from training to test and can therefore be controlled by modifying only the test set. Similarly, although $\hat{\kappa}_4$ also contributes to the structural cost, it affects the estimation noise of the sample scale at test time without changing the predictive mapping that the model needs to learn. It can therefore be controlled by reparameterizing only the innovation distribution of the test set. We use ETTh1 as an example to illustrate the synthetic reparameterizations. Figure 9

shows how the resulting series change visually after each intervention, while Table 2 summarizes the corresponding changes in the controlled statistics.

B.1 Controlling Distributional Drift

Intervention design. To control the two distributional-shift quantities δ_ρ and δ_s through distinct interventions, we construct a test-only intervention parameterized by α and β. We first decompose the test series as

$$x(t) = m(t) + \mathrm{dev}(t), \tag{8}$$

where $m(t)$ denotes the trailing mean over a window of length L, and $\mathrm{dev}(t)$ denotes the corresponding deviation. For each input window w in the test set, we compute its window scale $\hat{\sigma}_w$ and standardize these scales to obtain z_w. We then assign a gain $g_w = \exp(\alpha z_w)$. Because multiple input windows may cover the same future time point, we average the gains of all windows whose prediction intervals contain t, yielding a pointwise gain function $G(t)$. The modified test series is then defined as

$$x'(t) = m(t) + \beta\, G(t)\, \mathrm{dev}(t). \tag{9}$$

The two intervention parameters affect the scale dynamics through distinct mechanisms. The parameter α modulates the amplitude of future deviations as a function of the scale of the current window, thereby changing the test-set slope r_{te} and, consequently, the coupling drift $\delta_\rho = r_{\mathrm{te}} - r_{\mathrm{tr}}$. In contrast, β uniformly rescales the test deviations and therefore changes the overall scale drift δ_s. Because both a_i and $\hat{\sigma}_i$ are multiplied by the same factor under this transformation, β does not alter their least-squares slope. Since the intervention is applied only to the test set, the training-set quantities r_{tr} and $\hat{\kappa}_{4,\mathrm{tr}}$ remain unchanged across all synthetic configurations.

Configuration selection. The mapping from α to the resulting $r_{\mathrm{te}}(H = 720)$ is not one-to-one. Therefore, rather than uniformly spacing the intervention parameter, we select configurations based on the induced values of the quantity of interest. Specifically, we first evaluate Equation (9) over 81 densely spaced values of $\alpha \in [-1.2, 1.2]$ to determine the attainable range of $r_{\mathrm{te}}(H = 720)$. We then place nine target values uniformly across this range and, for each target, retain the value of α whose induced $r_{\mathrm{te}}(H = 720)$ is closest to the target. This procedure yields synthetic configurations that are approximately uniform in $r_{\mathrm{te}}(H = 720)$, rather than in the original intervention parameter.

Choice of the selection axis. We use $r_{\mathrm{te}}(H = 720)$ as the common selection axis for all forecast horizons. The main reason is that the intervention produces a wider attainable range of r_{te} at $H = 720$, allowing the selected configurations to cover a broader range of scale-amplitude coupling strengths. We therefore perform the inverse selection described above using $r_{\mathrm{te}}(H = 720)$ and use the resulting nine configurations as the common synthetic grid. Although the grid is selected at $H = 720$, the induced $r_{\mathrm{te}}(H = 96)$ values preserve a highly consistent ordering with those at $H = 720$, with a rank correlation of approximately 0.97. This supports using the same synthetic configurations for both forecast horizons.

B.2 Controlling Innovation Kurtosis

Intervention design. This intervention is designed to primarily vary the estimation-noise term $(\hat{\kappa}_4 - 1)/4L$ while suppressing the scale-amplitude coupling channel. We first whiten the test split by decomposing the series as $x = m + \mathrm{dev}$ and dividing the deviation by its rolling standard deviation s: $x'(t) = m(t) + \varepsilon(t), \varepsilon(t) = \mathrm{dev}(t)/s(t)$. This whitening substantially reduces the dependence between the current scale and future amplitude, driving the resulting slope toward $r_{\mathrm{te}} \approx 0$. We then reshape the distribution of the whitened residual using a signed power transform, $\varepsilon_p(t) = \mathrm{sign}(\varepsilon(t))|\varepsilon(t)|^p$, and re-standardize it using its own rolling standard deviation s_p:

$$x'(t) = m(t) + \varepsilon_p(t)/s_p(t). \tag{10}$$

The re-standardization preserves the local marginal scale of the residual, while the power transform changes its tail behavior and hence its window-level kurtosis. In particular, increasing p produces progressively heavier tails and increases $\hat{\kappa}_4$.

Configuration Selection. We select six values of p to obtain a range of $\hat{\kappa}_{4,\mathrm{te}}$ values while keeping the scale-amplitude coupling close to zero. The resulting configurations are then used as the synthetic doses for the kurtosis intervention.

Table 2: **Synthetic interventions on ETTh1.** (a) Structural-coupling intervention: δ_s is pinned to its baseline value throughout the intervention, while $\hat{\kappa}_4$ is measured on the training set. (b) Kurtosis sweep (test-only): $\hat{\kappa}_{4,\mathrm{te}}$ varies substantially, while the coupling drift is held fixed; δ_s decreases modestly with p but remains positive. (c) Drift grid (test-only): different rows correspond to different α values, while β varies δ_s without affecting r_{te}.

(a) Structural-coupling intervention (H=720)

α	r_{tr}	r_{te}	ρ_s	δ_ρ	δ_s	$\hat{\kappa}_4$
0	0.677	0.175	0.852	−0.502	+0.100	3.02
0.05	0.709	0.209	0.918	−0.500	+0.100	3.01
0.55	0.809	0.365	1.174	−0.444	+0.100	2.97
1.05	0.835	0.432	1.267	−0.403	+0.100	2.96
2.00	0.817	0.402	1.219	−0.415	+0.099	2.94

(b) Kurtosis intervention

p	r_{te}	δ_s	$\hat{\kappa}_{4,\mathrm{te}}$
1.0	0.074	+0.588	3.24
1.5	0.090	+0.587	6.18
2.0	0.101	+0.517	10.76
2.5	0.104	+0.421	15.06
3.0	0.102	+0.307	18.24
3.5	0.095	+0.161	20.33

(c) Drift intervention

	r_{te}		δ_s		
α	H=96	H=720	β=0.7	β=1.0	β=1.5
−0.33	0.361	0.024	−0.425	−0.068	+0.337
−0.09	0.310	0.096	−0.307	+0.050	+0.455
−0.48	0.445	0.169	−0.411	−0.054	+0.351
−0.57	0.491	0.235	−0.393	−0.037	+0.369
+0.24	0.586	0.311	−0.103	+0.254	+0.659
−0.81	0.630	0.379	−0.338	+0.019	+0.424
+0.54	0.765	0.460	+0.123	+0.479	+0.885
−1.05	0.726	0.526	−0.271	+0.085	+0.491
−1.17	0.756	0.611	−0.234	+0.123	+0.528

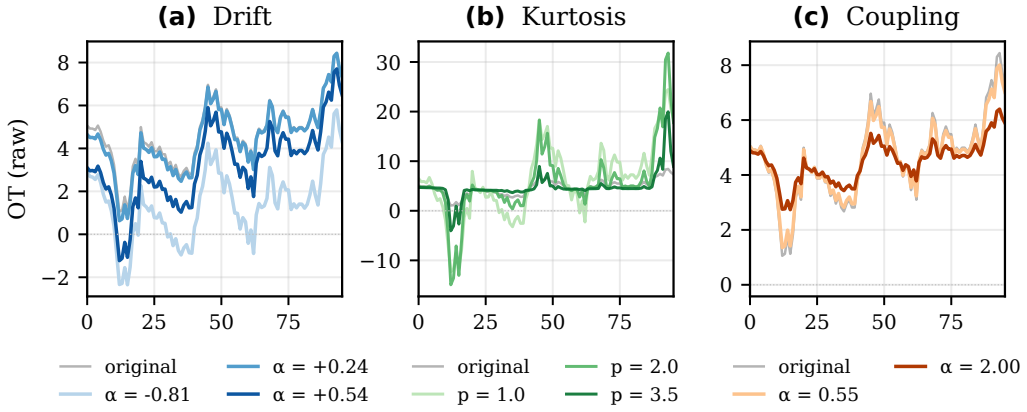

Figure 9: **Visualization of the synthetic control series on ETTh1.** Each panel shows the reparameterized OT channel over the same randomly selected 96-step window from the test segment. **(a) Drift:** the gain field $\exp(\alpha z_w)$ is applied to the segment with $\beta = 1.0$. Positive α amplifies the test amplitude, shifting the light-blue curves upward, while negative α compresses it. **(b) Kurtosis:** the series is first whitened, followed by a signed power transform ε_p and re-normalization. Darker curves indicate heavier tails, with the most extreme outliers stretched upward. **(c) Structural-coupling:** the stationary gain field is applied jointly to the training and test sets. Larger α compresses the amplitude in regions with high $\hat{\sigma}_w$, causing the orange curves to fall below the gray baseline.

B.3 Controlling Scale-Amplitude Coupling

Intervention design. This intervention targets the structural amplitude-coupling quantity $\rho_s = r_{\mathrm{tr}} + r_{\mathrm{te}}$. Unlike the previous two interventions, the gain field is applied simultaneously to the training and test sets. The pointwise gain field is constructed in the same way as in Appendix B.1,

but is applied to all splits, thereby keeping the training and test distributions aligned and introducing no train-test distribution shift. As a result, the resulting change in normalization performance can be attributed primarily to the structural cost associated with amplitude coupling. We take two steps to control potential confounding factors. First, within each split, we normalize the gain field to have unit geometric mean, thereby removing the level shift that would otherwise be introduced by $\exp(\alpha z) > 1$. Second, we iteratively adjust the bin-wise mean of $\log \hat{\sigma}$ to match that of the original dataset, ensuring that δ_s fixed at its baseline value throughout the intervention.

Configuration Selection. The mapping from the intervention parameter α to the induced scale-amplitude coupling is nonlinear. We therefore select dose levels based on the induced coupling rather than uniformly spacing the intervention parameter. Specifically, we evaluate a dense grid of α values, measure the resulting coupling strength, and then select configurations that are approximately evenly spaced along the induced coupling axis. This procedure provides a more uniform coverage of the structural factor of interest while avoiding over-representation of regions where the intervention parameter has little effect.

Choice of the selection axis. We use the induced coupling at $H = 720$ as the selection axis because the stationary intervention produces a substantially wider range of coupling strengths at $H = 720$ than at $H = 96$, where the baseline coupling is already relatively high. This provides better separation between dose levels and allows us to construct a more informative synthetic grid.

C A more detailed two-distribution risk decomposition

Fix a scheme T with envelope $\nu_T := (\mu_T, \sigma_T)$, normalized feature $\varphi_T := (x - \mu_T(x))/\sigma_T(x)$, and standardized target $z := (y - \mu_T(x))/\sigma_T(x)$. Since $\hat{y} - y = \sigma_T\,(g - z)$,

$$\mathcal{L}_Q(g) \;:=\; \frac{1}{H}\,\mathbb{E}_Q\big\|\hat{y} - y\big\|^2 \;=\; \frac{1}{H}\,\mathbb{E}_Q\Big[\sigma_T^2\,\big\|z - g\big\|^2\Big].$$

Theorem C.1 (Weighted Pythagorean decomposition)**.** *For any distribution Q with finite second moments,*

$$\mathcal{L}_Q(g) \;=\; R_Q^* \;+\; \frac{1}{H}\,\mathbb{E}_Q\Big[\sigma_T^2\,\big\|m_Q - g\big\|^2\Big] \qquad m_Q := \mathbb{E}_Q\big[z \mid \nu_T, \varphi_T\big] \tag{11}$$

with $R_Q^* := \frac{1}{H}\,\mathbb{E}_Q\big[\sigma_T^2\,\|z - m_Q\|^2\big]$.

Definition C.2 (Oracle, structural floor, transfer distance)**.** Let $\mathcal{G}$ be the hypothesis class of g.

- The *within-class oracle* under Q is $\bar{g}_Q^* := \arg\min_{g\in\mathcal{G}} \mathbb{E}_Q\big[\sigma_T^2\|m_Q - g\|^2\big]$.
- The *structural floor* of scheme T under Q is $\Delta_T(Q) := \frac{1}{H}\,\mathbb{E}_Q\big[\sigma_T^2\|m_Q - \bar{g}_Q^*\|^2\big] \geq 0$.
- The *transfer distance* of a trained predictor $\hat{g}$ is $\Theta_T(\mathrm{tr}\to\mathrm{te}) := \frac{1}{H}\,\mathbb{E}_{\mathrm{te}}\big[\sigma_T^2\|\bar{g}_{\mathrm{te}}^* - \hat{g}\|^2\big]$.

Both Δ_T and Θ_T split additively into level and shape components: $\Delta_T = \Delta_{T,\mathrm{level}} + \Delta_{T,\mathrm{shape}}$, $\Theta_T = \Theta_{T,\mathrm{level}} + \Theta_{T,\mathrm{shape}}$.

Theorem C.3 (Two-distribution master identity, restated)**.** *Let $\hat{g}$ be any predictor obtained by training scheme T on P_{tr}, and assume the oracle satisfies the projection orthogonality $\mathbb{E}_{\mathrm{te}}\big[\sigma_T^2\,\langle m_{\mathrm{te}} - \bar{g}_{\mathrm{te}}^*,\ \bar{g}_{\mathrm{te}}^* - g\rangle\big] = 0$ for all $g \in \mathcal{G}$. Then*

$$\mathcal{L}_{\mathrm{te}}(\hat{g}) \;=\; R_{\mathrm{te}}^* \;+\; \Delta_T(\mathrm{te}) \;+\; \Theta_T(\mathrm{tr}\to\mathrm{te}) \tag{12}$$

D Proofs for Section 4

D.1 Proof of Theorem C.1

Let $\mathcal{F} := \sigma(\nu_T, \varphi_T)$. By construction, $\sigma_T^2 \in \mathcal{F}$, $g \in \mathcal{F}$, and $m_Q = \mathbb{E}_Q[z \mid \mathcal{F}] \in \mathcal{F}$. Set $h := m_Q - g \in \mathcal{F}$. Insert $\pm m_Q$:

$$\begin{aligned}\|z - g\|^2 &= \|(z - m_Q) + (m_Q - g)\|^2 \\ &= \|z - m_Q\|^2 + \|m_Q - g\|^2 + 2\,\langle z - m_Q,\, h\rangle.\end{aligned}$$

Multiply by σ_T^2/H and take $\mathbb{E}_Q$. The cross term vanishes by the tower property:

$$\begin{aligned}\mathbb{E}_Q\big[\sigma_T^2\,\langle z - m_Q, h\rangle\big] &= \mathbb{E}_Q\big[\mathbb{E}_Q[\sigma_T^2\,\langle z - m_Q, h\rangle \mid \mathcal{F}]\big] \\ &= \mathbb{E}_Q\big[\sigma_T^2\,\langle \mathbb{E}_Q[z - m_Q \mid \mathcal{F}],\, h\rangle\big] \\ &= \mathbb{E}_Q\big[\sigma_T^2\,\langle 0,\, h\rangle\big] \;=\; 0,\end{aligned}$$

where the second line uses $\sigma_T^2, h \in \mathcal{F}$, and the third uses $\mathbb{E}_Q[z \mid \mathcal{F}] = m_Q$. Therefore

$$\mathcal{L}_Q(g) = \underbrace{\frac{1}{H}\mathbb{E}_Q[\sigma_T^2\,\|z - m_Q\|^2]}_{=:\,R_Q^*} + \frac{1}{H}\mathbb{E}_Q[\sigma_T^2\,\|m_Q - g\|^2].$$

The second term is ≥ 0 with equality iff $g = m_Q$ $\mathcal{F}$-a.s., so m_Q minimizes $\mathcal{L}_Q$ over $\mathcal{F}$-measurable predictors. □

D.2 Proof of Theorem 4.1

Apply Theorem C.1 with $Q = P_{\mathrm{te}}$:

$$\mathcal{L}_{\mathrm{te}}(\hat{g}) = R_{\mathrm{te}}^* + \frac{1}{H}\,\mathbb{E}_{\mathrm{te}}[\sigma_T^2\,\|m_{\mathrm{te}} - \hat{g}\|^2].$$

Set $d := \bar{g}_{\mathrm{te}}^* - \hat{g}$. Since $\bar{g}_{\mathrm{te}}^*, \hat{g} \in \mathcal{G}$ and $\mathcal{G}$ is a linear subspace, $d \in \mathcal{G}$. Insert $\pm\bar{g}_{\mathrm{te}}^*$:

$$\begin{aligned}\|m_{\mathrm{te}} - \hat{g}\|^2 &= \|(m_{\mathrm{te}} - \bar{g}_{\mathrm{te}}^*) + d\|^2 \\ &= \|m_{\mathrm{te}} - \bar{g}_{\mathrm{te}}^*\|^2 + \|d\|^2 + 2\,\langle m_{\mathrm{te}} - \bar{g}_{\mathrm{te}}^*,\, d\rangle.\end{aligned}$$

The cross term is zero by the first-order condition for the oracle. Define $\Phi(g) := \mathbb{E}_{\mathrm{te}}[\sigma_T^2\,\|m_{\mathrm{te}} - g\|^2]$. Since $\bar{g}_{\mathrm{te}}^* = \arg\min_{g\in\mathcal{G}} \Phi(g)$ and $\bar{g}_{\mathrm{te}}^* + td \in \mathcal{G}$ for all $t \in \mathbb{R}$,

$$0 = \frac{d}{dt}\Big|_{t=0} \Phi(\bar{g}_{\mathrm{te}}^* + td) = -2\,\mathbb{E}_{\mathrm{te}}\big[\sigma_T^2\,\langle m_{\mathrm{te}} - \bar{g}_{\mathrm{te}}^*,\, d\rangle\big].$$

Hence the cross term vanishes, and

$$\frac{1}{H}\mathbb{E}_{\mathrm{te}}[\sigma_T^2\,\|m_{\mathrm{te}} - \hat{g}\|^2] = \underbrace{\frac{1}{H}\mathbb{E}_{\mathrm{te}}[\sigma_T^2\,\|m_{\mathrm{te}} - \bar{g}_{\mathrm{te}}^*\|^2]}_{=:\,\Delta_T(\mathrm{te})} + \underbrace{\frac{1}{H}\mathbb{E}_{\mathrm{te}}[\sigma_T^2\,\|\bar{g}_{\mathrm{te}}^* - \hat{g}\|^2]}_{=:\,\Theta_T(\mathrm{tr}\to\mathrm{te})}.$$

Combining with R_{te}^* yields Equation (12). □

E Proof of Section 5

E.1 Proof of Theorem 5.1

By definition of the two population least-squares regressions,

$$\begin{aligned}u &= \mathbb{E}_s[u] + c_s\,(\bar{x} - \ell_s) + \eta, &\quad \mathbb{E}_s[\eta] = 0,\;\; \mathrm{Cov}_s(\bar{x}, \eta) = 0, \\ \bar{y} &= \mathbb{E}_s[\bar{y}] + \beta_s^*\,(\bar{x} - \ell_s) + \varepsilon_s, &\quad \mathbb{E}_s[\varepsilon_s] = 0,\;\; \mathrm{Cov}_s(\bar{x}, \varepsilon_s) = 0,\end{aligned}$$

with $c_s = \mathrm{Cov}_s(u, \bar{x})/V_s$ and $\beta_s^* = \mathrm{Cov}_s(\bar{y}, \bar{x})/V_s$. Subtracting and writing $b_s := \mathbb{E}_s[u] - \mathbb{E}_s[\bar{y}] = \mathbb{E}_s[u - \bar{y}]$,

$$u - \bar{y} = b_s + (c_s - \beta_s^*)\,(\bar{x} - \ell_s) + (\eta - \varepsilon_s).$$

Set $A := b_s$, $B := (c_s - \beta_s^*)(\bar{x} - \ell_s)$, $C := \eta - \varepsilon_s$. We claim $\mathbb{E}_s[AB] = \mathbb{E}_s[AC] = \mathbb{E}_s[BC] = 0$. Each is a one-line computation:

$$\begin{aligned}&\mathbb{E}_s[AB] = (c_s - \beta_s^*)\,b_s\,\mathbb{E}_s[\bar{x} - \ell_s] = 0, && (\mathbb{E}_s[\bar{x} - \ell_s] = 0) \\ &\mathbb{E}_s[AC] = b_s\,(\mathbb{E}_s[\eta] - \mathbb{E}_s[\varepsilon_s]) = 0, && (\mathbb{E}_s[\eta] = \mathbb{E}_s[\varepsilon_s] = 0) \\ &\mathbb{E}_s[BC] = (c_s - \beta_s^*)\,\big(\mathbb{E}_s[(\bar{x} - \ell_s)\eta] - \mathbb{E}_s[(\bar{x} - \ell_s)\varepsilon_s]\big) \\ &\qquad = (c_s - \beta_s^*)\,\big(\mathrm{Cov}_s(\bar{x}, \eta) - \mathrm{Cov}_s(\bar{x}, \varepsilon_s)\big) = 0.\end{aligned}$$

$$\mathbb{E}_s[(u - \bar{y})^2] = \mathbb{E}_s[A^2] + \mathbb{E}_s[B^2] + \mathbb{E}_s[C^2] = b_s^2 + (c_s - \beta_s^*)^2\,V_s + \mathbb{E}_s[(\eta - \varepsilon_s)^2],$$

which is Equation (3). □

E.2 Function Class Inclusion

A scheme's hypothesis class inherits the form of its forecast map. For Raw, $\hat{y}_{\text{raw}} = F_{\text{raw}}(x)$ for F_{raw} ranging over the backbone. For Mean-only, $\hat{y}_{\text{mean}} = \bar{x}\,\mathbf{1}_H + F_{\text{mean}}(\tilde{x})$ where $\tilde{x} = x - \bar{x}\,\mathbf{1}_H$ and F_{mean} ranges over the same backbone. Since $\bar{x}$ and $\tilde{x}$ are both deterministic functions of x, every Mean-only map admits the representation

$$\hat{y}_{\text{mean}}(x) = \bar{x}\,\mathbf{1}_H + F_{\text{mean}}(\tilde{x}) =: G(x),$$

for some measurable G. The class of all F_{raw} already ranges over every measurable map of x, so picking $F_{\text{raw}}(x) = G(x)$ realizes the same forecast. Hence $\mathcal{G}_{\text{mean}} \subseteq \mathcal{G}_{\text{raw}}$.

For the strictness, restrict to the level channel: the conditional level response $u_s(\bar{x}) := \mathbb{E}_s[\hat{\bar{y}} \mid \bar{x}]$. For Mean-only, $u_s(\bar{x}) = \mathbb{E}_s[\bar{x} + F_{\text{mean}}(\tilde{x}) \cdot \mathbf{1}_H/H \mid \bar{x}]$. When F_{mean} has zero-mean prediction steps (which holds for backbones with no output bias), $\mathbb{E}_s[F_{\text{mean}}(\tilde{x}) \mid \bar{x}] = 0$ under the standard independence $\tilde{x} \perp \bar{x}$ in P_s, so $u_s(\bar{x}) = \bar{x}$. The level response is the identity map. For Raw, $u_s(\bar{x}) = \mathbb{E}_s[F_{\text{raw}}(x) \mid \bar{x}]$ ranges over all measurable functions of $\bar{x}$ as F_{raw} varies, in particular over affine maps of arbitrary slope. Therefore $\mathcal{G}_{\text{mean}} \subsetneq \mathcal{G}_{\text{raw}}$: the level slope c is pinned at 1 for Mean-only but free for Raw.

E.3 Stationary Case

Let $P_{\text{tr}} = P_{\text{te}} =: P$ and drop the subscript s. Under Appendix E.2, $\mathcal{G}_{\text{raw}} \supseteq \mathcal{G}_{\text{mean}}$; with unlimited capacity, Raw's oracle attains the Bayes risk, so $\mathcal{L}_*^{\text{raw}} = R^*$.

Substitute Mean-only's pinned level response $u = \bar{x}$ (hence $c = 1$, $\eta \equiv 0$) into Equation (3):

$$C_{\text{mean}}^{\text{lev}} = \mathbb{E}[(\bar{x} - \bar{y})^2] = b^2 + (1-\beta^*)^2 V + \mathbb{E}[\varepsilon^2], \qquad b = \mathbb{E}[\bar{x} - \bar{y}] = -\gamma. \tag{13}$$

For Raw's oracle, the level channel is unrestricted, so the unconstrained minimum over (b, c, η) in Equation (3) is attained at $b = 0$, $c = \beta^*$, $\eta = 0$:

$$C_{\text{raw}}^{\text{lev, or}} = \mathbb{E}[\varepsilon^2], \qquad \text{i.e. } \mathcal{L}_*^{\text{raw}} = R^* = \mathbb{E}[\varepsilon^2]. \tag{14}$$

Under standardized training ($\ell = \mathbb{E}[\bar{x}] = 0$) and an intercept-free level response, both oracles carry bias $-\gamma$, contributing γ^2 to each side of Equation (13)/Equation (14); this term cancels in the gap. Subtracting,

$$C_{\text{mean}}^{\text{lev}} - C_{\text{raw}}^{\text{lev, or}} = (1-\beta^*)^2 V =: \Lambda_\mu.$$

The shape-channel risks coincide since both schemes operate identically on $\tilde{x}$. Therefore

$$\mathcal{L}_*^{\text{mean}} - \mathcal{L}_*^{\text{raw}} = \Lambda_\mu = V(1-\beta^*)^2 \geq 0,$$

with equality iff $\beta^* = 1$.

Remark E.1 (Shift γ). If $\ell \neq 0$, or if Raw learns an intercept, the γ^2 term in Equation (13) is no longer shared; it enters as an additional fixed cost on whichever scheme is not centered. Under a deployment shift, this is precisely the bias term tracked by Theorem 5.2 and is measured, not assumed.

E.4 Proof of Theorem 5.2 (Level Gap of Raw vs. Mean-only)

Apply Equation (3) under P_{te} to each scheme. For Raw, with $u = \hat{\bar{y}}_{\text{raw}}$ and $\eta = \eta_{\text{raw}}$,

$$C_{\text{raw}}(\text{te}) = b_{\text{raw}}^2 + (c - \beta_{\text{te}}^*)^2\, V_{\text{te}} + \mathbb{E}_{\text{te}}[\eta^2] - 2\,\mathbb{E}_{\text{te}}[\eta\,\varepsilon_{\text{te}}] + \mathbb{E}_{\text{te}}[\varepsilon_{\text{te}}^2]. \tag{15}$$

For Mean-only, $u = \bar{x}$. Three substitutions apply: $c = 1$ (the regression slope of $\bar{x}$ on $\bar{x}$ is 1), $\eta \equiv 0$ (the regression residual is identically zero), and

$$b_{\text{mean}} = \mathbb{E}_{\text{te}}[\bar{x} - \bar{y}] = \ell_{\text{te}} - \mathbb{E}_{\text{te}}[\bar{y}] = -\gamma_{\text{te}}.$$

$$C_{\text{mean}}(\text{te}) = \gamma_{\text{te}}^2 + (1 - \beta_{\text{te}}^*)^2\, V_{\text{te}} + \mathbb{E}_{\text{te}}[\varepsilon_{\text{te}}^2]. \tag{16}$$

Subtracting Equation (16) from Equation (15), the shared noise floor $\mathbb{E}_{\text{te}}[\varepsilon_{\text{te}}^2]$ cancels, leaving

$$\text{gap}_{\text{level}} = \left(b_{\text{raw}}^2 - \gamma_{\text{te}}^2\right) + \Big((c - \beta_{\text{te}}^*)^2\, V_{\text{te}} - (1 - \beta_{\text{te}}^*)^2\, V_{\text{te}}\Big) + \Big(\mathbb{E}_{\text{te}}[\eta^2] - 2\,\mathbb{E}_{\text{te}}[\eta\,\varepsilon_{\text{te}}]\Big),$$

which is Equation (6) after recognizing $(1 - \beta_{\text{te}}^*)^2 V_{\text{te}} = \Lambda_\mu(\text{te})$. □

E.5 Proof of Proposition 6.1: Scale Model

We work under the scale model

$$x = \sigma\,\varphi, \quad \sigma \perp \varphi, \quad y_{\mathrm{dev}} := y - \bar{x}\mathbf{1}_H = \rho(\sigma)\,\sigma\,\varphi_{\mathrm{fut}} + \varepsilon,$$

with $\mathbb{E}[\varepsilon \mid \sigma, \varphi] = 0$, φ_{fut} a deterministic function of φ, and $\hat{\sigma} = \sigma(1+\xi)$ where $\mathbb{E}[\xi] = 0$ and $\xi \perp (\sigma, \varphi, \varepsilon)$.

Mean-only oracle. The Mean-only oracle observes both (σ, φ), so $g^*(\sigma, \varphi) = \rho(\sigma)\,\sigma\,\varphi_{\mathrm{fut}}$ attains the Bayes risk R^* on the dev signal. Hence its floor over R^* is 0.

Full oracle, noiseless envelope. With $\xi = 0$, Full predicts $\hat{y} - \bar{x}\mathbf{1}_H = \sigma\,h(\varphi)$, and the dev-excess risk over R^* is

$$\Phi(h) := \mathbb{E}\big[\sigma^2\,\|\rho(\sigma)\,\varphi_{\mathrm{fut}} - h(\varphi)\|^2\big].$$

Φ is strictly convex in $h(\varphi)$ over each fiber $\{\varphi = \phi\}$. The minimizer satisfies the normal equation obtained by considering perturbations $h + t\,u(\varphi)$ for arbitrary u:

$$0 = \frac{d}{dt}\Big|_{t=0} \Phi(h_0 + tu) = 2\,\mathbb{E}\big[\sigma^2\,\langle \rho(\sigma)\,\varphi_{\mathrm{fut}} - h_0(\varphi),\, u(\varphi)\rangle\big].$$

Conditioning on φ and using $\sigma \perp \varphi$ and $\varphi_{\mathrm{fut}} \in \sigma(\varphi)$,

$$\mathbb{E}\big[\sigma^2\,\rho(\sigma)\,\varphi_{\mathrm{fut}} \mid \varphi\big] - \mathbb{E}[\sigma^2 \mid \varphi]\,h_0(\varphi) = 0 \implies h_0(\varphi) = \frac{\mathbb{E}[\sigma^2\rho(\sigma)]}{\mathbb{E}[\sigma^2]}\,\varphi_{\mathrm{fut}} =: \bar{\rho}_w\,\varphi_{\mathrm{fut}}.$$

Substituting h_0 back,

$$\begin{aligned}
\Lambda_\sigma := \Phi(h_0) &= \mathbb{E}\big[\sigma^2\,\|\rho(\sigma)\,\varphi_{\mathrm{fut}} - \bar{\rho}_w\,\varphi_{\mathrm{fut}}\|^2\big] \\
&= \mathbb{E}\|\varphi_{\mathrm{fut}}\|^2 \cdot \mathbb{E}\big[\sigma^2\,(\rho(\sigma) - \bar{\rho}_w)^2\big] \\
&= \mathbb{E}\|\varphi_{\mathrm{fut}}\|^2 \cdot \Big(\mathbb{E}[\sigma^2\rho^2] - \tfrac{(\mathbb{E}[\sigma^2\rho])^2}{\mathbb{E}[\sigma^2]}\Big) \\
&= \mathbb{E}\|\varphi_{\mathrm{fut}}\|^2 \cdot \mathbb{E}[\sigma^2] \cdot \mathrm{Var}_w(\rho),
\end{aligned}$$

where the tilted measure $dw/d\mathbb{P} = \sigma^2/\mathbb{E}[\sigma^2]$ makes $\bar{\rho}_w = \mathbb{E}_w[\rho]$ and $\mathrm{Var}_w(\rho) = \mathbb{E}_w[(\rho - \bar{\rho}_w)^2]$. In particular $\Lambda_\sigma \geq 0$, with equality iff ρ is w-a.s. constant (the constant-ρ case of Proposition 6.1).

Full oracle, noisy envelope. Now $\hat{\sigma} = \sigma(1+\xi)$. Dev-excess risk:

$$\Phi_\xi(h) := \mathbb{E}\big[\sigma^2\,\|(1+\xi)\,h(\varphi) - \rho(\sigma)\,\varphi_{\mathrm{fut}}\|^2\big].$$

The normal equation, derived by the same first-order perturbation argument, is

$$\mathbb{E}\big[\sigma^2(1+\xi)^2\,\varphi_{\mathrm{fut}}\big] - \mathbb{E}[\sigma^2(1+\xi)^2 \mid \varphi]\,h^*(\varphi) = 0.$$

Using $\xi \perp (\sigma, \varphi)$, $\mathbb{E}[(1+\xi)^2] = 1 + \mathrm{Var}(\xi) =: 1 + v_\xi$, so

$$h^*(\varphi) = \bar{\rho}_w\,\varphi_{\mathrm{fut}}/(1+v_\xi) = h_0(\varphi)/(1+v_\xi).$$

Expand $\Phi_\xi(h^*)$ by splitting $(1+\xi)h^* - \rho\,\varphi_{\mathrm{fut}} = \underbrace{(h_0 - \rho\,\varphi_{\mathrm{fut}})}_{=:\,r_0} + \underbrace{\big((1+\xi)h^* - h_0\big)}_{=:\,\delta}$:

$$\Phi_\xi(h^*) = \mathbb{E}[\sigma^2\|r_0\|^2] + 2\mathbb{E}[\sigma^2\langle r_0, \delta\rangle] + \mathbb{E}[\sigma^2\|\delta\|^2].$$

The three terms are, to leading order in v_ξ,

$$\begin{aligned}
\mathbb{E}[\sigma^2\|r_0\|^2] &= \Lambda_\sigma, \\
\mathbb{E}[\sigma^2\langle r_0, \delta\rangle] &= -\tfrac{v_\xi}{1+v_\xi}\,\mathbb{E}[\sigma^2\langle r_0, h_0\rangle] = -\tfrac{v_\xi}{1+v_\xi} \cdot 0 \quad \text{(normal equation at } \xi = 0), \\
\mathbb{E}[\sigma^2\|\delta\|^2] &= \tfrac{\mathbb{E}[(1+\xi)^2] - 2\mathbb{E}[1+\xi] + 1}{(1+v_\xi)^2}\,\mathbb{E}[\sigma^2\|h_0\|^2] \\
&= \tfrac{v_\xi}{1+v_\xi}\,\mathbb{E}[\sigma^2\|h_0\|^2] = v_\xi\,\mathbb{E}\|\sigma h_0\|^2 + O(v_\xi^2).
\end{aligned}$$

Thus

$$\Phi_\xi(h^*) - \Lambda_\sigma = v_\xi\,\mathbb{E}\|\sigma h_0\|^2 + O(v_\xi^2).$$

Delta-method for v_ξ. Write $\hat{\sigma}^2/\sigma^2 = \frac{1}{L}\sum_{j=1}^{L}\eta_j^2$ for i.i.d. centered innovations η_j with $\mathbb{E}[\eta_j^2] = 1$ and kurtosis $\kappa_4 = \mathbb{E}[\eta_j^4]$. Then

$$\mathrm{Var}\big(\hat{\sigma}^2/\sigma^2\big) = \tfrac{1}{L}\mathrm{Var}(\eta_j^2) = \tfrac{\kappa_4-1}{L}.$$

Set $S := \hat{\sigma}^2/\sigma^2$ and $\xi := \sqrt{S} - 1$. With $f(s) = \sqrt{s}$, $f'(1) = 1/2$ and $f''(1) = -1/4$; Taylor-expanding $f(1 + (S-1))$ about $S = 1$,

$$\xi = \tfrac{1}{2}(S-1) - \tfrac{1}{8}(S-1)^2 + O((S-1)^3).$$

Taking expectations (using $\mathbb{E}[S] = 1$ and $\mathbb{E}[(S-1)^2] = \mathrm{Var}(S) = (\kappa_4 - 1)/L$),

$$\mathbb{E}[\xi] = 0 + O(L^{-2}), \qquad \mathrm{Var}(\xi) = \tfrac{1}{4}\mathrm{Var}(S) + O(L^{-2}) = \tfrac{\kappa_4-1}{4L} + O(L^{-2}).$$

Combining. Substituting $v_\xi = (\kappa_4 - 1)/(4L) + O(L^{-2})$ into the Full oracle risk and subtracting the Mean-only floor of 0,

$$\mathcal{L}_*^{\mathrm{full}}(Q) - \mathcal{L}_*^{\mathrm{mean}}(Q) = \Lambda_\sigma + \frac{\kappa_4 - 1}{4L}\,\mathbb{E}\|\sigma h_0\|^2 + O(L^{-2}),$$

which is Equation (7). Both coefficients are non-negative: $\kappa_4 \geq 1$ by $\mathrm{Var}(\eta^2) \geq 0$, with $\kappa_4 = 3$ for Gaussian innovations.

Upper bound on the tax coefficient. Since $h_0 = \bar{\rho}_w\,\varphi_{\mathrm{fut}}$ is the σ^2-weighted L^2-projection of the signal $\rho(\sigma)\,\varphi_{\mathrm{fut}}$ onto the fiber $\{\varphi_{\mathrm{fut}} \in \sigma(\varphi)\}$, projection is a contraction:

$$\mathbb{E}\|\sigma h_0\|^2 \leq \mathbb{E}\|\sigma\,\rho(\sigma)\,\varphi_{\mathrm{fut}}\|^2.$$

Further, since $y_{\mathrm{dev}} = \sigma\rho(\sigma)\varphi_{\mathrm{fut}} + \varepsilon$ with $\varepsilon \perp \sigma, \varphi$, $\mathbb{E}\|\sigma\rho(\sigma)\varphi_{\mathrm{fut}}\|^2 \leq \mathbb{E}\|y_{\mathrm{dev}}\|^2$. Hence $\mathbb{E}\|\sigma h_0\|^2 \leq \mathbb{E}\|y_{\mathrm{dev}}\|^2$, yielding the closed-form upper bound stated in Proposition 6.1.

Role of shape persistence. Under M1, $\varphi_{\mathrm{fut}} \in \sigma(\varphi)$, so the Full class can in principle represent $\varphi_{\mathrm{fut}} = v(\varphi)$ exactly and the only obstruction is scale-invariance: this is what makes Λ_σ the clean object above. Without M1, write $v(\varphi) := \mathbb{E}[\varphi_{\mathrm{fut}} \mid \varphi]$ and decompose $\varphi_{\mathrm{fut}} = v(\varphi) + \zeta$ with $\mathbb{E}[\zeta \mid \varphi] = 0$. The Full input contains no future information, so ζ enters the irreducible noise R^* and is excluded from Λ_σ; equivalently, replacing φ_{fut} by $v(\varphi)$ throughout leaves Λ_σ and the tax term unchanged. The gap is therefore invariant under M1. □

F Additional experimental results

F.1 Extended validation of the exact level-gap identity

We extend the validation of Theorem 5.2 beyond the main-text configurations to two additional backbones, TimesNet Wu et al. (2023) and TSMixer Chen et al. (2023), and to a broader dataset suite: ETTh1, ETTh2, ETTm1, ETTm2, Exchange, Solar-Energy, and Weather, at $H \in \{96, 720\}$ (Figure 10). Following the main-text protocol, a single Raw model is trained per configuration, and the deployment gap is estimated from its audit quantities alone, without training the Mean-only model. We note that the estimated gaps track the measured ones in both the level component and the overall MSE across all seven datasets.

F.2 Synthetic control experiment

We replicate the synthetic control study of Figure 8 with TimesNet and TSMixer as backbones, using ETTh1 and ETTm1 with $L = 96$ and $H \in 96, 720$ (Figure 11). The top two rows report the deployment gap $g_{B|A=1}$ across the (δ_s, δ_ρ) drift grid. For both models and both horizons, the gap increases with either drift coordinate, with the largest gains from the Full scheme concentrated in the high-drift corner, consistent with the main-text findings. The bottom row reports the kurtosis scan, where the gap decreases monotonically with the innovation kurtosis κ_4, indicating that heavier-tailed innovations attenuate the benefit of variance scaling through the scale-estimation channel.

F.3 Detailed Numerical Results

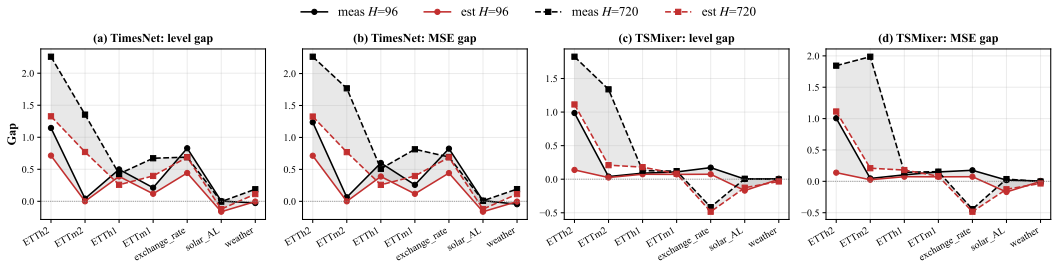

Figure 10: **Extended empirical validation of the exact level gap identity.** Estimated (red) versus measured (black) gaps for TimesNet (panels a, b) and TSMixer (panels c, d) on seven datasets at $H \in \{96, 720\}$, for the level component (a, c) and the overall MSE (b, d). The shaded region marks the estimation error.

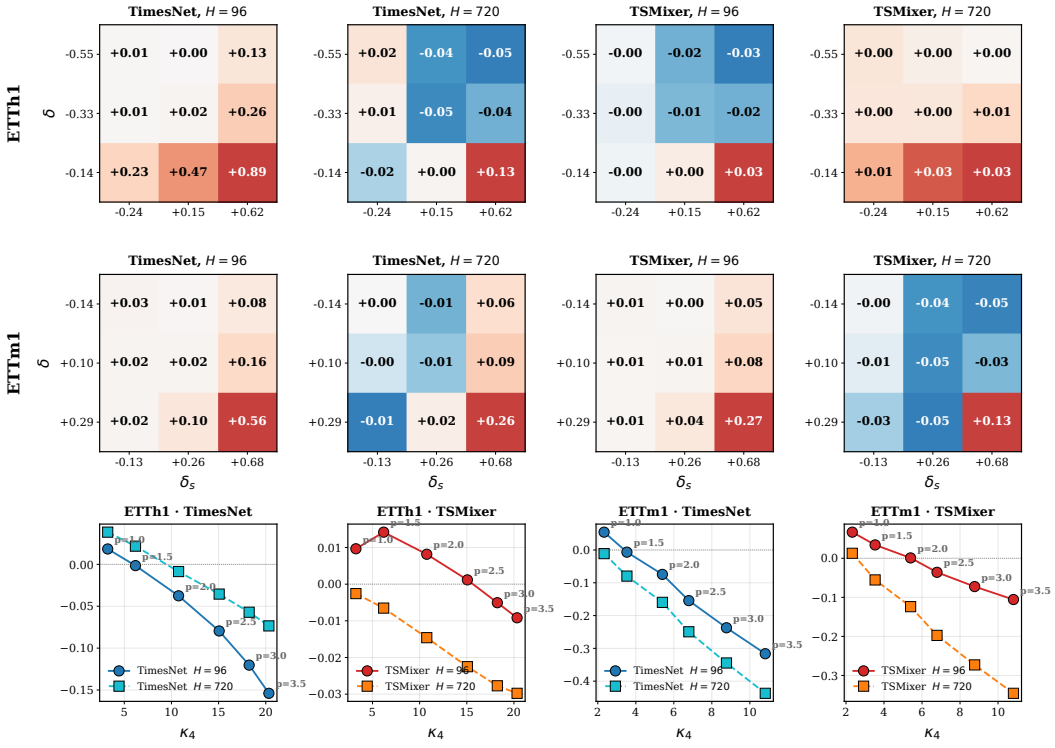

Figure 11: **Synthetic control validation of the variance division gate with TimesNet and TSMixer. Top two rows:** deployment gap $g_{B|A=1}$ over the (δ_s, δ_ρ) drift grid on reparameterized ETTh1 and ETTm1 (red: Full wins). **Bottom row:** gap versus innovation kurtosis $\hat{\kappa}_4$; point labels indicate the synthesis parameter p.

Table 3: Measurement of Dataset Drift Statistics

Dataset	δ_s	r^*_{tr}	r^*_{te}	δ_ρ	ρ_s	$\hat{\kappa}_4$
ETTh1	+0.100	+0.677	+0.175	-0.502	+0.852	3.02
ETTh2	-0.023	+0.884	+0.645	-0.240	+1.529	2.08
ETTm1	+0.072	+0.404	+0.177	-0.227	+0.581	2.44
ETTm2	-0.033	+1.158	+0.576	-0.582	+1.734	2.01
Weather	-0.496	+0.468	+0.407	-0.060	+0.875	2.04
Electricity	-0.025	+0.629	+0.496	-0.134	+1.125	1.81
Traffic	+0.075	+0.140	+0.005	-0.135	+0.145	5.89

Table 4: Estimated and measured gaps with raw audit quantities.

Model	H	Dataset	est_lvl	meas_lvl	meas_mse	bias	c_{eff}	β_T	V	γ	C_{raw}	C_{mean}
PatchTST	96	ETTh2	+0.0871	+0.0668	+0.0754	+0.2578	+0.7926	+0.9327	+1.3716	-0.0042	+0.0933	+0.0062
PatchTST	96	ETTm2	+0.0804	+0.0645	+0.0711	+0.2274	+0.8097	+0.9584	+1.4092	-0.0005	+0.0829	+0.0024
PatchTST	96	ETTh1	+0.0237	+0.0241	+0.0232	-0.0933	+0.7369	+0.9282	+0.4759	-0.0038	+0.0261	+0.0025
PatchTST	96	ETTm1	+0.0021	+0.0296	+0.0485	-0.0787	+0.8361	+0.8940	+0.5221	+0.0013	+0.0079	+0.0059
PatchTST	96	traffic	+0.0071	+0.0300	+0.0234	-0.0303	+0.3030	+0.6860	+0.1299	-0.0081	+0.0200	+0.0129
PatchTST	96	electricity	+0.0006	+0.0033	+0.0016	-0.0015	+0.7106	+0.8601	+0.2310	+0.0028	+0.0052	+0.0045
PatchTST	96	weather	+0.0006	+0.0095	+0.0076	+0.0288	+0.7734	+0.8857	+0.4671	-0.0025	+0.0067	+0.0061
PatchTST	720	ETTh2	+0.5230	+0.5441	+0.5818	+0.5201	+0.4812	+0.9271	+1.3048	+0.0031	+0.5299	+0.0070
PatchTST	720	ETTm2	+0.2476	+0.2432	+0.2696	+0.4206	+0.6354	+0.8846	+1.4491	-0.0046	+0.2669	+0.0193
PatchTST	720	ETTh1	+0.1046	+0.0961	+0.1155	-0.1526	+0.5002	+0.9444	+0.4295	+0.0458	+0.1080	+0.0034
PatchTST	720	ETTm1	+0.0255	+0.0544	+0.0780	-0.0572	+0.5834	+0.8428	+0.5238	-0.0055	+0.0385	+0.0130
PatchTST	720	traffic	+0.0142	+0.0343	+0.0298	-0.0526	+0.2298	+0.6730	+0.1282	-0.0020	+0.0279	+0.0137
PatchTST	720	electricity	-0.0070	-0.0011	-0.0012	-0.0009	+0.5895	+0.7562	+0.2212	+0.0025	+0.0062	+0.0132
PatchTST	720	weather	-0.0160	-0.0075	-0.0082	+0.1027	+0.5878	+0.7224	+0.4429	-0.0209	+0.0186	+0.0346
iTransformer	96	ETTh2	+0.1088	+0.2085	+0.2226	+0.2555	+0.7421	+0.9327	+1.3716	-0.0042	+0.1151	+0.0062
iTransformer	96	ETTm2	+0.0264	+0.0680	+0.0808	+0.1155	+0.8536	+0.9584	+1.4092	-0.0005	+0.0288	+0.0024
iTransformer	96	ETTh1	+0.0188	+0.0156	+0.0224	-0.1416	+0.8767	+0.9282	+0.4759	-0.0038	+0.0213	+0.0025
iTransformer	96	ETTm1	+0.0320	+0.0335	+0.0398	-0.1817	+0.7976	+0.8940	+0.5221	+0.0013	+0.0379	+0.0059
iTransformer	96	traffic	+0.0179	-0.0345	-0.0617	-0.1033	+0.2928	+0.6860	+0.1299	-0.0081	+0.0307	+0.0129
iTransformer	96	electricity	-0.0039	-0.0021	-0.0009	+0.0246	+0.8702	+0.8601	+0.2310	+0.0028	+0.0006	+0.0045
iTransformer	96	weather	-0.0050	-0.0046	-0.0068	+0.0172	+0.8429	+0.8857	+0.4671	-0.0025	+0.0011	+0.0061
iTransformer	720	ETTh2	+0.5786	+0.7149	+0.7246	+0.6343	+0.5523	+0.9271	+1.3048	+0.0031	+0.5856	+0.0070
iTransformer	720	ETTm2	+0.7540	+0.5522	+0.5729	+0.6997	+0.4421	+0.8846	+1.4491	-0.0046	+0.7733	+0.0193
iTransformer	720	ETTh1	+0.0776	+0.0918	+0.1121	-0.1972	+0.6314	+0.9444	+0.4295	+0.0458	+0.0810	+0.0034
iTransformer	720	ETTm1	+0.0427	+0.0640	+0.0763	-0.2162	+0.7120	+0.8428	+0.5238	-0.0055	+0.0557	+0.0130
iTransformer	720	traffic	+0.0097	-0.0076	-0.0878	-0.0237	+0.2505	+0.6730	+0.1282	-0.0020	+0.0234	+0.0137
iTransformer	720	electricity	-0.0086	-0.0027	-0.0015	+0.0414	+0.6424	+0.7562	+0.2212	+0.0025	+0.0046	+0.0132
iTransformer	720	weather	-0.0298	-0.0279	-0.0278	+0.0636	+0.7629	+0.7224	+0.4429	-0.0209	+0.0048	+0.0346

Table 5: Forecasting results for prediction horizons $H \in \{96, 192, 336, 720\}$ with input length $L = 96$. The **Mean** scheme is compared against the **Full RevIN** scheme. The "Mean" column is shaded with a gray background. The better result is bolded.

Dataset	H	Metric	SCINet (2022)		PatchTST (2023)		iTransformer (2024)		TiDE (2023)		Crossformer (2023)		RLinear (2023)	
			RevIN	Mean	RevIN	Mean	RevIN	Mean	RevIN	Mean	RevIN	Mean	RevIN	Mean
ETTh1	96	MSE	0.404	**0.396**	0.383	**0.382**	0.388	**0.382**	0.396	**0.395**	0.388	**0.373**	0.386	**0.385**
		MAE	0.410	**0.408**	**0.401**	0.402	**0.405**	0.408	**0.402**	**0.402**	0.404	**0.400**	**0.394**	**0.394**
	192	MSE	0.452	**0.444**	**0.430**	**0.430**	0.441	**0.429**	0.447	**0.446**	0.439	**0.411**	0.438	**0.436**
		MAE	0.437	**0.436**	**0.433**	0.434	0.438	**0.437**	0.433	**0.432**	0.436	**0.426**	**0.424**	**0.424**
	336	MSE	0.492	**0.483**	0.484	**0.459**	0.489	**0.464**	0.491	**0.487**	0.464	**0.450**	0.482	**0.478**
		MAE	0.456	**0.455**	0.467	**0.452**	0.465	**0.459**	0.455	**0.453**	0.450	**0.448**	0.448	**0.446**
	720	MSE	0.502	**0.493**	0.510	**0.491**	0.503	**0.485**	0.493	**0.483**	0.494	**0.478**	0.483	**0.473**
		MAE	**0.477**	0.479	0.497	**0.491**	0.493	**0.489**	0.479	**0.472**	**0.484**	0.488	0.471	**0.464**
ETTh2	96	MSE	0.330	**0.306**	0.297	**0.283**	0.295	**0.293**	**0.298**	**0.298**	0.290	**0.282**	**0.286**	**0.286**
		MAE	0.367	**0.364**	0.348	**0.336**	**0.347**	0.351	**0.347**	**0.347**	0.345	**0.341**	**0.335**	0.336
	192	MSE	0.423	**0.407**	0.374	**0.363**	0.385	**0.383**	**0.391**	**0.391**	0.394	**0.377**	**0.380**	**0.380**
		MAE	**0.420**	0.422	0.392	**0.389**	**0.400**	**0.400**	**0.401**	**0.401**	**0.406**	**0.406**	**0.391**	**0.391**
	336	MSE	0.453	**0.451**	**0.422**	0.424	**0.428**	0.431	0.429	**0.428**	**0.438**	0.458	**0.417**	**0.417**
		MAE	**0.447**	0.458	**0.433**	0.436	**0.435**	0.442	**0.436**	**0.436**	**0.439**	0.464	**0.426**	0.427
	720	MSE	**0.446**	0.455	0.440	**0.421**	0.430	**0.429**	**0.431**	0.432	**0.461**	0.476	**0.420**	**0.420**
		MAE	**0.456**	0.473	0.454	**0.452**	**0.447**	0.450	**0.447**	0.450	**0.462**	0.479	**0.439**	0.442
ETTm1	96	MSE	**0.332**	**0.332**	**0.327**	0.335	**0.338**	0.339	0.354	**0.352**	**0.312**	0.326	0.350	**0.348**
		MAE	**0.371**	0.378	**0.365**	0.370	**0.372**	0.377	0.375	**0.372**	**0.352**	0.359	0.372	**0.369**
	192	MSE	0.384	**0.365**	**0.366**	0.377	**0.380**	0.385	0.391	**0.390**	0.383	**0.365**	0.390	**0.388**
		MAE	**0.398**	**0.398**	**0.388**	0.396	**0.394**	0.408	0.393	**0.391**	0.390	**0.384**	0.391	**0.390**
	336	MSE	0.432	**0.398**	**0.398**	0.406	0.427	**0.414**	0.425	**0.424**	0.438	**0.411**	0.423	**0.422**
		MAE	0.425	**0.418**	**0.409**	0.413	**0.422**	0.427	0.415	**0.413**	0.420	**0.417**	0.413	**0.412**
	720	MSE	0.502	**0.451**	**0.457**	0.458	0.493	**0.465**	0.489	**0.486**	0.474	**0.440**	0.486	**0.483**
		MAE	0.461	**0.452**	**0.444**	**0.444**	**0.460**	**0.460**	0.450	**0.448**	0.451	**0.443**	0.448	**0.446**
ETTm2	96	MSE	0.184	**0.182**	0.179	**0.175**	0.184	**0.173**	**0.183**	**0.183**	**0.182**	0.190	**0.181**	**0.181**
		MAE	**0.265**	0.273	0.262	**0.259**	0.267	**0.259**	**0.266**	**0.266**	**0.258**	0.270	**0.264**	**0.264**
	192	MSE	**0.249**	**0.249**	**0.243**	0.244	0.250	**0.245**	0.248	**0.247**	0.252	**0.249**	0.246	**0.245**
		MAE	**0.309**	0.315	0.305	**0.301**	0.309	**0.305**	**0.306**	**0.306**	0.311	**0.309**	**0.304**	**0.304**
	336	MSE	**0.310**	0.316	**0.304**	**0.304**	0.313	**0.298**	0.309	**0.308**	**0.308**	0.311	**0.306**	**0.306**
		MAE	**0.347**	0.358	0.345	**0.340**	0.349	**0.341**	0.345	**0.344**	**0.347**	0.352	0.342	**0.341**
	720	MSE	0.416	**0.404**	0.410	**0.398**	**0.413**	0.424	0.410	**0.409**	**0.410**	0.432	**0.407**	**0.407**
		MAE	**0.408**	0.411	0.407	**0.399**	**0.407**	0.412	**0.400**	**0.400**	**0.405**	0.418	**0.398**	**0.398**
Electricity	96	MSE	**0.156**	0.164	0.184	**0.182**	**0.166**	0.167	**0.200**	**0.200**	0.143	**0.140**	**0.198**	**0.198**
		MAE	**0.260**	0.275	0.276	**0.275**	**0.257**	0.260	**0.278**	**0.278**	0.242	**0.241**	**0.275**	**0.275**
	192	MSE	**0.179**	0.185	0.191	**0.188**	**0.179**	0.180	**0.201**	**0.201**	0.164	**0.160**	**0.199**	**0.199**
		MAE	**0.283**	0.294	0.283	**0.282**	**0.268**	0.272	**0.281**	**0.281**	0.260	**0.258**	**0.278**	0.279
	336	MSE	0.204	**0.193**	0.207	**0.203**	0.196	**0.195**	**0.215**	**0.215**	0.185	**0.183**	**0.212**	**0.212**
		MAE	0.305	**0.303**	0.298	**0.297**	**0.285**	0.288	**0.296**	**0.296**	**0.279**	0.282	**0.293**	**0.293**
	720	MSE	0.241	**0.226**	0.248	**0.241**	0.237	**0.232**	-	-	**0.205**	0.206	**0.254**	**0.254**
		MAE	0.335	**0.330**	0.330	**0.329**	**0.319**	0.322	-	-	**0.296**	0.302	**0.326**	0.327
Traffic	96	MSE	**0.518**	0.531	**0.481**	0.559	**0.450**	0.874	-	-	**0.429**	0.550	**0.646**	0.648
		MAE	**0.343**	0.351	**0.312**	0.326	**0.312**	0.554	-	-	**0.291**	0.320	**0.386**	0.390
	192	MSE	**0.536**	0.549	**0.485**	0.566	**0.469**	0.837	-	-	**0.445**	0.536	**0.601**	0.602
		MAE	**0.354**	0.366	**0.314**	0.325	**0.319**	0.533	-	-	**0.280**	0.286	**0.364**	0.367
	336	MSE	**0.556**	0.565	**0.502**	0.582	**0.492**	0.882	-	-	**0.472**	0.530	**0.609**	0.610
		MAE	**0.363**	0.370	**0.322**	0.332	**0.331**	0.548	-	-	**0.287**	**0.287**	**0.367**	0.371
	720	MSE	**0.593**	0.602	**0.537**	0.617	**0.527**	0.917	-	-	**0.502**	0.554	**0.648**	0.649
		MAE	**0.379**	0.389	**0.339**	0.351	**0.353**	0.560	-	-	**0.302**	0.303	**0.386**	0.390
Weather	96	MSE	0.166	**0.157**	0.173	**0.166**	**0.173**	0.175	**0.195**	0.196	0.153	**0.146**	**0.194**	0.195
		MAE	**0.213**	0.225	**0.214**	0.217	**0.213**	0.228	**0.235**	**0.235**	**0.202**	0.209	**0.233**	0.234
	192	MSE	**0.214**	0.220	0.219	**0.213**	0.225	**0.224**	**0.240**	0.241	**0.203**	0.206	**0.240**	0.241
		MAE	**0.256**	0.286	**0.255**	0.259	**0.258**	0.276	**0.270**	0.272	**0.248**	0.271	**0.270**	0.271
	336	MSE	0.282	**0.275**	0.275	**0.271**	**0.282**	0.293	**0.292**	0.294	**0.267**	0.271	**0.291**	0.293
		MAE	**0.304**	0.328	**0.295**	0.304	**0.299**	0.331	**0.307**	0.308	**0.293**	0.323	**0.306**	0.307
	720	MSE	0.364	**0.358**	**0.352**	0.354	**0.358**	0.375	**0.365**	0.367	**0.342**	0.352	**0.364**	0.366
		MAE	**0.356**	0.385	**0.346**	0.361	**0.350**	0.385	**0.353**	0.356	**0.342**	0.374	**0.353**	0.355

Note: '-' marks an OOM or missing run.