FORECAST REVISIONS SHOULD BE NEWS: INSTRUMENTED MARTINGALE REGULARIZATION FOR MULTI-HORIZON FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasts are usually scored one issue date at a time, yet deployed systems repeatedly revise the same target. If yesterday's information predicts today's revision, the earlier forecast left usable signal on the table; revisions caused by genuinely new information are desirable, so penalizing their magnitude solves a different problem. We formalize *revision efficiency* through the martingale restriction of conditional-mean forecasts and turn it into an instrumented objective for direct multi-horizon models. A naive squared minibatch moment contains a positive diagonal term proportional to revision variance, silently rewarding stability. Moment-U removes this self-product with an off-diagonal U-statistic. We prove population compatibility, exact conditional unbiasedness for shuffled finite-series batches, and an $O(1/n)$ bias bound under summable temporal covariance. In a heteroskedastic-news control, Moment-U preserves legitimate news while diagonal and stability losses shrink it. A fixed validation rule selects Moment-U on five of six DLinear datasets with negligible MSE changes. Across five-seed PatchTST/iTransformer studies, a frozen richer kernel score falls 6.6–40.3% in all four backbone–dataset cells and is below validation-budget-matched stability in each; MSE shows one small cost, two unresolved changes, and one gain. Including DLinear, a frozen falsification gate passes in three of four headline cells. We reduce predictable revisions rather than claim complete conditional calibration.

1 INTRODUCTION

Forecasts do not arrive once: a rolling system may predict Friday's electricity demand on Monday and revise it on Tuesday and Wednesday. Each forecast can score well alone while the sequence exposes a failure. If Monday's information implies that Tuesday's forecast will move upward on average, Monday's forecast left usable signal on the table; if a new temperature observation reveals a heat wave, a large revision is appropriate. The question is not whether forecasts change, but whether their changes are news.

Standard multi-horizon benchmarks miss this failure because they score each origin–horizon pair separately. Recent work studies forecast *stability*, measuring or penalizing changes across issue dates (Van Belle et al., 2023; Potosnak et al., 2026; Ma et al., 2026). This is useful when revisions are costly, but it is not information efficiency: a constant forecast is perfectly stable, while an optimal forecast changes after informative observations.

The classical fixed-event efficiency principle makes this distinction precise (Nordhaus, 1987; Chang et al., 2013). Conditional-mean forecasts of one target form a martingale: the next revision may be large or heteroskedastic, but its conditional mean given the old information is zero. Thus old information should not predict the next revision in conditional mean.

Turning this principle into a loss creates a hidden obstacle. Squaring a minibatch moment includes each example's self-product, which penalizes revision variance even when revisions are pure news and silently recovers a stability objective. Our off-diagonal construction removes this shortcut.

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.

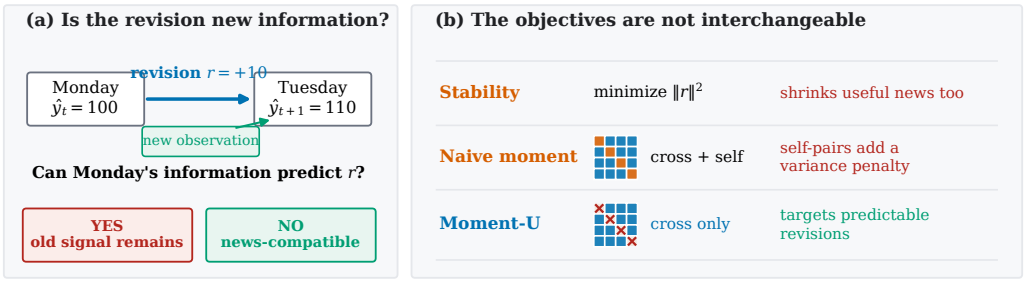


Figure 1: **Revision efficiency is not stability.** Old information should not predict the next update. Stability and naive minibatch moments also suppress useful news; Moment-U removes self-pairs and targets cross-example predictability.

Our contributions are:

1. **Problem and diagnostics.** We formulate revision efficiency for the overlapping forecasts produced by ordinary multi-horizon models and give an exact fixed-target alignment protocol. The resulting audit separates point accuracy, revision size, outcome-aware utility, and old-information predictability.
2. **Objective and estimator.** We introduce an instrumented conditional-moment penalty that targets only predictable revisions and preserves a correctly specified Bayes forecast. We identify the diagonal trap in its naive minibatch implementation, remove it with Moment-U, and prove the estimator's finite-series and dependent-process guarantees.
3. **Evidence and boundaries.** Synthetic data verify that Moment-U preserves genuine news. Across DLinear, PatchTST, and iTransformer, it repeatedly reduces both the trained moments and a frozen richer kernel diagnostic without mechanically smoothing the forecast path. Horizon transfer, strong stability baselines, an Exchange rejection, a mixed-sign tree probe, and remaining kernel-test rejections delimit the claim: repeatable partial improvement, not universal dominance or complete calibration.

2 Related Work

Direct long-horizon forecasting. Strong systems span linear models (Zeng et al., 2023), patch and channel-token Transformers (Nie et al., 2023; Liu et al., 2024), and multiscale mixers (Wang et al., 2024). Because normalized linear variants can have equivalent function classes and closed-form fits remain competitive (Toner & Darlow, 2024), we study a training-time plug-in across linear and nonlinear backbones rather than propose another architecture.

Conditional-moment learning and testing. Conditional restrictions can be learned with adversarial test functions (Lewis & Syrgkanis, 2018; Dikkala et al., 2020) and tested through RKHS maximum moments (Muandet et al., 2020). We instead train with a finite interpretable instrument bank, then avoid training-score circularity with a frozen raw-context kernel audit. The forecasting-specific pieces are fixed-target alignment, the news–stability distinction, and the minibatch diagonal correction.

Forecast stability and fixed-event revisions. Classical efficiency tests ask whether past information predicts forecast errors or revisions (Nordhaus, 1987; Chang et al., 2013; Giacomini & White, 2006). Forecast-evolution models instead specify innovation covariances for planning (Heath & Jackson, 1994); probability paths, threshold martingales, and measure-valued martingales model or assess evolving probability forecasts (Lin et al., 2021; Foster & Stine, 2021; Wilkinson & Ferro, 2026). We train point forecasters against predictable realized revisions. Stability work penalizes N-BEATS updates (Van Belle et al., 2023), jointly emits creation dates and measures excess volatility (Potosnak et al., 2026), or scores accuracy and coherence together (Ma et al., 2026). Look-ahead ensembles instead target within-vector consistency (Kim & Yoon, 2026). These are strong controls, but revision size and conditional-mean predictability are different estimands. Indeed, regularization can induce

efficiency violations (Frank et al., 2023), and conditional revision scale is a complementary second-moment target (Foo & Chang, 2026).

Closest work: martingale consistency. Gögl et al. (2026) apply the same tower property to stochastic coarse and refined views in self-supervised learning, using two conditionally independent imputations to estimate a pointwise conditional squared norm. We observe only one chronologically realized refinement per old origin, so use GMM-style instruments and pair distinct forecasting examples. Our endpoint is supervised fixed-target revision efficiency, audited against stability and without suppressing forecast news.

3 Forecasts across information sets

Let $Y_\tau \in \mathbb{R}^C$ be a fixed target at absolute time τ, and let $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ contain all information available when the forecast is issued at time $t < \tau$. The squared-loss Bayes forecast is

$$m_{t,\tau} = \mathbb{E}[Y_\tau \mid \mathcal{F}_t]. \tag{1}$$

A practical forecaster receives a lookback $X_t \in \mathbb{R}^{L\times C}$ and produces $f_\theta(X_t) = (\hat{y}_{t,1}, \dots, \hat{y}_{t,H})$. For $h \in \{2, \dots, H\}$, the aligned fixed-target revision is

$$r^\theta_{t,h} = \underbrace{f_{\theta,h-1}(X_{t+1})}_{\text{new forecast for } Y_{t+h}} - \underbrace{f_{\theta,h}(X_t)}_{\text{old forecast for } Y_{t+h}}. \tag{2}$$

This indexing is easy to get wrong: comparing the same horizon index across origins compares two different targets.

Proposition 1 (News property)**.** *If Y_τ is integrable and $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, then*

$$\mathbb{E}[m_{t+1,\tau} - m_{t,\tau} \mid \mathcal{F}_t] = 0. \tag{3}$$

If Y_τ is square-integrable, then

$$\mathbb{E}\|Y_\tau - m_{t,\tau}\|^2 - \mathbb{E}\|Y_\tau - m_{t+1,\tau}\|^2 = \mathbb{E}\|m_{t+1,\tau} - m_{t,\tau}\|^2. \tag{4}$$

Equation 3 says revisions are *news*, not that they are small. The finite-window inputs X_t and X_{t+1} need not themselves generate nested sigma-fields because a rolling window drops its oldest value. Our normative information sets are the full operational histories $\mathcal{F}_t$; functions of X_t are merely available test instruments. Exact Fisher consistency for a finite-window backbone therefore additionally requires that its input is sufficient for the full-history Bayes forecast. Without that condition, our moments remain necessary efficiency diagnostics but need not all be jointly attainable.

For arbitrary aligned forecasts $a = \hat{y}_{\tau|t}$ and $b = \hat{y}_{\tau|t+1}$, let $r = b - a$, $e_a = Y_\tau - a$, and $e_b = Y_\tau - b$. The identity

$$\mathbb{E}\|e_a\|^2 - \mathbb{E}\|e_b\|^2 = \mathbb{E}\|r\|^2 + 2\mathbb{E}[e_b^\top r] \tag{5}$$

motivates the information-gain ratio $\mathrm{IGR} = (\mathbb{E}\|e_a\|^2 - \mathbb{E}\|e_b\|^2)/(\mathbb{E}\|r\|^2 + \epsilon)$. An oracle conditional-mean sequence has IGR one. IGR is an outcome-aware diagnostic, not a replacement for MSE; it is unstable when revision energy is near zero.

4 Instrumented revision training

4.1 Conditional moments

For any bounded old-information-measurable instrument $\phi_j(\mathcal{F}_t) \in \mathbb{R}$, Proposition 1 implies

$$g_{j,h,c}(\theta) = \mathbb{E}\big[\phi_j(\mathcal{F}_t) r^\theta_{t,h,c}\big] = 0. \tag{6}$$

With K instruments we penalize the finite projection

$$\mathcal{J}(\theta) = \frac{1}{(H-1)CK} \sum_{h,c,j} g_{j,h,c}(\theta)^2, \qquad \mathcal{L}(\theta) = \mathcal{R}_{\text{forecast}}(\theta) + \lambda \mathcal{J}(\theta). \tag{7}$$

This is a GMM-style criterion (Hansen, 1982). A finite instrument bank cannot certify the entire conditional restriction; it tests and trains only a projection. Our transparent default bank contains an intercept, bounded last/mean/standard-deviation/slope/difference summaries, four segment means, normalized horizon, and the old forecast through stop-gradient. The latter makes training a semi-gradient method: it preserves the criterion's forward value and zero set while preventing instrument collapse. Every feature is example-wise and bounded with $\tanh$; no batch normalization is used.

Proposition 2 (Population compatibility). *Assume squared loss, square integrability, and that the model class contains the full-information Bayes forecast at every aligned origin and horizon. If every instrument is $\mathcal{F}_t$-measurable, the Bayes forecast is a population minimizer of Equation 7 for every $\lambda \geq 0$. If the squared-loss minimizer is unique almost surely, it remains the unique minimizer.*

By contrast, a stability objective adds $\lambda \mathbb{E}\|r^{\theta}_{t,h}\|^2$. At the Bayes forecast this term equals the variance of the news increment and is positive whenever new data change the conditional mean. It therefore lacks the zero-penalty compatibility above and generally shifts the population target toward smaller updates. This does not make stability wrong—revision costs may be operationally real—but it establishes that stability and efficiency are distinct estimands.

4.2 The diagonal trap

Let $z_i \in \mathbb{R}^d$ concatenate the instrument–revision products in one minibatch example, divided by $\sqrt{d}$ to match Equation 7, with $\mu = \mathbb{E}z_i$ and covariance Σ. The obvious estimator squares the sample moment:

$$\widehat{\mathcal{J}}_{\text{diag}} = \left\| \frac{1}{B} \sum_{i=1}^{B} z_i \right\|^2 . \tag{8}$$

Proposition 3 (Implicit revision shrinkage). *For IID examples with finite second moments,*

$$\boxed{\mathbb{E}\widehat{\mathcal{J}}_{\text{diag}} = \underbrace{\|\mu\|^2}_{\text{target moment}} + \underbrace{\frac{1}{B}\operatorname{tr}(\Sigma)}_{\text{unwanted self-term}} .} \tag{9}$$

The off-diagonal estimator

$$\widehat{\mathcal{J}}_U = \frac{1}{B(B-1)} \sum_{i \neq \ell} z_i^\top z_\ell = \frac{\|\sum_i z_i\|^2 - \sum_i \|z_i\|^2}{B(B-1)} \tag{10}$$

satisfies $\mathbb{E}\widehat{\mathcal{J}}_U = \|\mu\|^2$. Moreover, on every batch, $\widehat{\mathcal{J}}_{\text{diag}} - \widehat{\mathcal{J}}_U = \operatorname{tr}(S_z)/B \geq 0$, where S_z is the unbiased sample covariance.

Proposition 4 (Shuffled dependent origins). *Fix $z_{1:n}$ and sample a batch of $B \geq 2$ distinct indices uniformly without replacement. Then the batch estimator in Equation 10 is conditionally unbiased for*

$$\widehat{\mathcal{J}}_{n,U} = \frac{1}{n(n-1)} \sum_{s \neq t} z_s^\top z_t. \tag{11}$$

If $\{z_t\}$ is stationary with mean μ and scalar autocovariance $\gamma_k = \mathbb{E}[(z_0 - \mu)^\top (z_k - \mu)]$, then

$$\mathbb{E}\widehat{\mathcal{J}}_{n,U} - \|\mu\|^2 = \frac{2}{n(n-1)} \sum_{k=1}^{n-1} (n-k)\gamma_k. \tag{12}$$

Consequently, if $\sum_{k \geq 1} |\gamma_k| \leq G < \infty$, the absolute bias is at most $2G/(n-1)$.

Because z_i contains r_i, the extra term in Equation 9 rewards reducing all revision variation, including useful news. Equation 10 is a standard U-statistic construction (Hoeffding, 1948). It can be negative on a minibatch; unbiasedness and nonnegativity cannot both be expected from this construction. We therefore tune λ on validation data and expose high-weight failures. Proposition 4 matches our shuffled loader: temporal dependence does not bias a minibatch relative to the full empirical off-diagonal objective, while its population bias is controlled by the series' cumulative temporal

covariance. This expectation bound does not replace dependent-data uncertainty quantification; we use chronological evaluation and block resampling below.

The method requires forecasts at two consecutive origins during training and $O(BHCK)$ additional arithmetic. On an idle RTX 5090, complete Moment-U training-step overhead at the audited batch sizes was 1% for PatchTST, 8% for iTransformer, and 65% for the much lighter DLinear; only DLinear added measurable peak allocation (1.1 MiB). These hardware-specific measurements are detailed in Appendix G. The method changes neither the backbone nor inference: a deployed forecast still needs one ordinary forward pass.

5 Experiments

5.1 Questions and falsification criteria

The study is organized around four questions: (Q1) do well-tuned modern forecasters exhibit out-of-sample revision predictability; (Q2) does the proposed objective reduce that predictability beyond a validation-budget-matched stability loss; (Q3) are useful revisions retained; and (Q4) when does the method fail? Before the final audit we froze a falsification rule: Moment-U must lower an out-of-bank kernel conditional-moment score in at least three of four five-seed headline cells, with at least four paired seeds improving in each counted cell and no cell worsening by over 10%.

Orthogonal coverage design. We vary one generalization axis at a time: dataset breadth uses DLinear on ETTh1/2, ETTm1/2, Weather, and Exchange at $L = 336, H = 96$; horizon transfer freezes DLinear/ETTh1 and sweeps $H \in \{96, 192, 336, 720\}$; backbone transfer crosses PatchTST/iTransformer with ETTh1/ETTh2 at $L = 336, H = 96$. The first two slices use three seeds. Four frozen-confirmation cells use five; PatchTST/ETTh2 was extended from its recorded three-seed blind result with unchanged weights and all runs retained. No benchmark cell is a single run. This design tests each axis, not unobserved interactions; Appendix H maps claims, seed counts, and uncertainty.

Baselines and ablations. Every backbone is trained with ordinary ERM, raw squared-revision stability, and Equation 10; the biased moment estimator is isolated in the synthetic mechanism. Post-processing controls include validation-tuned fixed-target exponential smoothing and the externally fixed $\alpha = .9$ Forking-Sequences inference ensemble. The latter is compatible with stored window-model grids; we test the distinct all-origin training graph with a dual-path causal CNN rather than mislabel it as the same fixed-window backbone. Full weight curves retain rejected high-weight failures. Alignment/unit tests, two frozen out-of-bank audits, and matched training-compute controls target the main implementation confounds.

Metrics and inference. We report MSE/MAE by horizon; revision RMS and sFPC; scaled excess volatility and IGR; successive fixed-target revision correlation; a forecast-error-on-revision slope (Coibion & Gorodnichenko, 2015); and held-out old-information probes. The local probe uses the summary features optimized by the method. A deliberately different audit probe uses sparse raw anchors from every channel and ExtraTrees. Our primary out-of-bank audit is a scale-normalized maximum- moment V-statistic using a frozen mixture of global and additive RBF kernels on raw interior context anchors. It fits preprocessing on the first 40% of origins and evaluates on the final 50% after a one-horizon embargo. Test origins overlap, so MSE intervals use paired moving-block resampling (Künsch, 1989); kernel-null sensitivity uses block-wild multipliers, following the dependent-kernel bootstrap principle of Chwialkowski et al. (2014) but reported as a sensitivity audit rather than exact finite-sample calibration. Hyperparameters and checkpoints are selected on validation data only; no test diagnostic selects a forecast.

5.2 Synthetic news and estimator audit

We first isolate the estimand without a sequence-model confound. Draw X, η, ϵ independently with standard normal X, η and $\epsilon \sim \mathcal{N}(0, 0.25^2)$. Let $U = \sigma(X)\eta$, where $\sigma(X) = 0.35 + 0.55\,\mathrm{sigmoid}(X)$, and set $Y = X + U + \epsilon$. The old and new oracle forecasts are X and $X + U$. Thus U is genuine news: $\mathbb{E}[U \mid X] = 0$, although its conditional variance is predictable from X.

At the oracle forecasts, Figure 2 estimates each loss over 50,000 independent minibatches. The ordinary squared sample moment is positive and follows the theoretical $1/B$ self term. The off-diagonal estimate is statistically near zero for every batch size, both when news is homoskedastic and when its variance is predictable. This is the precise failure that the U-statistic corrects; it should not erase conditional heteroskedasticity.

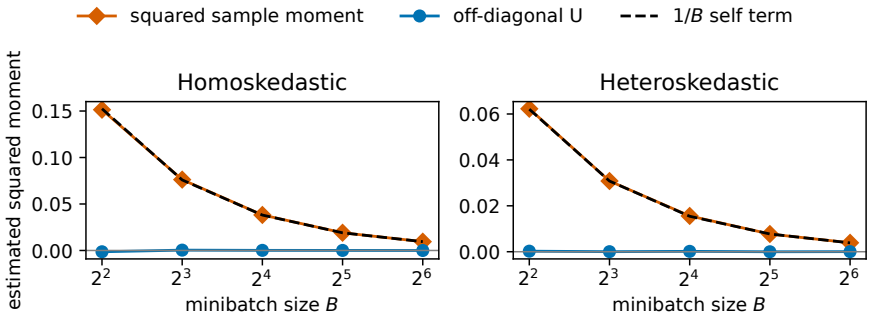


Figure 2: Monte Carlo expectation of the moment estimators when the oracle revision is pure news. The diagonal estimator follows its $1/B$ self-product term; the off-diagonal estimator remains centered at zero.

We then fit the transparent paired forecaster $\hat{y}_{\text{old}} = a_o X$ and $\hat{y}_{\text{new}} = a_n X + bU$. Here $a_n - a_o$ is predictable drift and b is the fraction of genuine news retained. Table 1 averages 100 independent training sets, each containing 512 minibatches of size eight; all regularizer weights are one and evaluation uses a shared 200,000-example test sample. Moment-U preserves $b \simeq 1$ and ERM-level MSE. The diagonal estimator shrinks b to 0.867, while stability gives the population value $b = 1/(1 + 2\lambda) = 1/3$ and substantially worsens MSE.

Table 1: Heteroskedastic-news audit (mean $\pm$ Monte Carlo s.e., 100 fits). News loading $b = 1$ is ideal. Smaller revision RMS can mean that useful news was discarded.

Objective	Test MSE ↓	News loading $b \to 1$	Revision RMS
ERM	.26561±.00001	1.00072±.00065	.63801±.00042
Stability	.35577±.00006	.33357±.00022	.21267±.00014
Moment-diagonal	.26917±.00004	.86735±.00068	.55297±.00043
Moment-U	.26562±.00001	.99981±.00088	.63743±.00056

This experiment also exposes what the moment restriction cannot do alone. A deliberately smoothed forecast $X + 0.3U$ still has zero predictable revision mean, despite worse MSE and IGR 5.66 rather than one. Consequently, revision moments are necessary rather than sufficient for a conditional-mean forecast; the supervised risk and outcome-aware diagnostics in Equation 5 are essential. In this correctly specified toy problem Moment-U does not improve ERM's already-zero population moment—its role is to avoid corrupting it.

5.3 Baseline-fidelity DLinear study

We audit ETTh1/2, ETTm1/2, Weather, and Exchange with $L = 336, H = 96$ and canonical chronological splits (Zhou et al., 2021). DLinear uses its official random initialization and moving-average width 25, Adam, at most 10 epochs, and validation-MSE early stopping. Dataset-specific batch sizes and learning rates are selected by three-seed validation MSE; the resulting ERM test MSEs are $0.380, 0.289, 0.302, 0.166, 0.175, 0.089$, closely reproducing five of the six published DLinear references $(0.375, 0.289, 0.299, 0.167, 0.176, 0.081)$.

For each regularizer, seed 0 screens the full weight curve $\{0.03, 0.1, 0.3, 1, 3, 10, 30\}$. We rerun both the best validation-target point and the lowest validation-MSE point on seeds 1 and 2, then minimize the three-seed mean validation target subject to mean validation MSE being within 1% of ERM and

the target improving over ERM. A failed constraint returns ERM. No test value enters either stage. Moment RMS is computed with the exact bounded training instruments: moments remain separate for each horizon and channel, average only over origins, and are normalized by the corresponding revision-cell RMS before the final aggregation. The probe is a temporally held-out ridge model and is averaged across eight fixed horizons.

Table 2: DLinear $L = 336, H = 96$ results (three-seed means). ΔMSE is relative to paired ERM; arrows show ERM→method. Moment-U is rejected on Exchange by the validation rule. Revision RMS is a magnitude, not an efficiency score.

Data	ERM MSE	λ_U	$\Delta\text{MSE}_U\ 10^5$	Moment RMS_U	λ_S	$\Delta\text{MSE}_S\ 10^3$	Rev. RMS_S	Probe R^2_S
ETTh1	.3804	.3	$-.80$	$.1508 \to .1497$	3	-3.97	$.1305 \to .0499$	$.0197 \to .1072$
ETTh2	.2889	3	$+1.12$	$.1123 \to .1057$	1	-2.46	$.1149 \to .0653$	$.0690 \to .0547$
ETTm1	.3022	10	-1.44	$.1875 \to .1684$	.3	$-.41$	$.0717 \to .0543$	$.0387 \to .0595$
ETTm2	.1663	1	-2.81	$.1722 \to .1668$	30	$+4.08$	$.0704 \to .0234$	$.0406 \to .5011$
Weather	.1745	10	$-.86$	$.2087 \to .1635$	3	$+2.91$	$.0309 \to .0171$	$.0984 \to .2920$
Exchange	.0893	–	–	$.3095 \to$–	–	–	$.0452 \to$–	$.2235 \to$–

Moment-U is accepted on five of six datasets. Its test moment RMS decreases by $0.7, 5.9, 10.2, 3.2,$ and 21.7%, respectively, while every absolute MSE change is below 2.9×10^{-5} and revision RMS changes by less than 0.7%. The independent local probe falls on ETTm1, ETTm2, and Weather and is effectively unchanged on ETTh1/2. Exchange is a necessary failure: neither screened Moment-U candidate improves the three-seed ERM validation moment, so the rule makes no method claim there. In a confirmation frozen before seeds 3–4, the ETTh1 anchor's five-seed moment changes $.1518 \to .1507$ (-0.7%), while its MSE difference remains -8.9×10^{-6} (95% hierarchical block CI $[-1.30, -.58] \times 10^{-5}$).

Stability reduces revision RMS by 24–67% in every accepted cell. It also improves test MSE on ETTh1/2 and ETTm1, showing that revision shrinkage can be useful regularization. However, its local probe worsens on four of five datasets—dramatically on ETTm2—and its exact moment worsens on four of five; test MSE also degrades on ETTm2 and Weather. Thus lower revision magnitude neither implies nor uniformly prevents efficiency gains. The completed slice supports the narrower claim that the instrumented objective changes the targeted conditional mean without achieving its effect by smoothing. It does *not* establish cross-backbone dominance.

A validation-tuned fixed-target exponential smoother gives an even sharper control. Under the same 1% validation-MSE budget it is rejected on ETTh1/2 and ETTm1, and selects $\alpha = 0.1, 0.3, 0.7$ on ETTm2, Weather, and Exchange. The selected smoothers reduce test revision RMS by $84\%, 54\%,$ and 26%, respectively, but the held-out probe rises in all three ($.041 \to .453$, $.098 \to .375$, and $.224 \to .287$). Test moment RMS improves only on Weather; test MSE is higher in all three, although Exchange changes by only 6.4×10^{-5}. Mechanical temporal averaging can therefore meet a raw stability target while leaving, or amplifying, a predictable direction in the residual updates.

The externally fixed $\alpha = .9$ Forking-Sequences inference ensemble (Potosnak et al., 2026) lowers both volatility metrics in all nine cells and improves MSE in eight, but improves the exact moment and local probe in only two cells each. A separate dual-path causal-CNN control holds examples and updates fixed: all-origin Forking and window execution select the same learning rate and both obtain test MSE $.4184 \pm .0088$; their paired difference is -7.0×10^{-6} (95% block-bootstrap CI $[-2.26, 0.36] \times 10^{-5}$). Forking is nevertheless $5.40\times$ faster per epoch and cuts peak training allocation by 99.2%. Conventional batch-32 training is more accurate ($.4051 \pm .0008$) but uses 257 rather than 33 updates and takes $8.8\times$ longer per epoch. Thus the controlled training result supports computational reuse, not a statistical gain; Appendix F reports both audits.

Appendix E plots all accepted interventions on common test axes. Full weight curves retain high-weight Moment-U failures and the ETTm2 stability collapse; these motivate the shuffled-origin analysis and frozen audit below.

5.4 FROZEN HORIZON TRANSFER ON ETTH1

Before running any longer-horizon cell, we froze the H96 seed-0 validation-target choices $\lambda_U = .3$ and $\lambda_S = 3$, as well as the DLinear optimizer configuration. We then ran every method at $H \in \{192, 336, 720\}$ with seeds 0–2 and no horizon-specific selection. All six transferred method–horizon cells remain within the fixed 1% validation-MSE budget and improve their respective validation target.

Table 3: DLinear ETTh1 frozen-weight transfer (three-seed means). Arrows denote ERM→method; no higher-horizon value is used for tuning.

H	ERM MSE	$\Delta\text{MSE}_U\ 10^5$	Moment RMS_U	$\Delta\text{MSE}_S\ 10^3$	Rev. RMS_S	Moment RMS_S	Probe R^2_S
96	.3804	$-.80$	.1508 → .1497	-3.97	.1305 → .0499	.1508 → .1930	.0197 → .1072
192	.4182	$-.45$	.1543 → .1527	-4.15	.1259 → .0468	.1543 → .1987	.0261 → .1060
336	.4499	$-.72$	.1488 → .1470	-5.32	.1233 → .0437	.1488 → .2078	.0362 → .1753
720	.5005	-2.83	.1341 → .1329	-9.30	.1425 → .0594	.1341 → .1802	.0413 → .0594

Across the three transferred horizons, Moment-U lowers the exact test moment RMS by $1.1, 1.2$, and 0.9%, respectively, while changing test MSE by at most 2.9×10^{-5} and revision RMS by less than 0.02%. Stability improves test MSE at every horizon and cuts revision RMS by 58–65%, but its exact moment worsens by 28–40% and its local probe rises in every cell. Thus a smoother forecast path can be more accurate yet less efficient under the old- information moment. Frozen optimization is a deliberately stringent transfer test: ERM MSE at $H = 192, 336, 720$ is $.418, .450, .500$ versus published DLinear references $.405, .439, .472$, so the claim here is paired objective transfer rather than state-of-the-art long-horizon accuracy.

5.5 NONLINEAR-BACKBONE CONFIRMATIONS

We repeat the $L = 336, H = 96$ protocol with channel-independent PatchTST and channel-token iTransformer on ETTh1/2. PatchTST's five-seed ERM MSE is .3770 on ETTh1 and .2766 on ETTh2, near the published $.375/.274$ references (Nie et al., 2023). iTransformer's official presets use $L = 96$, so our common-$L = 336$ cells are architecture transfers (Liu et al., 2024). ETTh1 and both iTransformer cells use validation-selected weights frozen before seeds 3–4. For PatchTST/ETTh2, weights transfer unchanged from ETTh1: its three-seed blind result was recorded before a post-result seeds-3–4 precision extension, with all methods retained.

Table 4: Nonlinear $L = 336, H = 96$ confirmations (five-seed mean ± s.d.). MSE intervals hierarchically resample seeds and paired circular origin blocks (length 96; 5,000 draws).

Backbone	Data	Method	Test MSE	ΔMSE [95% CI]	Rev. RMS	Probe R^2	Moment RMS
PatchTST	ETTh1	ERM	.3770±.0016	–	.0911±.0051	.0239±.0069	.0615±.0020
		Moment-U ($\lambda = 30$)	.3773±.0015	+.00032[+.00016, +.00050]	.0898±.0049	.0182±.0072	.0479±.0038
		Stability ($\lambda = .3$)	.3780±.0019	+.00104[+.00038, +.00173]	.0717±.0041	.0250±.0063	.0583±.0054
	ETTh2	ERM	.2766±.0018	–	.0970±.0044	.0349±.0178	.0700±.0067
		Moment-U ($\lambda = 30$)	.2764±.0019	−.00019[−.00052, +.00011]	.0944±.0046	.0241±.0156	.0471±.0055
		Stability ($\lambda = .3$)	.2749±.0018	−.00168[−.00293, −.00029]	.0784±.0034	.0175±.0144	.0584±.0044
iTransformer	ETTh1	ERM	.3984±.0030	–	.1447±.0168	.0563±.0110	.0450±.0026
		Moment-U ($\lambda = 10$)	.3986±.0037	+.00017[−.00045, +.00080]	.1459±.0174	.0535±.0106	.0418±.0018
		Stability ($\lambda = 1$)	.3925±.0021	−.00592[−.00933, −.00271]	.0730±.0022	.0502±.0056	.0452±.0041
iTransformer	ETTh2	ERM	.3171±.0049	–	.1153±.0014	.0495±.0162	.0708±.0026
		Moment-U ($\lambda = 30$)	.3126±.0044	−.00448[−.00601, −.00304]	.1159±.0016	.0447±.0192	.0456±.0014
		Stability ($\lambda = .3$)	.3088±.0024	−.00831[−.01295, −.00401]	.0900±.0011	.0404±.0227	.0648±.0022

On PatchTST, Moment-U lowers the finite-bank moment by 22.1% (ETTh1) and 32.8% (ETTh2 transfer), while revision RMS changes by only -1.5% and -2.7%. Its ETTh1 MSE cost is .00032 (CI excludes zero); the ETTh2 change is unresolved. Stability improves ETTh2 MSE and cuts revision RMS by 19.2%, but lowers the moment by only 16.6%.

On iTransformer, Moment-U lowers the moment by 7.0% on ETTh1 at unresolved MSE change. On ETTh2 it lowers the moment by 35.6% and MSE by .00448 (CI $[-.00601, -.00304]$), with revision

RMS changing by $+0.5\%$. Stability improves MSE and sharply smooths both cells, yet changes the ETTh1 moment by $+0.4\%$ and lowers the ETTh2 moment by only 8.5%. This is a target-specific Pareto distinction, not dominance.

Frozen out-of-bank kernel audit. We froze this richer test before seeing any kernel result. It uses raw interior old-context anchors absent from training, fits centering and bandwidth on the first 40% of origins, then evaluates the last 50% after a one-horizon embargo. Early-period RMS scaling prevents uniform smoothing from passing automatically. The score is a nonnegative maximum-moment V-statistic (Muandet et al., 2020) with a synthetic-preselected mixture of global and additive RBF kernels.

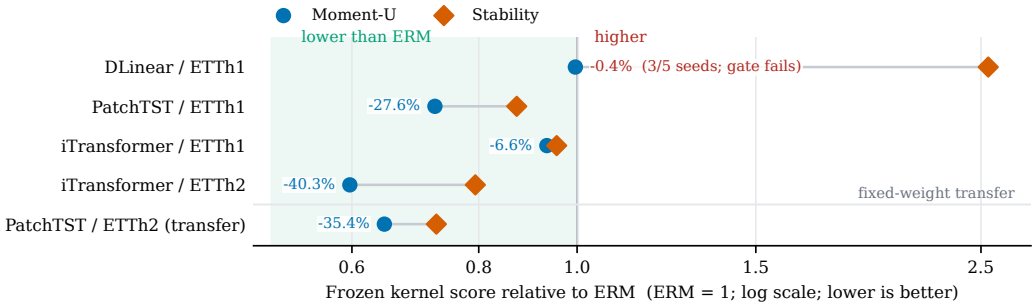


Figure 3: **The frozen richer diagnostic improves across backbones.** Scores are normalized by ERM; lower is better. Moment-U is below matched Stability in all five cells. The frozen gate passes $3/4$: DLinear improves in only $3/5$ seeds. PatchTST/ETTh2 is a fixed-weight extension; Appendix C, Table 5 gives exact means.

The rule passes exactly three of four gate rows; none is over 10% adverse. Moment-U is below Stability in every reported cell, although Stability often has lower MSE, and both kernel components preserve the Moment-U-versus-ERM direction in every gate cell. Block-wild tests still reject the martingale null for most seeds: this is partial transfer, not full calibration. Appendix C reports exact values and sensitivity checks; the frozen ExtraTrees probe in Appendix D retains mixed signs as a complementary failure boundary.

6 Conclusion, Limitations, and Scope

Fixed-target revisions expose an information-use failure that point accuracy and raw stability do not identify. Moment-U targets this failure without mechanically suppressing news, and the diagonal correction is essential in both theory and the synthetic control. Across three backbone families, the frozen rich audit shows repeatable but incomplete improvement: revision efficiency is a complementary forecasting criterion, not a replacement for MSE or operational stability costs.

The boundaries are concrete. A finite instrument bank controls only selected projections of $\mathbb{E}[r_{t,h} \mid \mathcal{F}_t]$, and the richer audit still rejects the null in most cells. Fixed-window inputs are not nested information sets, so exact compatibility needs sufficiency or correct specification. The population bias bound assumes stationarity and summable autocovariance; block- wild p-values are sensitivity checks. Revision efficiency is relative to squared-loss conditional means and the recorded information set: it neither identifies causal "news" nor removes revision costs. Finally, coverage is orthogonal rather than factorial: dataset breadth fixes DLinear and $H = 96$; horizon transfer fixes DLinear/ETTh1; and backbone transfer fixes ETTh1/2 and $H = 96$. The all-origin control is a small causal CNN rather than MQForecaster. These slices test transfer along three separate axes, but do not establish interactions across their full Cartesian product.

AI USE STATEMENT

Generative AI tools were used only for code implementation and debugging, and for copy-editing and language polishing of the manuscript. The authors take full responsibility for all research content, claims, results, figures, tables, citations, and artifacts.

ETHICS STATEMENT

The current study uses public aggregate time-series benchmarks and synthetic data, with no human subjects or personally identifiable information. Forecast revision objectives can nevertheless affect high-stakes operational decisions. Efficiency diagnostics should therefore complement, not replace, domain-specific cost, safety, calibration, and fairness evaluation.

REPRODUCIBILITY STATEMENT

Each run produced by the supplementary code records exact configurations, seeds, environment metadata, checkpoints, absolute forecast origins, contexts, predictions, and targets. Unit tests verify fixed-target alignment, the diagonal/U-statistic identity, and exact agreement between the training and evaluation instrument moments. Synthetic positive, nonlinear-drift, null, and scale-invariance tests audit the frozen probes and kernel score. Appendix A gives proofs, and Appendix B specifies data splits, metrics, hyperparameter grids, and selection and inference protocols. Every numerical table has a machine-readable source summary regenerated from stored predictions and cross-checked during the submission audit.

REFERENCES

Chia-Lin Chang, Bert de Bruijn, Philip Hans Franses, and Michael McAleer. Analyzing fixed-event forecast revisions. *International Journal of Forecasting*, 29(4):622–627, 2013. doi: 10.1016/j.ijforecast.2013.04.002.

Kacper Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://papers.neurips.cc/paper/5452-a-wild-bootstrap-for-degenerate-kernel-tests.

Olivier Coibion and Yuriy Gorodnichenko. Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–2678, 2015. doi: 10.1257/aer.20110306.

Nishanth Dikkala, Greg Lewis, Lester W. Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8fcd9e5482a62a5fa130468f4cf641ef-Abstract.html.

Hui-Mean Foo and Yuan-chin Ivan Chang. Estimating the conditional forecast-revision scale in sequential models: Local-smoothing limits, matched models, and cost–accuracy trade-offs. *arXiv preprint arXiv:2608.03163*, 2026. URL https://arxiv.org/abs/2608.03163.

Dean P. Foster and Robert A. Stine. Threshold martingales and the evolution of forecasts. *arXiv preprint arXiv:2105.06834*, 2021. URL https://arxiv.org/abs/2105.06834.

Murray Z. Frank, Jing Gao, and Keer Yang. Behavioral machine learning? regularization and forecast bias. *arXiv preprint arXiv:2303.16158*, 2023. URL https://arxiv.org/abs/2303.16158.

Raffaella Giacomini and Halbert White. Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578, 2006. doi: 10.1111/j.1468-0262.2006.00718.x.

Moritz Gögl, Hanwen Xing, and Christopher Yau. Martingale-consistent self-supervised learning. *arXiv preprint arXiv:2605.11846*, 2026. URL https://arxiv.org/abs/2605.11846.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. doi: 10.2307/1912775.

David C. Heath and Peter L. Jackson. Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Transactions*, 26(3):17–30, 1994. doi: 10.1080/07408179408966604.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Sunho Kim and Susik Yoon. Back to the future: Look-ahead augmentation and parallel self-refinement for time series forecasting. *arXiv preprint arXiv:2602.02146*, 2026. URL https://arxiv.org/abs/2602.02146. Accepted as a short paper at The Web Conference 2026.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. doi: 10.1214/aos/1176347265.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018. URL https://arxiv.org/abs/1803.07164.

Zhiyuan Jerry Lin, Hao Sheng, and Sharad Goel. Probability paths and the structure of predictions over time. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7f53f8c6c730af6aeb52e66eb74d8507-Abstract.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Chutian Ma, Grigorii Pomazkin, Giacinto Paolo Saggese, and Paul Smith. Beyond accuracy: A stability-aware metric for multi-horizon forecasting. *arXiv preprint arXiv:2601.10863*, 2026. URL https://arxiv.org/abs/2601.10863.

Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 41–50. PMLR, 2020. URL https://proceedings.mlr.press/v124/muandet20a.html.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

William D. Nordhaus. Forecasting efficiency: Concepts and applications. *The Review of Economics and Statistics*, 69(4):667–674, 1987. doi: 10.2307/1935962.

Willa Potosnak, Malcolm Wolff, Mengfei Cao, Ruijun Ma, Tatiana Konstantinova, Dmitry Efimov, Michael W. Mahoney, Boris Oreshkin, and Kin G. Olivares. Forking-sequences: Statistically and computationally efficient multi-horizon forecasting with reduced volatility. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=dXdycy7WCX.

William Toner and Luke Nicholas Darlow. An analysis of linear time series forecasting models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48404–48427. PMLR, 2024. URL https://proceedings.mlr.press/v235/toner24a.html.

Jente Van Belle, Ruben Crevits, and Wouter Verbeke. Improving forecast stability using deep learning. *International Journal of Forecasting*, 39(3):1333–1350, 2023. doi: 10.1016/j.ijforecast.2022.06.007.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Thomas Wilkinson and Christopher Ferro. Calibrated probability forecast sequences and measure-valued martingales. *arXiv preprint arXiv:2606.31621*, 2026. URL https://arxiv.org/abs/2606.31621.

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

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.

A Proofs

A.1 Proof of Proposition 1

By the tower property and $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$,

$$\mathbb{E}[m_{t+1,\tau} \mid \mathcal{F}_t] = \mathbb{E}[\mathbb{E}[Y_\tau \mid \mathcal{F}_{t+1}] \mid \mathcal{F}_t] = \mathbb{E}[Y_\tau \mid \mathcal{F}_t] = m_{t,\tau},$$

which proves Equation 3. For the second statement, write $Y_\tau - m_{t,\tau} = (Y_\tau - m_{t+1,\tau}) + (m_{t+1,\tau} - m_{t,\tau})$. The cross term has expectation zero because $m_{t+1,\tau} - m_{t,\tau}$ is $\mathcal{F}_{t+1}$-measurable and $\mathbb{E}[Y_\tau - m_{t+1,\tau} \mid \mathcal{F}_{t+1}] = 0$. Expanding the square yields Equation 4.

A.2 Proof of Proposition 2

For squared loss, conditional-expectation projection gives

$$\mathcal{R}_{\text{forecast}}(f) = \mathcal{R}_{\text{forecast}}(m) + \mathbb{E}\|f - m\|^2,$$

with the sum understood across origins, horizons, and channels. Proposition 1 and $\mathcal{F}_t$-measurability of each instrument imply $g_{j,h,c}(m) = 0$, so $\mathcal{J}(m) = 0$. Since $\mathcal{J}(f) \geq 0$ at population level,

$$\mathcal{R}_{\text{forecast}}(f) + \lambda \mathcal{J}(f) \geq \mathcal{R}_{\text{forecast}}(m).$$

Thus m is a minimizer. If the squared-loss projection is unique, equality requires $f = m$ almost surely.

A.3 Proof of Proposition 3

For IID z_i,

$$\mathbb{E}\|\bar{z}\|^2 = \|\mathbb{E}\bar{z}\|^2 + \operatorname{tr}\operatorname{Var}(\bar{z}) = \|\mu\|^2 + \operatorname{tr}(\Sigma)/B.$$

For $i \neq \ell$, independence gives $\mathbb{E}[z_i^\top z_\ell] = \mathbb{E}[z_i]^\top \mathbb{E}[z_\ell] = \|\mu\|^2$; averaging all ordered off-diagonal pairs proves unbiasedness of Equation 10. Finally, direct algebra using $\operatorname{tr}(S_z) = (B-1)^{-1} \sum_i \|z_i - \bar{z}\|^2$ gives $\|\bar{z}\|^2 - \widehat{\mathcal{J}}_U = \operatorname{tr}(S_z)/B$.

A.4 Proof of Proposition 4

Condition on $z_{1:n}$. Any ordered pair of distinct indices appears in a uniform size-B batch with probability $B(B-1)/[n(n-1)]$. Multiplying this inclusion probability by the $1/[B(B-1)]$ normalization and summing over pairs gives Equation 11. For the population statement, stationarity gives $\mathbb{E}[z_s^\top z_t] = \|\mu\|^2 + \gamma_{|t-s|}$. There are $2(n-k)$ ordered pairs at lag k; grouping them proves Equation 12. Finally,

$$\left|\mathbb{E}\widehat{\mathcal{J}}_{n,U} - \|\mu\|^2\right| \leq \frac{2}{n(n-1)} \sum_{k=1}^{n-1} (n-k)|\gamma_k| \leq \frac{2G}{n-1}.$$

B EXACT FIRST-SLICE PROTOCOL

Data. We retain every numeric channel in ETTh1/2, ETTm1/2, Weather, and Exchange. A per-channel standardizer is fitted only on the training period. Validation and test targets remain inside their respective chronological splits, although their lookbacks may use preceding observations. For origin t, the implementation asserts that the old targets at horizons $2{:}H$ exactly equal the new targets at horizons $1{:}H-1$.

Optimization. The forecast loss averages MSE at origins t and $t+1$. Instrument-derived old forecasts are detached, and all remaining instruments are deterministic per-example functions. The default does not standardize across a minibatch. Each run stores the validation-selected checkpoint, raw test arrays, its full configuration (with a deterministic hash in the artifact path), Torch/CUDA/device metadata, and runtime.

Selection and inference. Seed 0 screens every weight; both its validation-target and validation-MSE candidates advance to three seeds. Final selection uses the three-seed mean validation quantities and the 1% accuracy budget stated in the main text. This two-stage resource allocation can miss a weight that is noisy on seed 0, but it cannot select on test data. Before any H192/H336/H720 run, the H96 seed-0 target-selected weights were frozen; all higher-horizon cells are reported without reselection.

The MSE interval is hierarchical: seeds are resampled with replacement, then a circular moving-block bootstrap resamples paired per-origin loss differences within each sampled seed. The block length is 96 for the original slices and equals the forecast horizon in the transfer sweep. We use 3,000 draws for three-seed slices. The frozen headline confirmation fixes DLinear ETTh1, PatchTST ETTh1, and iTransformer ETTh1/2 before seeds 3–4, then resamples all five seeds with 5,000 draws; neither new seed enters selection. PatchTST/ETTh2 transfers ETTh1 weights before its first regularized run; seeds 3–4 are explicitly a post-three-seed precision extension and use 9,999 draws. These intervals are not multiplicity-adjusted simultaneous intervals.

C FROZEN KERNEL CONDITIONAL-MOMENT AUDIT

For each stored forecast grid, we form revisions at eight equally spaced horizons. The conditioning vector contains only raw values at up to 16 equally spaced *interior* lookback positions per channel and at most 128 scalars in total; it excludes old forecasts and every training-instrument summary. The first 40% of chronological origins fit feature centering and scaling, RBF bandwidths, and each horizon–channel revision RMS. The final 50% are evaluated after an embargo of H origins. Thus neither outcomes nor evaluation revisions choose a feature or bandwidth.

Let u_i concatenate evaluation revisions divided by their early-period RMS. The reported score is

$$\widehat{V} = \frac{1}{n^2} \sum_{i,j} k(x_i, x_j) \frac{u_i^\top u_j}{\dim(u_i)},$$

with k an equal mixture of a global RBF and the mean of coordinate-wise RBFs. Each bandwidth is one half of its corresponding calibration median distance. An initial global median RBF failed a synthetic sparse-square-drift power test; the mixture was fixed after that synthetic failure and before any benchmark score. We also report the off-diagonal U-statistic and each kernel component, but neither can change the gate.

Table 5: Exact five-seed means underlying Figure 3. Parentheses give the Moment-U change from ERM and number of paired seeds lower. The first four rows are the frozen falsification gate; PatchTST/ETTh2 is the fixed-weight extension.

Backbone	Data	ERM	Moment-U	Stability	U vs. Stability
DLinear	ETTh1	.015391	.015331 (-0.4%; $3/5$)	.039118	-60.8%
PatchTST	ETTh1	.004647	.003366 (-27.6%; $5/5$)	.004052	-16.9%
iTransformer	ETTh1	.005372	.005016 (-6.6%; $4/5$)	.005129	-2.2%
iTransformer	ETTh2	.009638	.005755 (-40.3%; $5/5$)	.007651	-24.8%
PatchTST	ETTh2	.005145	.003323 (-35.4%; $5/5$)	.003739	-11.1%

For dependence sensitivity, one Rademacher multiplier is shared by each contiguous block and the resulting quadratic form is recomputed with 4,999 draws at block lengths 48, 96, and 192. This is a block-wild audit motivated by Chwialkowski et al. (2014), not a claim of exact finite-sample size. Table 6 reports component changes and rejection counts at nominal .05; the persistence of rejection is why the main text claims reduction rather than calibration.

Table 6: Kernel-audit robustness. Global/additive columns are Moment-U changes from ERM. Remaining entries are numbers of rejections among five seeds for blocks 48/96/192.

Backbone	Data	Global	Additive	ERM	Moment-U	Stability
DLinear	ETTh1	-0.4%	-0.4%	5/5/5	5/5/5	5/5/5
PatchTST	ETTh1	-17.2%	-32.4%	5/5/5	5/5/5	5/5/5
iTransformer	ETTh1	-3.0%	-9.4%	5/5/5	5/5/4	5/5/5
iTransformer	ETTh2	-10.8%	-45.0%	5/5/4	4/3/0	5/5/3
PatchTST	ETTh2	-5.2%	-41.2%	5/5/5	5/4/4	5/5/5

D FROZEN INDEPENDENT NONLINEAR PROBE

The primary local probe uses channel-local summaries related to the training instruments. As a deliberately different audit, we freeze an ExtraTrees regressor with 128 trees, minimum leaf size 20, and no feature subsampling. Its only inputs are at most 16 evenly spaced anchors from the old context across all channels and the complete old-forecast vector. For eight fixed horizons, it is fit on the first 40% of origins and evaluated on the final 50%, with at least one full forecast horizon embargo. Hyperparameters are shared by every dataset, backbone, method, and seed; neither this probe nor its test values select a forecast. Before inspecting benchmark results, a synthetic square-drift sensitivity test motivated changing feature subsampling from square-root to all features; a separate independent-noise test guards against a spurious positive result.

Table 7: Frozen nonlinear old-information probe (three-seed mean held-out, variance-weighted R^2 across channels, then averaged over eight horizons). Negative values are retained.

Backbone	Dataset	ERM	Moment-U	Stability
DLinear	ETTh1	.0168	.0169	.0515
DLinear	ETTh2	.0136	.0125	$-.0031$
DLinear	ETTm1	.0300	.0309	.0563
DLinear	ETTm2	$-.0268$	$-.0274$	.0935
DLinear	Weather	.0755	.0774	.1944
PatchTST	ETTh1	.0599	.0584	.0631
iTransformer	ETTh1	.0427	.0455	.0263
iTransformer	ETTh2	$-.0138$	$-.0078$	$-.0147$

Moment-U's absolute paired change is at most .0060; it decreases R^2 in three cells and increases it in five. Stability increases R^2 in five cells, including the large ETTm2 and Weather changes $-.0268 \to .0935$ and $.0755 \to .1944$, but improves DLinear ETTh2 and both iTransformer cells. The complete per-seed, per-horizon and energy-weighted values are stored in the supplementary artifact; the table is not used for model selection.

E ACCURACY–REVISION TRADE-OFF MAP

Figure 4 places every validation-accepted DLinear intervention and fixed-target smoother on common test axes. Moment-U changes revision RMS by less than 0.7% while generally lowering the targeted moment; Stability and smoothing move farther left but often raise the frozen probe. These descriptive curves include every evaluated weight and are not used for reselection.

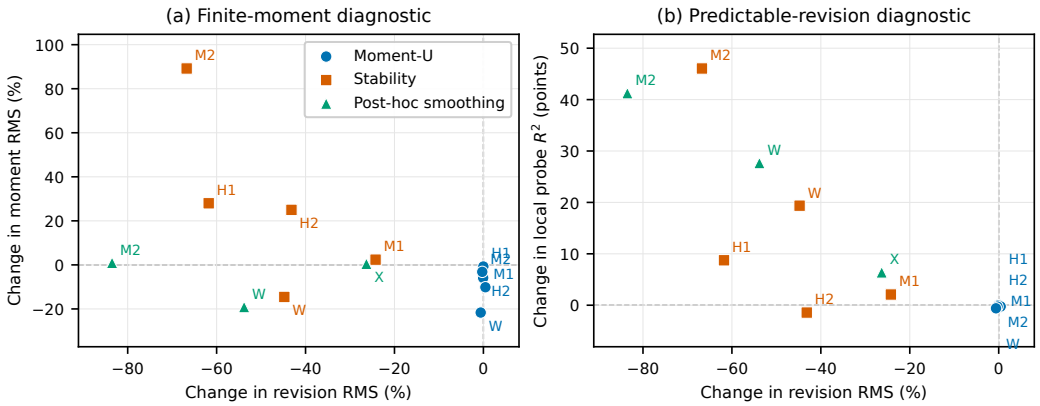


Figure 4: Test trade-offs for DLinear at $L = 336$, $H = 96$. Each point is a three-seed mean; labels give the regularization weight or smoother coefficient. Filled points satisfy the validation accuracy budget and target-improvement rule, while faint points retain rejected high-weight failures. Lower is better on every axis.

F Fixed Forking-Sequences inference control

Forking-Sequences has distinct training and inference components. The training architecture emits the full origin–horizon grid while reusing encoder states, so replacing a fixed-window backbone by that graph is not an architecture-identical loss comparison. Its inference ensemble, however, is explicitly applicable to pretrained window-sampling models. We therefore apply the published $\alpha = .9$ exponential ensemble to each stored ERM grid without tuning. For a fixed target, the recursion is $q_{o,h} = \alpha p_{o,h} + (1 - \alpha) q_{o-1,h+1}$; the oldest available forecast receives the truncated residual weight. Scaled EV below is the point-median specialization of the paper's quantile metric. No value in this audit selects a model or α.

Table 8: Fixed $\alpha = .9$ inference ensemble (three-seed means, ERM→ensemble). Lower is better for every column.

Backbone	Dataset	MSE	sFPC	Scaled EV	Moment RMS	Local probe R^2
DLinear	ETTh1	.3804 → .3792	32.41 → 29.21	.0584 → .0507	.1508 → .1558	.0197 → .0259
DLinear	ETTh2	.2889 → .2880	20.63 → 18.63	.0278 → .0241	.1123 → .1159	.0690 → .0629
DLinear	ETTm1	.3022 → .3021	18.24 → 16.60	.0274 → .0241	.1875 → .1888	.0387 → .0452
DLinear	ETTm2	.1663 → .1660	11.88 → 10.85	.0159 → .0140	.1722 → .1750	.0406 → .0490
DLinear	Weather	.1745 → .1746	7.14 → 6.79	.0062 → .0057	.2087 → .2055	.0984 → .1191
DLinear	Exchange	.0893 → .0893	5.76 → 5.34	.0093 → .0084	.3095 → .3116	.2235 → .2426
PatchTST	ETTh1	.3764 → .3760	22.18 → 19.97	.0373 → .0325	.0616 → .0619	.0227 → .0287
iTransformer	ETTh1	.3984 → .3968	32.15 → 28.98	.0641 → .0555	.0438 → .0438	.0554 → .0572
iTransformer	ETTh2	.3197 → .3188	20.79 → 18.69	.0288 → .0249	.0697 → .0690	.0421 → .0406

The fixed ensemble improves both volatility measures in all nine cells and MSE in eight, closely matching its intended role. It improves the exact moment and local probe in only two cells each. This is not evidence against volatility-aware forecasting; it is evidence that volatility and the conditional first moment are different targets.

F.1 Architecture-native training control

We implement both execution paths in the same 6,416-parameter channel-shared causal residual CNN. Nine width-two dilated blocks have receptive field 336 and a width-one head emits all 96 horizons. The window path recomputes the encoder for each origin; the grid path evaluates each block once and reuses its states across origins. Unit tests verify matched outputs and gradients before training and no future-input leakage. On ETTh1 we compare four prespecified regimes. Seed 0 selects the learning rate for each regime from $\{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$ by validation MSE; seeds 1–2 confirm it, and

the test set never selects anything. Times exclude validation and are three-seed means on an otherwise idle RTX 5090; peak MiB is training-only CUDA allocation. An initial instrumentation pass that allowed the large validation forward into the next epoch's peak was discarded, the counter was reset at every epoch boundary, and the complete experiment rerun.

Table 9: Architecture-native all-origin control on ETTh1, $L = 336, H = 96$. Test MSE is mean$\pm$standard deviation across three seeds.

Regime	LR	Origins/epoch	Updates/epoch	Seconds/epoch	Peak MiB	Test MSE
WS32-update	.003	1,040	33	.074	187.2	$.4103 \pm .0143$
WS32-full	.001	8,208	257	.576	187.2	$.4051 \pm .0008$
WS256-full	.003	8,208	33	.354	1476.6	$.4184 \pm .0088$
FS256-full	.003	8,208	33	.066	11.3	$.4184 \pm .0088$

FS256-full and WS256-full see identical contiguous origin blocks in every update. Their paired test-MSE difference (FS minus WS) is -0.00000702 with a hierarchical circular-block bootstrap 95% interval $[-0.00002261, 0.00000363]$ (block 96, 3,000 draws); revision RMS, exact moment, and probe R^2 agree to four decimals (.1503, .2021, .0861). FS is $5.40\times$ faster and uses 99.2% less peak allocation. It does not improve mean MSE over the update-matched sparse control, while WS32-full is more accurate at 257 updates. This cell therefore validates the computational mechanism and numerical equivalence, but supplies no evidence for a dense-supervision generalization benefit or for cross-dataset claims of the original architecture.

G Controlled compute audit

We time the validation-selected ETTh1 $L = 336, H = 96$ graphs on an otherwise idle NVIDIA RTX 5090. Every objective receives the same real paired-origin batch, architecture, initialization seed, zero-learning-rate Adam optimizer, and eager CUDA path. Zero learning rate holds weights fixed across repetitions without omitting the optimizer update. Each reported training step includes two forecast forwards, loss construction, backward, and Adam. Each of five trials warms every objective for at least 500 ms, then interleaves 100 timed steps in five blocks whose objective order rotates. Inference is warmed for at least 250 ms and averages 200 ordinary forwards per trial. A process-table guard found no competing CUDA process at either endpoint. The values are implementation- and hardware-specific microbenchmarks, not throughput claims.

Table 10: Controlled compute audit. Times are trial means in milliseconds per batch. Parentheses give training-step time relative to ERM; memory lists ERM/Moment-U/Stability peak allocated MiB.

Backbone	Batch	Parameters	ERM	Moment-U	Stability	Peak MiB / inference ms
DLinear	8	64,704	.880	1.449 (1.65×)	1.020 (1.16×)	17.8/18.9/17.8 / .051
PatchTST	128	81,728	10.060	10.189 (1.01×)	10.067 (1.00×)	635.5/635.5/635.5 / .964
iTransformer	32	903,008	5.329	5.758 (1.08×)	5.390 (1.01×)	48.5/48.5/48.5 / .360

The U-statistic's absolute loss-side work is small relative to both Transformer backbones but is visible when the backbone and batch are extremely small. Peak allocation is unchanged to the reported precision for both Transformers. All deployed methods use the identical single-forward graph, so the inference column is shared within each backbone rather than repeated by objective.

H Artifact and claim-boundary checklist

The released artifact records:

1. oracle news and diagonal-estimator controls;
2. validation-only selection, full weight curves, failures, and three-seed broad/horizon cells;

3. five-seed headline estimates and dependence-aware paired intervals, with the timing of the PatchTST/ETTh2 precision extension stated explicitly;
4. frozen kernel and ExtraTrees audits that cannot select a forecast;
5. validation-budget-matched Stability, post-hoc smoothing, and compatible Forking-Sequences controls;
6. peak memory, training-step overhead, and unchanged one-pass inference; and
7. citation, leakage, anonymity, AI-disclosure, and page-limit checks.

Table 11: Claim–evidence map. Each empirical slice changes one generalization axis while holding the others fixed; negative outcomes remain in the record. The final column states the supported scope and its explicit boundary.

Question	Falsification slice	Recorded outcome	Supported scope / explicit boundary
Estimator mechanism	Oracle homoskedastic and heteroskedastic news	Moment-U stays centered and retains news; diagonal and Stability shrink it	Removes the self-pair shortcut; does not certify calibration
Dataset transfer	Six datasets, fixed DLinear/$H = 96$	Validation accepts Moment-U on $5/6$; Exchange is retained as a failure	Dataset breadth for one backbone and horizon
Horizon transfer	Four ETTh1 horizons, fixed DLinear	Direction persists at all three transferred horizons with negligible MSE change	Horizon transfer for one dataset and backbone
Backbone transfer	PatchTST/iTransformer on ETTh1/2, five seeds	Frozen kernel score changes by -6.6% to -40.3%; MSE has one cost, two unresolved changes, and one gain	Transfer across three backbone families; not accuracy dominance
Richer restriction	Pre-frozen raw-context kernel gate	Gate passes $3/4$; DLinear fails and most Moment-U seeds still reject the null	Repeatable partial improvement, not complete conditional calibration

Table 12: Replication and uncertainty accounting. Every stochastic benchmark cell uses multiple seeds, and every aggregate is backed by released per-seed values. Hierarchical intervals resample seeds and paired circular origin blocks.

Evidence block	Independent fits	Across-seed report	Dependence-aware uncertainty
Synthetic mechanism	100 fits; 50,000 minibatches per batch size	Mean $\pm$ Monte Carlo s.e.	Independent synthetic replications
Six-dataset DLinear slice	3 seeds per dataset; ETTh1 headline extended to 5	Means, s.d., and all per-seed values released	Five-seed ETTh1 uses a 5,000-draw hierarchical MSE interval
DLinear horizon transfer	3 seeds at each of 4 horizons	Means, s.d., and all per-seed values released	3,000-draw hierarchical MSE intervals at every horizon
Nonlinear headline cells	5 seeds in each of 4 backbone–dataset cells	Mean $\pm$ s.d. in Table 4; all values released	5,000-draw hierarchical MSE intervals with paired length-96 blocks
Frozen kernel gate	5 paired seeds in each gate cell	Mean, s.d., per-seed values, and paired direction counts	4,999 block-wild draws per seed at block lengths 48/96/192