MOSTLY THE NORM: DECOMPOSING FREQUENCY-DOMAIN FORECAST LOSSES INTO BASIS, NORM, AND RESIDUAL WEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Frequency-domain forecast losses such as FreDF change three things at once: the basis in which the residual is measured, the norm applied to it and the weight of each coordinate. We separate the three. (i) Under a squared norm any unitary basis is inert (Parseval), and training confirms it. (ii) Under ℓ_1, on iTransformer and at an equal mix of the two terms, the same ℓ_1 term with no transform recovers most of FreDF's gain and is behind FreDF on none of nineteen held-out dataset–horizon cells. At FreDF's released weights it is ahead on ten of twenty cells and behind on four, mostly at the short horizons of the hourly datasets. (iii) To measure explicit weighting we use WHITEDF as an instrument, a squared loss on the residual whitened by detached, shrunk, Kronecker-factored second-moment estimates. Linear theory gives the invariance condition under which a fixed quadratic metric can move the fitted predictor. Empirically, the estimated weighting is ahead of the time-domain ℓ_1 term on no iTransformer cell. Its largest gain, 5.3% on PatchTST at ETTh1 and horizon 720 over three seeds, does not survive a Holm correction, depends on the metric's eigenvectors rather than its spectrum, and disappears once the test period moves three months later. A time-domain ℓ_1 term at matched strength is therefore the baseline against which frequency-domain and whitened losses should be judged.

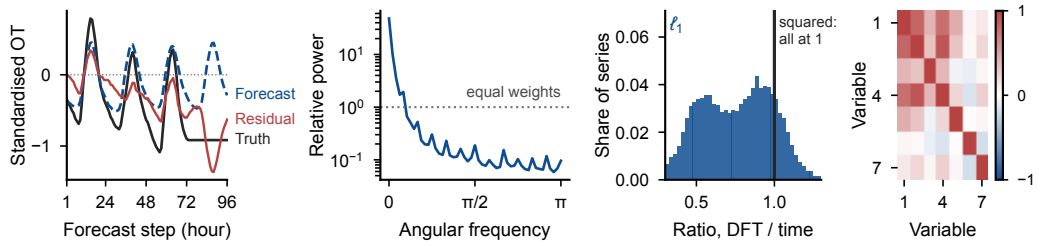
1 INTRODUCTION

Most deep forecasters are trained by *direct forecasting* (DF). The model outputs all T future steps of all D variables at once, and the loss sums the squared error over them (Liu et al., 2022; Zhou et al., 2021). Recent progress has been architectural, with attention over time or across variables (Zhou et al., 2021; 2022; Nie et al., 2023; Liu et al., 2024a), convolution (Wu et al., 2023; Luo & Wang, 2024; Li et al., 2025) and multi-scale mixing (Wang et al., 2024; 2025c), and most of it keeps the same objective.

Squared error targets the conditional mean whatever the correlation and scale of the residuals. It weights every step and variable equally, and forecast residuals are correlated along the horizon and across variables (Figure 1). FreDF (Wang et al., 2025a) addresses the temporal part of this structure. It computes the loss after a discrete Fourier transform (DFT) of forecast and label, because the DFT coefficients of a stationary sequence are asymptotically uncorrelated. Relative to DF this changes the basis, the norm (an ℓ_1 frequency term in place of a squared one) and, implicitly, the weight of each coordinate (Figure 1b,c). The decorrelation argument leaves open which of the three produces the gain. We measure each of them, and we measure what an explicit residual weighting adds.

A change of loss can be judged at three levels, and none of them implies the next. The objective may change, the fitted predictor may change, and test error may change. Classical results settle the first two levels for linear predictors, and controlled experiments address the third. Whitening a residual by a covariance Σ and measuring it in the squared norm is generalised least squares (GLS). For explicit weighting we use WHITEDF, a whitened direct forecast. It transforms the residual by the inverse square roots of estimated temporal and cross-variable second-moment factors, $\Sigma_c \otimes \Sigma_t$, before the squared loss. The factors are moving averages of training-batch residuals, shrunk towards

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.



(a) Residual drifts for hours (b) Spectrum not flat (c) ℓ_2 ignores the basis (d) Across variables

Figure 1: **Forecast residuals are correlated, but a change of basis alone cannot act on them.** Residuals of a closed-form linear forecaster on ETTh2 at $\mathrm{T} = 96$ (Appendix E). (a) Truth, forecast and residual of the OT variable over the first test window. (b) Residual power per frequency relative to its mean. An equal-weight frequency loss treats every frequency alike (dotted). (c) Loss after the unitary DFT divided by the loss in time, for the 19,985 test residual series. The squared loss is unchanged (Corollary 1), the ℓ_1 loss is not. (d) Normalised second moment across the seven variables. Details in Appendix L.

the identity and detached from the gradient. DF is the $\gamma = 1$ member of the family. Three questions organise the paper (Figure 2).

1. **What do the basis and the norm contribute?** *On iTransformer, mostly the norm.* Under the squared norm the basis is inert (Corollary 1). Under ℓ_1 the same term without a transform, which we call time-ℓ_1, recovers 0.0036 of the 0.0042 MSE that our equal-mix FreDF gains on the development cell and is behind it on none of nineteen held-out cells. At FreDF's released weights, time-ℓ_1 at matched strength is ahead on ten of twenty cells and behind on four, mostly at short hourly horizons (Section 3.2).
2. **When does a fixed weighting change the fitted predictor?** *Only where an invariance condition fails, and finite training adds a cost of its own.* Channel independence is the case in which the cross-variable factor can matter (Section 4.1). For the shared-weight linear class a temporal factor cannot move the optimum, yet a frozen one costs up to 0.044 MSE on DLinear after a fixed training budget (Section 4.3).
3. **When does an estimated weighting help?** *On some cells, depending on dataset, backbone and test period.* On iTransformer its six held-out gains all share the development cell's station and test months, and it is ahead of time-ℓ_1 on no cell. On PatchTST its largest gain (5.3% at ETTh1, $\mathrm{T} = 720$, three seeds) does not survive Holm and fades as the test period moves two to four months later (Section 5).

The formal results are classical, and WHITEDF is a measurement instrument rather than a proposed training loss. The contribution is the decomposition and the controls that separate basis, norm and weighting. **Its practical conclusion is that a time-domain ℓ_1 term at matched strength is the baseline against which frequency-domain and whitened losses should be judged.**

2 FORECAST LOSSES AND EVALUATION SETUP

2.1 FORECASTS AND LOSSES

Following Wang et al. (2025a), a multivariate series $[X(1), \ldots, X(N)]$ has $X(n) \in \mathbb{R}^{1\times \mathrm{D}}$. At step n the input is $L = [X(n-\mathrm{H}+1), \ldots, X(n)] \in \mathbb{R}^{\mathrm{H}\times \mathrm{D}}$ and the label is $Y = [X(n+1), \ldots, X(n+\mathrm{T})] \in \mathbb{R}^{\mathrm{T}\times \mathrm{D}}$, with context length H and horizon T. A forecaster g_θ outputs $\hat{Y} = g_\theta(L)$. We write $E = Y - \hat{Y}$ for the residual matrix, $e = \mathrm{vec}(E) \in \mathbb{R}^{\mathrm{TD}}$ for its column-stacked vector, $\Sigma := \mathrm{Cov}(e \mid L)$ for the conditional residual covariance, and $\Sigma_t \in \mathbb{R}^{\mathrm{T}\times \mathrm{T}}$, $\Sigma_c \in \mathbb{R}^{\mathrm{D}\times \mathrm{D}}$ for its temporal and variable factors. DF optimises

$$\mathcal{L}^{(\mathrm{tmp})} := \|E\|_F^2 = \sum_{t=1}^{\mathrm{T}} \sum_{d=1}^{\mathrm{D}} \left(Y_{t,d} - \hat{Y}_{t,d}\right)^2 = e^\top e. \tag{1}$$

Definition 1 (Whitened forecast loss) *For a positive definite $\Sigma \in \mathbb{R}^{\mathrm{TD}\times\mathrm{TD}}$, the whitened forecast loss is $\mathcal{L}_{\Sigma}^{(\mathrm{gls})} := e^{\top}\Sigma^{-1}e = \|We\|_2^2$ for any whitening transform W of Σ, that is any W with $W^{\top}W = \Sigma^{-1}$, so that $\mathrm{Cov}(We) = \mathbf{I}$ when $\mathrm{Cov}(e) = \Sigma$.*

The loss is the same for every whitening transform W. The symmetric square root $\Sigma^{-1/2}$ is one choice and QW another for any orthogonal Q. A unitary transform such as the DFT changes the basis and leaves the eigenvalues of Σ in place, so on its own it does not whiten.

Both objectives target the conditional mean. For any Σ fixed or depending on L alone, and any square-integrable Y, $\mathbb{E}[\mathcal{L}_{\Sigma}^{(\mathrm{gls})} \mid L]$ is minimised over all predictors by $\mathbb{E}[Y \mid L]$. A metric estimated from the model's own residuals (Section 4.2) is outside this statement. The two objectives weight the coordinates of the error differently whenever Σ is not a multiple of the identity (Proposition 1).

Proposition 1 (When the two objectives coincide) *$\mathcal{L}^{(\mathrm{tmp})}$ and $\mathcal{L}_{\Sigma}^{(\mathrm{gls})}$ coincide as functions of the residual e, up to a positive affine transformation, if and only if $\Sigma \propto \mathbf{I}_{\mathrm{TD}}$.*

2.2 Related Objectives

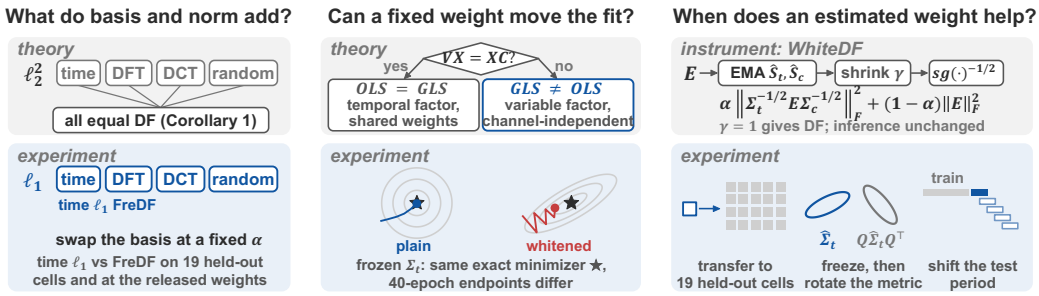
The architectures above, among them TimesNet, ModernTCN, TVNet, TimeMixer and TimeMixer++, train on time-domain squared error, and the foundation model Time-MoE (Shi et al., 2025b) on the Huber loss. Several represent the input in the frequency domain (Zhou et al., 2022; Yi et al., 2023b;a; Wang et al., 2025c), and all of them measure the error in the time domain. DF has largely displaced iterative rollouts (Taieb & Atiya, 2016) in this literature (Zhou et al., 2021; Wu et al., 2023; Liu et al., 2024a; Nie et al., 2023). FreDF (Wang et al., 2025a) keeps DF, computes the error after a DFT of forecast and label, and mixes the frequency term with the time-domain one by a weight α. OLMA (Shi et al., 2025a) applies the same ℓ_1 term along the variable axis as well. FreDF's authors have since replaced the fixed basis by an SVD of the training labels (Time-o1 (Wang et al., 2025b)) and by a $\mathrm{T} \times \mathrm{T}$ quadratic form learned through the model by bilevel optimisation (QDF (Wang et al., 2026a)). SFALoss (Zhao et al., 2026) weights a frequency-domain ℓ_1 loss by a learned spectral filter, and FreST (Wang et al., 2026d) carries the decorrelation argument to spatio-temporal forecasting. None of these derives the weight from the residual spectrum.

Another line models error correlation during training, with an AR(1) whitening learned jointly with the model (Sun et al., 2021), learned covariances under the full Gaussian likelihood (Zheng et al., 2025; 2024; Zheng & Sun, 2024), joint predictive distributions (Salinas et al., 2019; Rasul et al., 2021), conditionally whitened generative models (Yang et al., 2026) and a fixed cross-variable graph penalty (Ding et al., 2026). The covariance of WhiteDF is instead a detached, shrunk plug-in estimate from training residuals, so the criterion remains a reweighting of point-forecast error (Appendix A).

2.3 Evaluation Protocol and Vocabulary

We use ETT (four subsets), Weather, ECL and, in Section 5.3, Traffic with chronological splits, input length 96, horizons $\{96, 192, 336, 720\}$ and MSE on the standardised scale. The protocol is that of Liu et al. (2024a), with the drop-last trick disabled at test time (Qiu et al., 2024). iTransformer with its reference hyperparameters is the main backbone, and DLinear and PatchTST are added in Sections 4.3 and 5.2 (Appendices H and I).

A *cell* is one dataset at one horizon. Of the twenty ETT and Weather cells, ETTh1 at $\mathrm{T} = 96$ is the *development cell*, on whose test error the WhiteDF setting was chosen, and the other nineteen are *held out*. An *arm* is one training criterion run on a cell. *Time-ℓ_1* is $\alpha\,\mathrm{mean}|E| + (1-\alpha)\,\mathrm{mean}\,E^2$, FreDF's ℓ_1 term without its transform. *FreDF-eq* is our re-implementation at an equal mix and *FreDF-rel* is FreDF at its released weights (Section 3.2). Every arm is compared with a DF baseline retrained on the same backbone, seeds and code path, paired on seed, by a two-sided paired t-test. A difference with $p < 0.05$ is *beyond noise*, and one with $p \geq 0.05$ is *not separated*, which means the test cannot reject equality and is not evidence of equivalence. For scale, DF's test MSE on the development cell varies by 0.0015 across five seeds, 0.4% of its value. Headline effects carry 95% confidence intervals. Group statements count per-cell verdicts, and each count is also given under a Holm correction within its *family*, one backbone, protocol and contrast over its cells



(a) Norm and basis (b) Fixed weighting (c) Estimated weighting

Figure 2: **Theory decides what a change of loss can do, and paired runs measure what it does.** (a) Section 3, (b) Sections 4.1 and 4.3, (c) Sections 4.2 and 5.

(Appendix I.2). Cells are never pooled into one test, because the horizons of one dataset are not independent replicates. The reference code stops early on each criterion's own validation loss (the *reference rule*). Several families were rerun with the checkpoint chosen on validation MSE for every criterion (the *common rule*).

3 Norm and Basis

The formal results of Sections 3.1 and 4.1 are classical. Proofs are in Appendix B and further discussion in Appendix C.

3.1 The Squared Norm Makes the Basis Inert

FreDF (Wang et al., 2025a) computes its frequency term after a fixed unitary transform. Under a circulant temporal covariance, GLS is a squared spectral error weighted by the inverse residual power λ_k^{-1} at each frequency (Theorem 1, Appendix C.1), the quadratic term of the Whittle likelihood. Reading a squared frequency loss as this GLS needs two separate assumptions. A circulant covariance makes the DFT the diagonalising basis, and a flat spectrum makes the weights equal. With equal weights the transform drops out.

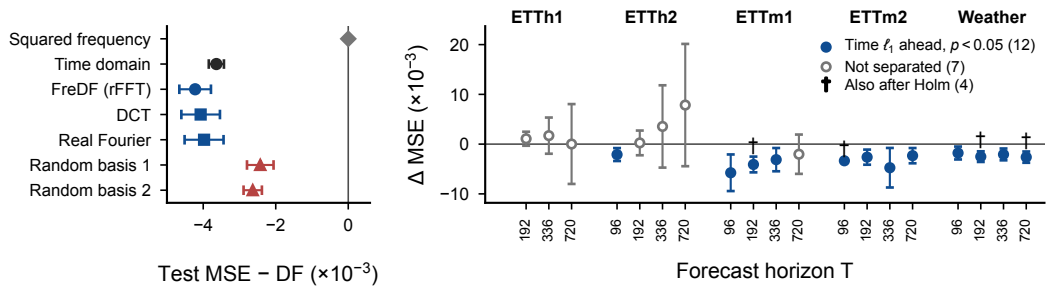
Corollary 1 (Parseval: equal weights make the transform inert) *Let $\mathcal{F}$ be the normalised DFT, which is unitary. Then $\|\mathcal{F}(E)\|_F^2 = \|E\|_F^2$ for every E. Consequently the unweighted squared frequency loss is numerically identical to the time-domain DF objective (1). Under a squared norm, a change of basis without a change of weights leaves the loss unchanged. (Full unitary DFT; half-spectrum: Appendix B.)*

The corollary holds for any unitary transform, along time or across variables, and only for the squared norm. FreDF, and OLMA (Shi et al., 2025a) on both axes, measure the frequency term under ℓ_1, which is not unitarily invariant, so the corollary does not apply to them. Each of the three ingredients therefore needs its own test. The corollary settles the basis under a squared norm. A time-domain ℓ_1 control isolates the norm. The per-frequency weight is λ_k^{-1} under a Gaussian likelihood, or the scale b_k^{-1} of the Laplace likelihood that an ℓ_1 term implies (Remark 4, Appendix F.1), and WhiteDF supplies it from an estimate.

3.2 What Does the Basis Add under ℓ_1?

The corollary predicts that an equal-weight squared frequency loss cannot differ from the time-domain loss. At a fixed point the two gradients agree to float32 rounding. Over five seeds on ETTh1 at $\mathrm{T} = 96$ the paired difference in test MSE is **zero to six decimals on every seed** (Figure 3a; per seed in supplementary Figure S1).

FreDF-eq is our re-implementation, $\alpha\,\mathrm{mean}|\mathcal{F}(E)| + (1-\alpha)\,\mathrm{mean}\,E^2$ with the orthonormal rFFT and $\alpha = 0.5$, and time-ℓ_1 drops $\mathcal{F}$. The released code uses the unnormalised transform, $\sqrt{\mathrm{T}}$ heavier,



(a) Six bases on the development cell

(b) Time-domain ℓ_1 minus FreDF on the held-out cells

Figure 3: **Under ℓ_1 the basis adds little.** iTransformer, five seeds per cell. (a) Test MSE minus DF on ETTh1 at $\mathrm{T} = 96$ for an ℓ_1 term at $\alpha = 0.5$ in six orthonormal bases, and for the squared frequency loss, with 95% intervals. The time-domain row uses no transform. (b) Filled markers have $p < 0.05$, and a dagger marks a cell that survives Holm.

and its iTransformer scripts set the frequency weight to 0.6–0.9 per dataset, so FreDF-eq measures FreDF's basis and norm at an equal mix (Appendix I), and FreDF-rel, the released operating point, is tested below. FreDF-eq improves on DF by 0.0042 MSE (1.1%; 95% CI $[0.0038, 0.0047]$), and time-ℓ_1, at the same α and with no transform, by 0.0036 (0.9%; $[0.0034, 0.0039]$). The difference of 0.0006 ($[0.0000, 0.0012]$, $p = 0.047$) is what the transform adds on this cell. Other structured bases share it (Figure 3a). A DCT is not separated from FreDF-eq, and two random orthonormal bases fall behind time-ℓ_1 by 0.0010 to 0.0012. A random rotation makes the coordinates nearly Gaussian, so its ℓ_1 term stays within a few percent of $\sqrt{2\mathrm{T}/\pi}\,\|e\|_2$ and acts as a Euclidean norm, while a trained model's residuals keep heavier tails in time and are sparse in the DCT (Figure 11). Structured bases are interchangeable within the resolution of five seeds (Appendix F.3).

On the nineteen held-out iTransformer cells, with five seeds, time-ℓ_1 beats DF on eighteen and is worse on none (Figure 3b). Against FreDF-eq it is ahead on twelve and behind on none, and four of the twelve survive Holm (Appendix I.2). The ordering is the same under the common checkpoint rule (Appendix F.8). Against FreDF-rel, with DF retrained on the same GPU and time-ℓ_1 scaled by $\sqrt{\mathrm{T}}$ to the same gradient strength, time-ℓ_1 is ahead on ten of twenty cells and behind on four, and all fourteen verdicts survive Holm (Appendix F.7). With each arm's weight chosen on validation it is ahead on ten and behind on six. All arms use the paper's reference iTransformer configuration rather than FreDF's own hyperparameters, and in it FreDF-rel is worse than DF at ETTh1, $\mathrm{T} = 720$ and at three Weather horizons (Appendix F.7). FreDF's edge is at the shorter horizons of ETTh1 and ETTh2 and at ETTm2 $\mathrm{T} = 336$, as in its authors' repository on TQNet (Appendix F.6). **The norm carries most of the gain, and the transform keeps a measurable part at the shortest hourly horizons.**

4 Fixed Weighting: Exact Fits and Finite Training

4.1 When Does a Fixed Metric Change the Estimator?

Stack N training windows into the linear model $y = X\beta + \varepsilon$. Here X is a fixed design of full column rank that encodes the predictor's parameterisation, β holds its parameters, $\mathbb{E}[\varepsilon \mid X] = 0$, and $\mathrm{Cov}(\varepsilon \mid X) = V \succ 0$ is known. Only finite second moments are assumed. Aitken's theorem (Theorem 2, Appendix C.2) gives $\mathrm{Cov}(\hat{\beta}_{\mathrm{gls}}) \preceq \mathrm{Cov}(\hat{\beta}_{\mathrm{ols}})$ in the Loewner order. The two coincide if and only if the column space of X is an invariant subspace of V, that is $VX = XC$ for some nonsingular C (Zyskind, 1967; Rao, 1967; Kruskal, 1968). We call this the invariance condition. For an unconstrained linear map from the full context to the full horizon, $X = \mathbf{I}_{\mathrm{TD}} \otimes Z$, and the condition holds for every $V = \Sigma \otimes \mathbf{I}_N$. **OLS and GLS then coincide, and no fixed quadratic metric of this form alters the fitted predictor** (Corollary 3).

Corollary 2 (Cross-variable weighting in channel-independent linear models) *If the predictor is channel-independent, so that output channel d is a function of channel d's own history alone*

and X is block diagonal with distinct per-channel blocks $\{X_d\}_{d=1}^{\mathrm{D}}$, and if $\Sigma = \Sigma_c \otimes \Sigma_t$, then the invariance condition holds when Σ_c is diagonal and each per-channel block is unconstrained; for generic per-channel designs it fails whenever Σ_c carries off-diagonal mass. In that case the covariance difference in Aitken's inequality is non-zero, and the GLS estimator is more efficient because of the cross-variable residual correlation.

Whitening changes which member of a channel-independent predictor's fixed hypothesis class is selected, the seemingly-unrelated-regressions gain of Zellner (1962) read for forecasters. Channel independence is the prevailing design (Nie et al., 2023; Zeng et al., 2023; Shi et al., 2025b), so the corollary suggests that the variable factor should matter more for channel-independent backbones than for channel-mixing ones. For deep backbones this is a heuristic, since iTransformer and PatchTST both end in an output head shared across variables. For the shared-weight models we test, temporal whitening alone cannot move the estimator at any fusion weight, and the pure two-factor GLS solution depends on Σ_c alone. The mixed objective we train couples both factors (Remark 2). The results concern linear predictors with a known V, and Section 5 tests the estimated metric on nonlinear backbones. They order estimator covariances, not test errors: in a two-output, one-parameter example the true covariance raises test MSE once the misfit exceeds the noise variance over n (Proposition 4).

4.2 WHITEDF: A WEIGHTING WE CAN VARY

WHITEDF is the instrument with which we test residual weighting. It has to apply $\Sigma^{-1/2}$ when Σ is too large to form, with a metric that stays fixed with respect to the parameters being trained (Figure 2c and design notes in Appendix D). We assume the residual covariance is separable,

$$\Sigma \;\approx\; \Sigma_c \otimes \Sigma_t, \qquad \Sigma_t \in \mathbb{R}^{\mathrm{T}\times\mathrm{T}},\;\; \Sigma_c \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}, \tag{2}$$

the matrix-normal assumption of Dutilleul (1999). Storage falls from $\mathcal{O}(\mathrm{T}^2\mathrm{D}^2)$ to $\mathcal{O}(\mathrm{T}^2+\mathrm{D}^2)$.

Proposition 2 (Closed-form separable whitening) *If (2) holds exactly, $\Sigma^{-1/2}e = \mathrm{vec}\left(\Sigma_t^{-1/2} E\, \Sigma_c^{-1/2}\right)$, and hence, applied to the approximation,*

$$\mathcal{L}^{(\mathrm{wht})} \;=\; \left\| \Sigma_t^{-1/2}\, E\, \Sigma_c^{-1/2} \right\|_F^2. \tag{3}$$

$\Sigma_t^{-1/2}$ reweights error across forecast steps and $\Sigma_c^{-1/2}$ across variables. Setting $\Sigma_c = \mathbf{I}$ gives the temporal variant WHITEDF$_\mathrm{t}$, and estimating both factors gives WHITEDF$_{\mathrm{t+c}}$. Diagonal factors give a weighted-least-squares rung that recovers a tenth of the temporal variant's gain (Appendix F.4). On each training minibatch $\mathcal{B}$ we update exponential moving averages (EMA) $\hat{S}_t$ and $\hat{S}_c$ of the residual second moments $E_b E_b^\top/\mathrm{D}$ and $E_b^\top E_b/\mathrm{T}$ at rate β (Eqs. (12)–(13), Appendix D), initialised at the identity, so training starts at the DF objective. The updates run only on training batches with gradients enabled, and validation and test passes leave $\hat{S}_t$ and $\hat{S}_c$ untouched. Each factor is shrunk towards a scaled identity (Ledoit & Wolf, 2004),

$$\Sigma_t^{(\gamma)} = (1-\gamma)\,\hat{S}_t + \gamma\,\frac{\operatorname{tr}\hat{S}_t}{\mathrm{T}}\,\mathbf{I}_\mathrm{T}, \qquad \Sigma_c^{(\gamma)} = (1-\gamma)\,\hat{S}_c + \gamma\,\frac{\operatorname{tr}\hat{S}_c}{\mathrm{D}}\,\mathbf{I}_\mathrm{D}, \tag{4}$$

with one $\gamma \in [0,1]$ for both axes. For $\gamma > 0$ this bounds the condition number, and at $\gamma = 1$ the whitened loss equals $\mathcal{L}^{(\mathrm{tmp})}$ up to a scalar (Proposition 3). The factors are then rescaled to $\operatorname{tr}\Sigma_t^{(\gamma)} = \mathrm{T}$ and $\operatorname{tr}\Sigma_c^{(\gamma)} = \mathrm{D}$ (Appendix F.9).

The factors enter through a stop-gradient $\mathrm{sg}(\cdot)$, so the metric is a plug-in quantity (Appendix C.8). Optimised jointly with θ, the quadratic form would inflate Σ without bound. Following Wang et al. (2025a), the objective keeps a share of the unwhitened loss:

$$\mathcal{L}^{\alpha} \;=\; \alpha \cdot \underbrace{\left\| \mathrm{sg}\left(\Sigma_t^{(\gamma)}\right)^{-1/2} E\; \mathrm{sg}\left(\Sigma_c^{(\gamma)}\right)^{-1/2} \right\|_F^2}_{\mathcal{L}^{(\mathrm{wht})}} + (1-\alpha)\cdot \underbrace{\|E\|_F^2}_{\mathcal{L}^{(\mathrm{tmp})}}, \qquad 0 \le \alpha \le 1. \tag{5}$$

The criterion is sensitive to both γ and α and is worse than DF at some settings (Figure 15). Algorithm 1 adds five lines to a DF trainer and nothing at inference.

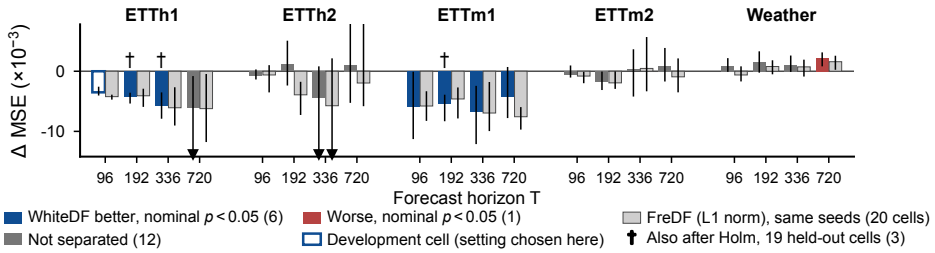


Figure 4: **One setting chosen on the development cell improves six held-out cells, all from its own station and test months.** The setting ($\alpha = 0.5$, $\gamma = 0.5$) was chosen on the hollow cell and applied unchanged to nineteen held-out iTransformer cells, five seeds each. Colour marks a two-sided paired t-test at $p < 0.05$, a dagger marks a gain that survives Holm over the nineteen, and light grey bars are FreDF. Details in Appendix L.

4.3 A Factor That Cannot Move the Optimum Still Costs Test Error

The two-stage experiment separates the metric from its online estimate. A factor is estimated once from closed-form training residuals and frozen, and DLinear is then trained for 40 epochs on a fixed schedule (Appendix F.11). A frozen variable factor lowers test MSE at all twelve settings on ETTh2, ETTm2 and Weather, seven beyond noise, and at none of four on ETTh1. **This follows the sign of the closed-form gain**, which is positive on those three datasets and -0.3% on ETTh1 (Table 2).

A frozen temporal factor cannot move this estimator (Remark 2). It still costs up to $+0.044$ on ETTh2 after the fixed schedule. We measured the gap on the class the remark covers, a shared-weight linear map trained with the same optimiser and budget (Appendix F.12). Under plain MSE, training ends within 0.02 of the closed-form minimiser in relative parameter distance, and under the pure whitened loss 0.05 to 0.37 away, in the order of the frozen factor's condition number. **Keeping half of the plain term removes the gap.** The bare linear map reproduces the direction of the DLinear cost, and DLinear's cost is an order of magnitude larger.

5 Estimated Weighting across Backbones and Datasets

5.1 On iTransformer, the Estimated Weighting Trails the Norm

On the development cell (ETTh1, T = 96, five seeds) WhiteDF$_t$ at $(\alpha, \gamma) = (0.5, 0.5)$ improves on DF by 0.0035 MSE ($[0.0028, 0.0042]$). The setting was chosen on this cell's test error, so the evidence for it comes from nineteen cells that did not influence it, to which Figure 4 applies it unchanged. Six improve beyond noise, twelve are not separated and one is worse (Weather at T = 720, $+0.0022$). All six gains are on ETTh1 and ETTm1, which record the development cell's transformer station over the same test months. On ETTh2, ETTm2 and Weather none of the twelve cells improves. Under Holm three of the six gains remain and the harm does not (Appendix I.2).

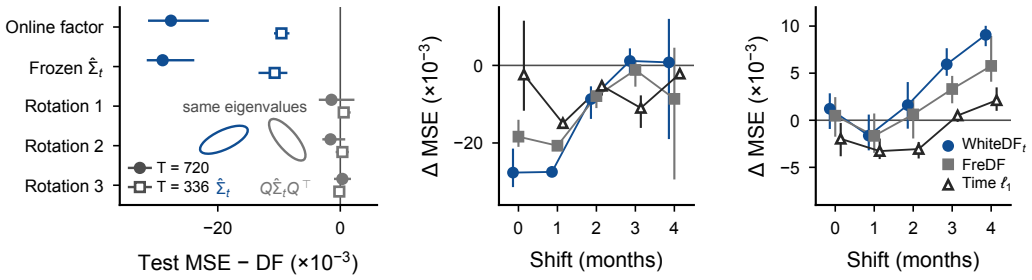
On this backbone the estimated weighting recovers part of what time-ℓ_1 gains and passes it on no cell. Against time-ℓ_1 it is not separated on the development cell ($+0.0001$, $[-0.0007, 0.0010]$), and on the held-out cells, with five seeds, it is behind on fourteen and ahead on none, seven of the fourteen surviving Holm (Appendix I.2). Against FreDF-eq on the same five seeds and all twenty cells it is not separated on fourteen, worse on five and better on one, and none of these verdicts survives Holm on the held-out cells (Table 17). An AR(1) whitening (Sun et al., 2021), run only at the setting $(0.5, 0.1)$, and QDF are compared in Table 18 and Appendix F.6.

5.2 On PatchTST, the Largest Gain Is Tied to Direction and Period

We add two channel-independent backbones, PatchTST (Nie et al., 2023) and DLinear (Zeng et al., 2023), from the Time-Series-Library scripts (Wu et al., 2023). On the hourly ETT cells of PatchTST all three criteria gain against DF (Table 1).

Table 1: **On PatchTST every criterion lowers test MSE on most hourly-ETT cells, and whitening leads the other criteria only at ETTh1,** $\mathrm{T} = 720$**.** That lead is uncorrected; the one whitening gain over DF that survives Holm is at $\mathrm{T} = 336$. Common checkpoint rule, three seeds. DF is the mean test MSE, and the other columns are paired differences from DF. Bold marks $p < 0.05$, uncorrected, and † marks a cell that also survives Holm within its column. Paired t-values and the direct contrasts are in Table 9.

		DF	Difference from DF ($\times 10^{-3}$)			
Dataset	T	MSE	FreDF	Time ℓ_1	WHITEDF$_{\mathrm{t}}$	WHITEDF$_{\mathrm{t+c}}$
ETTh1	96	0.379	+0.5	−2.0	+1.2	+2.8
	192	0.431	**−3.2**	**−3.6**†	**−3.3**	−1.9
	336	0.474	**−8.8**	**−8.8**†	**−9.6**†	**−8.4**†
	720	0.521	**−18.4**	−2.4	**−27.6**	**−30.0**
ETTh2	96	0.300	**−4.6**	**−5.4**	**−5.4**	**−7.1**
	192	0.377	−1.9	−3.4	−3.0	−4.0
	336	0.433	−9.0	−3.9	−11.3	−9.3
	720	0.442	−6.6	−3.2	−11.4	−12.6

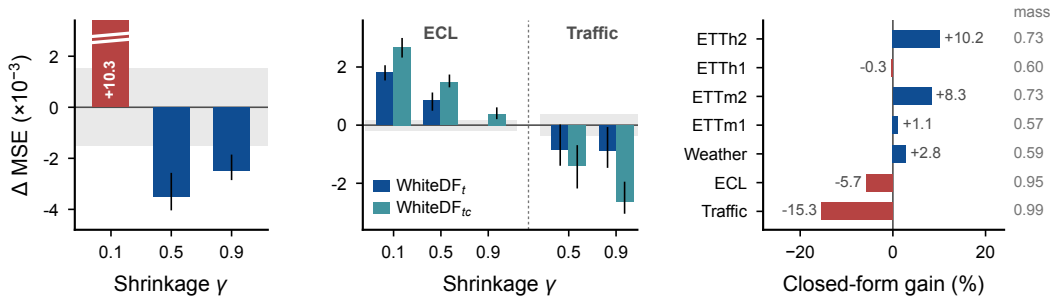


(a) Frozen and rotated metrics (b) $\mathrm{T} = 720$ on shifted folds (c) $\mathrm{T} = 96$ on shifted folds

Figure 5: **The PatchTST gain on ETTh1 depends on the metric's directions and on the test period.** Three seeds. Markers are mean differences from the DF of the same fold, and whiskers span the seeds. The ellipses in (a) sketch the control: a rotation Q keeps the eigenvalues of $\hat{\Sigma}_t$ and turns its eigenvectors. (a) At $\mathrm{T} = 720$ (filled) and $\mathrm{T} = 336$ (hollow), a temporal factor frozen at the DF model's training residuals keeps the gain, and three rotations with the same spectrum lose it. (b, c) Folds shifted by one to four months, whose test periods overlap the reference one by 75, 50, 25 and 0%.

The largest effect in the paper, at ETTh1 $\mathrm{T} = 720$, rests on three seeds and does not survive Holm within the eight-cell family, where only the $\mathrm{T} = 336$ gain does (Appendix I.2). We report it with the three checks that bound it. There DF's MSE is 0.521. WHITEDF$_{\mathrm{t}}$ lowers it by 0.0276 ($[0.0142, 0.0411]$, 5.3%), WHITEDF$_{\mathrm{t+c}}$ by 0.0300 (5.8%), FreDF-eq by 0.0184 ($[0.0089, 0.0278]$) and time-ℓ_1 by 0.0024 ($[-0.0281, 0.0329]$), and WHITEDF$_{\mathrm{t}}$ is ahead of both beyond noise, uncorrected (Table 9). Freezing the temporal factor at the DF model's own training residuals reproduces the gain (-0.0290, $t = -11.2$), and three rotations with the same spectrum lose it. The same holds at $\mathrm{T} = 336$, where the frozen factor gives -0.0107 and the rotations -0.0002 to $+0.0006$, so **at both cells the gain depends on the metric's directions** (Figure 5a, Appendix F.10). On folds shifted by one to four months the $\mathrm{T} = 720$ gain is -0.0275, -0.0087, $+0.0012$ and $+0.0008$. **The gain holds while the test period overlaps the reference one by three quarters and is gone once the overlap falls to a quarter** (Figure 5b, Appendix F.13). At $\mathrm{T} = 96$ WHITEDF$_{\mathrm{t}}$ is not separated from DF up to two months and worse at three and four ($+0.0059$, $+0.0091$; Figure 5c).

Over all twenty-four PatchTST cells (Figure 9, Appendix F.2) the temporal factor is better than DF beyond noise on four, worse on seven and not separated on thirteen. The two-factor variant is at 3/7/14. The harm is on ETTm1 (beyond noise at $\mathrm{T} = 336$ and 720 for both factors), on Weather at $\mathrm{T} = 192$ and on ECL at every horizon ($+0.011$ to $+0.038$). On DLinear the online estimate gains



(a) ETTh1 across shrinkage (b) ECL and Traffic (c) Closed-form fit, seven datasets

Figure 6: **The sign of residual weighting changes with dataset and backbone.** $\mathrm{T} = 96$. (a, b) iTransformer, test MSE minus DF against the shrinkage γ at $\alpha = 0.5$. Bars are means, whiskers span the seeds (five in a, three in b), and grey bands span $\pm$ DF's seed range. (c) Channel-whitening gain of the closed-form linear forecaster, ordered by $\mathrm{D}/(n\mathrm{T})$, with the off-diagonal mass of Σ_c at the right. Details in Appendix L.

beyond noise on one and two cells, by at most 0.0007, and is worse on seven each (Figure 9), so the frozen factors of Section 4.3 are the informative experiment there.

5.3 The Sign Changes with Dataset and Backbone

On ECL ($\mathrm{D} = 321$) the criterion is not below DF on iTransformer at any shrinkage. The difference shrinks monotonically towards $\gamma = 1$ under both checkpoint rules (Table 10). ETTh1 has the opposite shape, an interior optimum (Figure 6a,b; the full (α, γ) grid is in Figure 15). On Traffic ($\mathrm{D} = 862$) the same backbone gains. The two-factor variant at $\gamma = 0.9$ beats DF by 0.0026 ($[0.0012, 0.0041]$), the other three settings lie below zero short of noise, and validation moves the same way (Table 19). In QDF's repository, with the TQNet backbone, every arm beats DF on ECL (Appendix F.6).

The closed-form linear fit loses 5.7% on ECL and 15.3% on Traffic and gains 10.2% on ETTh2 and 8.3% on ETTm2 (Figure 6c). Its sign on Traffic is the opposite of iTransformer's, so the fit orders datasets for the linear class only. ETTh1 and ETTh2 share the dimension-to-sample ratio $\mathrm{D}/(n\mathrm{T})$ and differ by ten points. The off-diagonal mass of Σ_c is largest where whitening fails (Figure 6c), and the Ledoit–Wolf intensity does not explain the failures (Appendix E). **Whether whitening helps therefore depends on more than dimension and the linear sign.**

6 Conclusion

A frequency-domain forecast loss changes the basis, the norm and the weighting of the residual at once. Under the squared norm the basis is inert. Under ℓ_1 on iTransformer, time-ℓ_1 recovers most of the gain of FreDF-eq and is ahead of it on twelve of nineteen held-out cells and behind on none. Against FreDF at its released or validation-tuned weights it is ahead on ten of twenty cells and behind on four or six, and FreDF's edge is at the shorter hourly horizons and ETTm2 $\mathrm{T} = 336$. **A time-domain ℓ_1 term at matched strength is therefore the baseline against which frequency-domain losses should be compared.** For linear predictors a fixed metric moves the estimator only where the invariance condition fails. Estimated weighting helps on some cells and harms others, and its largest gain depends on the metric's directions and fades on later test periods. The reported uncertainty is seed variation on fixed chronological splits, with Holm corrections within families defined after the results, and under them the largest PatchTST gain is not retained. A rule that predicts when whitening helps remains open (Appendix M).

Reproducibility statement

https://anonymous.4open.science/r/norm-basis-weighting-forecast-losses-056A

Code and per-run records are in this anonymous repository and in the supplementary material. Datasets are in Appendix H, implementation details and loss-module versions in Appendix I, and proofs in Appendix B.

AI use statement

Generative AI tools were used to aid and polish the writing.

References

Alexander C Aitken. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936.

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

James K. Binkley and Carl H. Nelson. A note on the efficiency of seemingly unrelated regression. *The American Statistician*, 42(2):137–139, 1988.

George E P Box and Gwilym M Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, revised edition, 1976.

Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. In *International Conference on Machine Learning*, 2017.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Kuiye Ding, Yifan Hu, Hanchen Wang, and Hao Xue. Multivariate time series forecasting needs cross variable loss. *arXiv preprint arXiv:2608.05742*, 2026.

Pierre Dutilleul. The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123, 1999.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Ming Jin, Yaxuan Kong, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, Vincent S. Tseng, Yu Zheng, Lei Chen, Hui Xiong, Shirui Pan, and Qingsong Wen. Large models for time series and spatio-temporal data: A survey and outlook. *ACM Computing Surveys*, 58(14):1–38, 2026. arXiv:2310.10196.

Jan Kmenta and Roy F. Gilbert. Small sample properties of alternative estimators of seemingly unrelated regressions. *Journal of the American Statistical Association*, 63(324):1180–1200, 1968.

William Kruskal. When are Gauss-Markov and least squares estimators identical? A coordinate-free approach. *The Annals of Mathematical Statistics*, 39(1):70–75, 1968.

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

Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. In *Advances in Neural Information Processing Systems*, 2019.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Hyunwook Lee, Chunggi Lee, Hongkyu Lim, and Sungahn Ko. TILDE-Q: A transformation invariant loss function for time-series forecasting. *arXiv preprint arXiv:2210.15050*, 2022.

Chenghan Li, Mingchen Li, and Ruisheng Diao. TVNet: A novel time series analysis method based on dynamic convolution and 3D-variation. In *International Conference on Learning Representations*, 2025. arXiv:2503.07674.

Shiyu Liu, Rohan Ghosh, and Mehul Motani. Towards better long-range time series forecasting using generative forecasting. *arXiv preprint arXiv:2212.06142*, 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*, 2024a.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *International Conference on Machine Learning*, 2024b.

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

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417, 2015.

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.

Walter Oberhofer and Jan Kmenta. A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica*, 42(3):579–590, 1974.

Sigbert J Prais and Christopher B Winsten. Trend estimators and serial correlation. *Cowles Commission Discussion Paper No. 383*, 1954.

Simo Puntanen and George P. H. Styan. The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician*, 43(3):153–161, 1989.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, and Bin Yang. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17 (9):2363–2377, 2024.

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

C. Radhakrishna Rao. Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 355–372, 1967.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, 2021.

David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank Gaussian copula processes. In *Advances in Neural Information Processing Systems*, 2019.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020.

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

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025b. arXiv:2409.16040.

Fan-Keng Sun, Christopher I Lang, and Duane S Boning. Adjusting for autocorrelated errors in neural networks for time series. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29806–29819, 2021. arXiv:2101.12578.

Souhaib Ben Taieb and Amir F Atiya. A bias and variance analysis for multistep-ahead time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 27(1):62–76, 2016.

Hao Wang, Lichen Pan, Yuan Shen, Zhichao Chen, Degui Yang, Yifei Yang, Sen Zhang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. FreDF: Learning to forecast in the frequency domain. In *International Conference on Learning Representations*, 2025a. arXiv:2402.02399.

Hao Wang, Licheng Pan, Zhichao Chen, Xu Chen, Qingyang Dai, Lei Wang, Haoxuan Li, and Zhouchen Lin. Time-o1: Time-series forecasting needs transformed label alignment. In *Advances in Neural Information Processing Systems*, 2025b. arXiv:2505.17847.

Hao Wang, Licheng Pan, Yuan Lu, Zhichao Chen, Tianqiao Liu, Shuting He, Zhixuan Chu, Qingsong Wen, Haoxuan Li, and Zhouchen Lin. Quadratic direct forecast for training multi-step time-series forecast models. In *International Conference on Learning Representations*, 2026a. arXiv:2511.00053.

Hao Wang, Licheng Pan, Qingsong Wen, Jialin Yu, Zhichao Chen, Chunyuan Zheng, Xiaoxi Li, Zhixuan Chu, Chao Xu, Mingming Gong, Haoxuan Li, Yuan Lu, Zhouchen Lin, Philip Torr, and Yan Liu. Deep autocorrelation modeling for time-series forecasting: Progress and prospects. *arXiv preprint arXiv:2603.19899*, 2026b.

Jieting Wang, Xiaolei Shang, Feijiang Li, and Furong Peng. RI-Loss: A learnable residual-informed loss for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26277–26284, 2026c.

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.

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 *International Conference on Learning Representations*, 2025c. arXiv:2410.16032.

Zepu Wang, Bowen Liao, and Xuegang (Jeff) Ban. Decorrelating the future: Joint frequency domain learning for spatio-temporal forecasting. *arXiv preprint arXiv:2603.04418*, 2026d.

Mark W Watson. Vector autoregressions and cointegration. In *Handbook of Econometrics*, volume 4, pp. 2843–2915. Elsevier, 1994.

Peter Whittle. Estimation and information in stationary time series. *Arkiv för Matematik*, 2(5): 423–434, 1953.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, 2024.

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.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020.

Yanfeng Yang, Siwei Chen, Pingping Hu, Zhaotong Shen, Yingjie Zhang, Zhuoran Sun, Shuai Li, Ziqi Chen, and Kenji Fukumizu. Conditionally whitened generative models for probabilistic time series forecasting. In *International Conference on Learning Representations*, 2026. arXiv:2509.20928.

Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. FourierGNN: Rethinking multivariate time series forecasting from a pure graph perspective. In *Advances in Neural Information Processing Systems*, 2023a.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain MLPs are more effective learners in time series forecasting. In *Advances in Neural Information Processing Systems*, 2023b.

Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368, 1962.

Arnold Zellner. Estimators for seemingly unrelated regression equations: Some exact finite sample results. *Journal of the American Statistical Association*, 58(304):977–992, 1963.

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*, 2023.

Yuang Zhao, Tianyu Li, Jiadong Chen, Shenrong Ye, Fuxin Jiang, and Xiaofeng Gao. Disentangled parameter-efficient linear model for long-term time series forecasting. In *Database Systems for Advanced Applications (DASFAA)*, pp. 222–238, 2026. arXiv:2411.17257 (2024); introduces SFALoss.

Vincent Zhihao Zheng and Lijun Sun. Multivariate probabilistic time series forecasting with correlated errors. In *Advances in Neural Information Processing Systems*, 2024.

Vincent Zhihao Zheng, Seongjin Choi, and Lijun Sun. Better batch for deep probabilistic time series forecasting. In *International Conference on Artificial Intelligence and Statistics*, pp. 91–99, 2024.

Vincent Zhihao Zheng, Seongjin Choi, and Lijun Sun. Probabilistic traffic forecasting with dynamic regression. *Transportation Science*, 59(4):689–707, 2025. arXiv:2301.06650 (2023), circulated as "Enhancing Deep Traffic Forecasting Models with Dynamic Regression".

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*, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, 2022.

George Zyskind. On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *The Annals of Mathematical Statistics*, 38(4):1092–1109, 1967.

A Extended Related Work

A.1 Architectures for Time Series Forecasting

Statistical predecessors such as VAR (Watson, 1994) and ARIMA (Box & Jenkins, 1976) gave way to neural encoders of input autocorrelation: recurrent models (Salinas et al., 2020; Gu et al., 2022), convolutional models that reshape the series to expose periodicity, such as TimesNet's frequency-guided 2D folding (Wu et al., 2023), ModernTCN's large-kernel design (Luo & Wang, 2024), and TVNet's 3D reshaping into intra-patch, inter-patch and cross-variable axes driven by video-style dynamic convolution (Li et al., 2025), and graph models (Wu et al., 2020; Yi et al., 2023a). The current frontier is contested between Transformers, which scale well but cost more (Nie et al., 2023; Liu et al., 2024a), and MLP or mixing architectures, which are cheaper (Zeng et al., 2023; Wang et al.,

2024). A prominent recent synthesis is TimeMixer++ (Wang et al., 2025c), which downsamples the series across scales, converts each scale into multi-resolution "time images" using the top-K FFT amplitudes, disentangles seasonality and trend by dual-axis attention, and mixes the results across scales and resolutions, reporting state of the art on eight analytical tasks. In parallel, general pre-training has produced forecasting foundation models (Woo et al., 2024; Ansari et al., 2024; Das et al., 2024; Liu et al., 2024b), surveyed by Jin et al. (2026), of which Time-MoE (Shi et al., 2025b) is the largest, scaling a decoder-only backbone to 2.4B parameters with sparsely activated experts, point-wise tokenisation and multi-resolution forecasting heads.

A.2 Forecast Paradigms and Shape-Based Objectives

Multi-step forecasts are produced either iteratively (IF), one step at a time with previous predictions fed back, or directly (DF), all steps at once through a multi-output head (Liu et al., 2022). IF respects label autocorrelation by construction but accumulates error along the rollout, which is severe at long horizons (Taieb & Atiya, 2016); DF is faster, simpler and empirically more accurate (Zhou et al., 2021), and is now essentially universal (Wu et al., 2023; Liu et al., 2024a; Nie et al., 2023). Its squared error targets the conditional mean without assuming independent steps, but weights every step and variable equally, which Sections 2.1 and 4.1 examine. Shape-based objectives replace pointwise error with an alignment-tolerant discrepancy, as in soft-DTW (Cuturi & Blondel, 2017) and DILATE (Le Guen & Thome, 2019). These target temporal misalignment rather than residual correlation, and they alter which forecasts are considered equivalent. TILDE-Q (Lee et al., 2022) adds terms invariant to amplitude and phase distortions, PS Loss (Kudrat et al., 2025) compares forecast and label patch by patch through their correlation, variance and mean, and DBLoss (Qiu et al., 2025) splits the horizon into trend and seasonal parts by exponential moving averages and weights a loss on each. None of them estimates the residual's covariance, and each mixes its term with a point-wise loss, so the time-domain ℓ_1 control of Section 3.2 is a natural baseline for them as well. FreDF (Wang et al., 2025a) keeps DF and changes the metric instead. An appendix of that work explores substituting other fixed bases for the DFT, and our reading is that the choice of basis is not the essential variable. What the decorrelation argument requires is a metric under which residuals are uncorrelated, and Section 3.1 states the conditions under which a fixed basis supplies one, and benchmark data meets neither (Remark 5, Appendix E).

A.3 Error Correlation as a Training Signal

Sun et al. (2021) adjust deterministic neural forecasters for autocorrelated errors by learning an autocorrelation coefficient $\hat{\rho}$ jointly with the model and applying the transform $X_t - \hat{\rho}X_{t-1} = f(X_{t-1} - \hat{\rho}X_{t-2}, \ldots; \theta)$ to inputs and targets alike, benchmarking against the classical Prais–Winsten procedure (Prais & Winsten, 1954). That transform is GLS whitening under an AR(1) error model, so whitening a neural forecaster's residual is an established strategy. Zheng et al. (2025) extend the idea to traffic forecasting via dynamic regression, and Zheng et al. (2024); Zheng & Sun (2024) carry it into probabilistic forecasting, the latter learning a low-rank-plus-diagonal contemporaneous covariance together with cross-lag structure and training by the full Gaussian likelihood, log-determinant included. In probabilistic forecasting more broadly, GP-Copula (Salinas et al., 2019), DeepVAR and normalising-flow or diffusion forecasters (Rasul et al., 2021) *predict* a joint predictive distribution, so there the covariance is an output of the model and part of the deliverable.

A.4 Weighting the Frequency Loss: SFALoss and FreST

Closest to our prescription in form, though not in derivation, is SFALoss (Zhao et al., 2026), which weights a frequency-domain ℓ_1 loss element-wise by the model's own learned spectral filter and fuses it with a temporal MSE under the same α convention we adopt. Per-frequency weighting of the forecast loss is therefore established. Section 3.1 adds the argument that equal weights make the transform inert and identifies the weight that GLS prescribes, λ_k^{-1}, derived from the residual spectrum rather than learned as a relevance filter. Zhao et al. (2026) also detach their weights, so that the module does not learn to suppress frequencies it cannot predict. Our detachment in Section 4.2 serves a related purpose: $\Sigma \mapsto e^\top \Sigma^{-1} e$ has infimum zero as Σ is inflated, so a covariance optimised jointly under this loss alone would grow without bound. The decorrelation framing has since been carried to spatio-temporal forecasting by FreST (Wang et al., 2026d), whose spectral terms are

combined by a learned softmax over three transforms rather than weighted per frequency, so the per-frequency weight that Section 3.1 identifies as the missing ingredient is absent there as well.

A.5 Neighbours from 2025–2026

Two follow-ups by FreDF's authors are the closest prior work on the temporal axis. Time-o1 (Wang et al., 2025b) replaces the DFT by the right singular vectors of the training-label matrix, keeps the leading fraction of components and aligns them under ℓ_1, noting that the DFT decorrelates only as $\mathrm{T} \to \infty$ (our Remark 5). The transform is fixed before training and the weighting is by selection, not by residual power. Quadratic Direct Forecast (QDF) (Wang et al., 2026a) trains under $\operatorname{tr}\left[(Y - g_\theta(L))^\top \Sigma^{-1} (Y - g_\theta(L))\right]$ with one $\mathrm{T} \times \mathrm{T}$ matrix per horizon shared across variables, obtained by bilevel optimisation: the outer objective is the same quadratic form on a held-out split, differentiated through the trained model into a Cholesky factor of Σ. Its theorem identifies the quadratic form with the Gaussian negative log-likelihood. The object is the same as our temporal factor. The estimator is not. QDF learns Σ through the model on held-out data; WhiteDF takes it from training residuals with shrinkage and a stop-gradient. The shrinkage makes DF the $\gamma = 1$ point of the family, and a plug-in estimate can fail in the way Section 5.3 reports on ECL. QDF reports gains on ECL and on two PEMS traffic sets with TQNet as its main backbone and PDF, Fredformer and iTransformer in a generalisation study. On iTransformer our temporal factor does not gain on ECL (Section 5.3); Appendix F.6 runs the two side by side in QDF's own repository on ETTh1 and on ECL, where it does. Neither work models the variable axis or asks when the correction fails. Four further neighbours are close in form, and none of them estimates and inverts the residual covariance of a point forecaster. CvLoss (Ding et al., 2026) diagnoses the same gap on the variable axis that Section 4.1 does, that point-wise objectives leave the cross-variable structure of the residual unconstrained, and closes it with a plug-in penalty on the differences of residuals along the edges of a cross-variable graph, over forecast patches. Read as a quadratic form on the residual, that is a precision structure supplied by the graph, the variable-axis counterpart of the fixed DFT basis on the temporal axis. WhiteDF estimates the structure from the residuals instead, so the shrinkage γ and the ratio $\mathrm{D}/(n\mathrm{T})$ enter. OLMA (Shi et al., 2025a) applies FreDF's ℓ_1 frequency supervision along the variable axis as well as the temporal one and motivates it by an entropy argument; Corollary 1 covers any unitary transform, so under a squared norm the channel-axis DFT would be inert too, and a DFT over the variable index presupposes a circulant covariance over an arbitrary ordering of the variables. RI-Loss (Wang et al., 2026c) replaces MSE by a Hilbert–Schmidt independence criterion between the residual sequence and a random series, a kernel account of residual structure that neither estimates nor inverts a covariance. Conditionally whitened generative models (Yang et al., 2026) replace the terminal Gaussian of a diffusion or flow forecaster by one with a jointly learned conditional mean and sliding-window covariance. This is whitening in the probabilistic line, where the covariance is learned with the likelihood and belongs to the model. WhiteDF instead applies a detached metric to a point criterion. Wang et al. (2026b) survey the objective side of this literature under the heading of label autocorrelation.

A.6 Classical Antecedents

The classical statistical antecedents are Aitken's generalised least squares (Aitken, 1936) and the Gauss–Markov theory around it. Separable (Kronecker) covariance estimation for matrix-valued data goes back to Dutilleul (1999), and the shrinkage we use is that of Ledoit & Wolf (2004). Kronecker factorisation also appears in K-FAC (Martens & Grosse, 2015) and its relatives, but there it preconditions the optimiser using curvature, whereas here it structures a metric on the error, so the two share only the algebraic form.

B Proofs

B.1 Proof of Proposition 1

Both criteria are quadratic forms in the residual, $\mathcal{L}^{(\mathrm{tmp})} = e^\top \mathbf{I} e$ and $\mathcal{L}^{(\mathrm{gls})}_\Sigma = e^\top \Sigma^{-1} e$. Two quadratic forms $e^\top A_1 e$ and $e^\top A_2 e$ with A_1, A_2 symmetric satisfy $e^\top A_1 e = c\, e^\top A_2 e + b$ for all $e \in \mathbb{R}^{\mathrm{TD}}$ with $c > 0$ if and only if $b = 0$ and $A_1 = c A_2$: setting $e = 0$ gives $b = 0$, and the identity

$e^\top(A_1 - cA_2)e = 0$ for all e forces $A_1 - cA_2 = 0$ by symmetry. Hence $\mathbf{I} = c\,\Sigma^{-1}$, i.e. $\Sigma = c\,\mathbf{I}$. Conversely $\Sigma = c\mathbf{I}$ gives $\mathcal{L}_\Sigma^{(\mathrm{gls})} = c^{-1}\mathcal{L}^{(\mathrm{tmp})}$. □

B.2 Proof of Theorem 1

With $\Sigma = \mathbf{I}_\mathrm{D} \otimes \Sigma_t$ and $e = \mathrm{vec}(E)$ stacking columns, the quadratic form separates over variables: $e^\top\Sigma^{-1}e = \sum_{d=1}^{\mathrm{D}} E_{:,d}^\top \Sigma_t^{-1} E_{:,d}$. A real symmetric circulant matrix is diagonalised by the unitary DFT, $\Sigma_t = F^*\Lambda F$ with Λ real, positive and diagonal, so $\Sigma_t^{-1} = F^*\Lambda^{-1}F$ and

$$E_{:,d}^\top \Sigma_t^{-1} E_{:,d} = (FE_{:,d})^*\Lambda^{-1}(FE_{:,d}) = \sum_{k=1}^{\mathrm{T}} \lambda_k^{-1}\left|\mathcal{F}(E_{:,d})_k\right|^2,$$

where $\mathcal{F}(\cdot) = F(\cdot)$ is the normalised DFT. Summing over d gives (6). □

B.3 Proof of Corollary 1

F is unitary, so $\|Fv\|_2^2 = v^*F^*Fv = \|v\|_2^2$ for every v. Applying this column-wise, $\|\mathcal{F}(E)\|_F^2 = \sum_d \|FE_{:,d}\|_2^2 = \sum_d \|E_{:,d}\|_2^2 = \|E\|_F^2$. This is the $\Lambda = \lambda\mathbf{I}$ case of Theorem 1 up to the constant λ^{-1}. □

Remark 1 *The corollary applies to the squared frequency loss. It does not apply to the ℓ_1 frequency loss $|\mathcal{F}(\hat{Y}) - \mathcal{F}(Y)|_1$ used by Wang et al. (2025a), since the ℓ_1 norm is not unitarily invariant, so what separates FreDF from DF is measured by experiment. In our implementation on the development cell most of the difference is reproduced by the ℓ_1 norm in the time domain and a small remainder is attributable to the basis (Appendix F.1).*

Half-spectrum implementations. *The corollary is stated for the full unitary DFT, and our experiments use it. Implementations that keep only the half spectrum of a real signal, as FreDF's does (its authors note that "the forecast length in the frequency domain is halved"), count the DC and Nyquist bins once and every other bin once out of its conjugate pair, so a squared half-spectrum loss is not equal-weight. It gives those two bins twice the relative weight of the rest, and differs from DF by that reweighting alone.*

B.4 Proof of Theorem 2

Unbiasedness of both estimators is immediate from $\mathbb{E}[y] = X\beta$ and the fact that both are of the form My with $MX = \mathbf{I}$. For the covariance ordering, set

$$A = V^{1/2}X(X^\top X)^{-1}, \qquad B = V^{-1/2}X(X^\top V^{-1}X)^{-1},$$

so that $A^\top A = \mathrm{Cov}(\hat{\beta}_{\mathrm{ols}})$ and $B^\top B = \mathrm{Cov}(\hat{\beta}_{\mathrm{gls}})$. Then

$$A^\top B = (X^\top X)^{-1}X^\top V^{1/2}V^{-1/2}X(X^\top V^{-1}X)^{-1} = (X^\top V^{-1}X)^{-1} = B^\top B,$$

and symmetrically $B^\top A = B^\top B$. Therefore

$$0 \preceq (A-B)^\top(A-B) = A^\top A - A^\top B - B^\top A + B^\top B = A^\top A - B^\top B,$$

which is the claimed ordering. Equality holds iff $A = B$, i.e. $VX(X^\top X)^{-1} = X(X^\top V^{-1}X)^{-1}$, i.e. $VX = XC$ with $C = (X^\top V^{-1}X)^{-1}(X^\top X)$ nonsingular. Conversely, substituting $VX = XC$ into either covariance expression yields $C(X^\top X)^{-1}$, the same matrix. The condition is due to Zyskind (1967), Rao (1967) and Kruskal (1968). See Puntanen & Styan (1989) for a survey. □

B.5 Proof of Corollary 3

Order the stacked observations by output coordinate, so that $y = (y^{(1)}; \ldots; y^{(\mathrm{TD})})$ with $y^{(j)} \in \mathbb{R}^N$. If every output coordinate is regressed on the same $N \times \mathrm{HD}$ design Z, then $X = \mathbf{I}_{\mathrm{TD}} \otimes Z$ and $V = \Sigma \otimes \mathbf{I}_N$. Using $(A \otimes B)(C \otimes D) = AC \otimes BD$,

$$VX = (\Sigma \otimes \mathbf{I}_N)(\mathbf{I}_{\mathrm{TD}} \otimes Z) = \Sigma \otimes Z = (\mathbf{I}_{\mathrm{TD}} \otimes Z)(\Sigma \otimes \mathbf{I}_{\mathrm{HD}}) = XC$$

with $C = \Sigma \otimes \mathbf{I}_{\mathrm{HD}}$ nonsingular. The invariance condition of Theorem 2 holds for every positive definite Σ, so the two estimators coincide. This is Zellner's observation that seemingly unrelated regressions with identical regressors gain nothing from joint estimation (Zellner, 1962). □

B.6 Proof of Corollary 2

Order the observations by variable, so $X = \mathrm{blkdiag}(X_1, \ldots, X_{\mathrm{D}})$ where X_d collects the T output coordinates of variable d regressed on that variable's own history, and $V = \Sigma_c \otimes \Sigma_t \otimes \mathbf{I}_N$ in the matching ordering. If $\Sigma_c = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{\mathrm{D}}^2)$ then V is block diagonal with blocks $\sigma_d^2(\Sigma_t \otimes \mathbf{I}_N)$, so $VX = \mathrm{blkdiag}\big(\sigma_d^2(\Sigma_t \otimes \mathbf{I}_N)X_d\big)$. If in addition each X_d is unconstrained, its T output coordinates share the design of variable d and Corollary 3 applies within the block, giving $(\Sigma_t \otimes \mathbf{I}_N)X_d = X_d C_d$. Hence $VX = XC$ with $C = \mathrm{blkdiag}(\sigma_d^2 C_d)$ and the estimators coincide.

If Σ_c has a nonzero off-diagonal entry $(\Sigma_c)_{dd'}$ with $d \neq d'$, then VX acquires a block in position (d, d') equal to $(\Sigma_c)_{dd'}(\Sigma_t \otimes \mathbf{I}_N)X_{d'}$, whereas every off-diagonal block of XC is $X_d C_{dd'}$, whose columns lie in $\mathrm{col}(X_d)$. Equality therefore requires $\mathrm{col}\big((\Sigma_t \otimes \mathbf{I}_N)X_{d'}\big) \subseteq \mathrm{col}(X_d)$ for every such pair, which fails whenever the per-variable designs $\{X_d\}$ span distinct column spaces, which is the generic situation, since distinct variables have distinct histories. In that case the covariance difference in Theorem 2 is positive semidefinite and non-zero, so some linear combinations of the parameters, not necessarily every parameter, are estimated with strictly smaller variance. □

Remark 2 (Shared weights: when the temporal factor cancels) *The corollary is stated for per-variable unconstrained blocks. Channel-independent models that also share one map across variables, as DLinear does in its non-individual configuration, are a further special case, and for them the temporal axis drops out of the temporal-only objective and of the pure two-factor GLS objective, but not of the mixed two-factor objective that Section 4.2 trains. Write* $\hat{Y}_n = WX_n$ *with* $X_n \in \mathbb{R}^{\mathrm{H}\times\mathrm{D}}$ *holding the* D *histories as columns and* $W \in \mathbb{R}^{\mathrm{T}\times\mathrm{H}}$ *shared. Then* $\mathrm{vec}(WX_n) = (X_n^\top \otimes \mathbf{I}_{\mathrm{T}})\,\mathrm{vec}(W)$*, so the stacked design is* $X = \mathrm{stack}_n(X_n^\top \otimes \mathbf{I}_{\mathrm{T}})$ *and under temporal whitening alone,* $V = \mathbf{I} \otimes (\mathbf{I}_{\mathrm{D}} \otimes \Sigma_t)$*,*

$$VX = \mathrm{stack}_n\big(X_n^\top \otimes \Sigma_t\big) = \mathrm{stack}_n\big(X_n^\top \otimes \mathbf{I}_{\mathrm{T}}\big)(\mathbf{I}_{\mathrm{H}} \otimes \Sigma_t) = XC,$$

so the invariance condition of Theorem 2 holds for every Σ_t*. Temporal whitening alone cannot move a shared-weight linear channel-independent forecaster. Directly, with* $A = \Sigma_t^{-1}$*, the stationarity condition of the mixed temporal-only objective* $\sum_n \big[\alpha\|\Sigma_t^{-1/2}(Y_n - WX_n)\|_F^2 + (1-\alpha)\|Y_n - WX_n\|_F^2\big]$ *is* $\big(\alpha A + (1-\alpha)\mathbf{I}\big)\big[\sum_n Y_n X_n^\top - W\sum_n X_n X_n^\top\big] = 0$*, and the invertible left factor cancels for every* $\alpha \in [0,1]$*. The channel factor does not cancel: with* $V = \mathbf{I} \otimes (\Sigma_c \otimes \Sigma_t)$ *one needs* $\Sigma_c X_n^\top = X_n^\top C_1$ *for every* n*, which fails for generic histories. Writing* $B = \Sigma_c^{-1}$*, the pure two-factor GLS objective (*$\alpha = 1$*) has stationarity condition* $A\big[\sum_n Y_n B X_n^\top - W\sum_n X_n B X_n^\top\big] = 0$*, so* A *cancels again and* $W = \big[\sum_n Y_n B X_n^\top\big]\big[\sum_n X_n B X_n^\top\big]^{-1}$ *depends on* Σ_c *alone. For the mixed two-factor objective with* $0 < \alpha < 1$ *the condition is* $\alpha A\big(G_B - WS_B\big) + (1-\alpha)\big(G - WS\big) = 0$ *with* $G_B = \sum_n Y_n B X_n^\top$*,* $S_B = \sum_n X_n B X_n^\top$*,* $G = \sum_n Y_n X_n^\top$ *and* $S = \sum_n X_n X_n^\top$*, a Sylvester-type equation in which* A *does not cancel. Once the variable factor is present, the temporal factor enters the solution as well. Appendix E verifies the temporal-only cancellation numerically to* 10^{-13} *on all seven benchmarks. For such a backbone, therefore, whatever* WhiteDF$_t$ *does during training is an effect of the optimiser's path and not of the estimator, which is one reason we report* WhiteDF$_t$ *and* WhiteDF$_{t+c}$ *separately and the reason Section 4.3 reads the DLinear numbers the way it does. For* WhiteDF$_{t+c}$ *the estimator itself moves, and the two factors cannot be attributed separately.*

B.7 Proof of Proposition 2

For matrices of compatible shape, $(B^\top \otimes A)\,\mathrm{vec}(X) = \mathrm{vec}(AXB)$. With $\Sigma = \Sigma_c \otimes \Sigma_t$ and both factors symmetric positive definite, $\Sigma^{-1/2} = \Sigma_c^{-1/2} \otimes \Sigma_t^{-1/2}$, and taking $A = \Sigma_t^{-1/2}$ and $B = \Sigma_c^{-1/2} = B^\top$ gives $\Sigma^{-1/2}\,\mathrm{vec}(E) = \mathrm{vec}(\Sigma_t^{-1/2} E \Sigma_c^{-1/2})$. Since $\|\mathrm{vec}(M)\|_2 = \|M\|_F$, (3) follows. □

B.8 Proof of Proposition 3

At $\gamma = 1$, (4) gives $\Sigma_t^{(1)} = \frac{\mathrm{tr}\,\hat{S}_t}{\mathrm{T}}\mathbf{I}_{\mathrm{T}}$ and $\Sigma_c^{(1)} = \frac{\mathrm{tr}\,\hat{S}_c}{\mathrm{D}}\mathbf{I}_{\mathrm{D}}$, so by Proposition 2 the whitened loss is $\big(\frac{\mathrm{tr}\,\hat{S}_t}{\mathrm{T}}\frac{\mathrm{tr}\,\hat{S}_c}{\mathrm{D}}\big)^{-1}\|E\|_F^2$, a positive multiple of $\mathcal{L}^{(\mathrm{tmp})}$. Trace normalisation fixes the multiple at one.

For the second statement, projecting $\hat{S}_t$ onto the symmetric circulant cone and replacing its spectrum by a constant yields $\Sigma_t \propto \mathbf{I}$ (a circulant matrix with flat spectrum is a multiple of the identity), and Theorem 1 with $\Lambda \propto \mathbf{I}$ together with Corollary 1 identifies the resulting objective with the equal-weight spectral loss. □

C Additional Theory

This appendix states the formal results that Sections 3 and 4 summarise (proofs in Appendix B), with the second-moment decomposition, the Laplace scale of an ℓ_1 frequency loss, two departures from the circulant model, a two-output example and the objective of the detached update.

C.1 Spectral Representation of GLS

Theorem 1 (Spectral representation of GLS under a circulant covariance) *Let $\Sigma = \mathbf{I}_\mathrm{D} \otimes \Sigma_t$ with $\Sigma_t \in \mathbb{R}^{\mathrm{T}\times\mathrm{T}}$ real, symmetric and circulant, and let F denote the unitary DFT matrix, so that $\Sigma_t = F^* \Lambda F$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_\mathrm{T})$, $\lambda_k > 0$ the residual power spectrum. Then*

$$\mathcal{L}_\Sigma^{(\mathrm{gls})} = \sum_{d=1}^{\mathrm{D}} \sum_{k=1}^{\mathrm{T}} \lambda_k^{-1} \left| \mathcal{F}(E_{:,d})_k \right|^2. \tag{6}$$

That is, GLS under a circulant temporal covariance is a squared spectral error weighted by the inverse residual power at each frequency.

Equation (6) is the quadratic term of the Whittle likelihood of the residual process (Whittle, 1953) under a circulant approximation. Neither a flat spectrum nor the circulant model holds on benchmark data (Remark 5, Appendix E).

C.2 Aitken's Theorem and Unconstrained Channel Mixing

Theorem 2 (Aitken; efficiency and the invariance condition) *Both the OLS estimator $\hat{\beta}_\mathrm{ols} = (X^\top X)^{-1} X^\top y$ and the GLS estimator $\hat{\beta}_\mathrm{gls} = (X^\top V^{-1} X)^{-1} X^\top V^{-1} y$ are unbiased, and*

$$\mathrm{Cov}(\hat{\beta}_\mathrm{gls}) = (X^\top V^{-1} X)^{-1} \preceq (X^\top X)^{-1} X^\top V X (X^\top X)^{-1} = \mathrm{Cov}(\hat{\beta}_\mathrm{ols}), \tag{7}$$

with equality if and only if the column space of X is an invariant subspace of V, i.e. $VX = XC$ for some nonsingular C.

Corollary 3 (Unconstrained channel mixing: whitening is a no-op) *If the predictor is an unconstrained linear map from the full context to the full horizon, so that every one of the* TD *output coordinates is regressed on the same features, then $X = \mathbf{I}_\mathrm{TD} \otimes Z$ and the invariance condition holds for every V of the form $\Sigma \otimes \mathbf{I}_N$. The OLS and GLS minimisers coincide, and no reweighting of the loss, whether in the time domain, the frequency domain, or any other basis, can alter the fitted predictor.*

Consecutive windows overlap, so $V = \Sigma \otimes \mathbf{I}_N$ is an approximation: the invariance statements are algebraic and hold exactly, while the efficiency ordering holds approximately.

C.3 DF as a Boundary Case

Proposition 3 (DF as a boundary case) *At $\gamma = 1$, $\Sigma_t^{(1)} \propto \mathbf{I}_\mathrm{T}$ and $\Sigma_c^{(1)} \propto \mathbf{I}_\mathrm{D}$, and $\mathcal{L}^{(\mathrm{wht})}$ equals $\mathcal{L}^{(\mathrm{tmp})}$ up to a positive scalar. If in addition $\hat{S}_t$ is projected onto the symmetric circulant cone and its spectrum flattened, $\mathcal{L}^{(\mathrm{wht})}$ recovers the equal-weight spectral objective of Corollary 1.*

The second clause says that flattening a circulant estimate returns the $\gamma = 1$ point, since by Corollary 1 the equal-weight spectral objective is DF itself.

C.4 Covariance or Second Moment?

The quantity a practitioner can estimate is not the same as the quantity the theory calls for, and the gap matters. Write $\bar{g}(L) := \mathbb{E}[Y \mid L]$ for the true conditional mean and decompose the residual into an irreducible and a reducible part,

$$e \;=\; \underbrace{\big(Y - \bar{g}(L)\big)}_{\text{noise } \eta} \;+\; \underbrace{\big(\bar{g}(L) - g_\theta(L)\big)}_{\text{bias } b_\theta(L)}. \tag{8}$$

Conditionally on L the forecast is deterministic, so $\mathrm{Cov}(e \mid L) = \mathrm{Cov}(Y \mid L) = \mathrm{Cov}(\eta \mid L)$ carries no dependence on θ. Under $Y \mid L \sim \mathcal{N}(\bar{g}(L), \Sigma)$ the $\log\det\Sigma$ term of the negative log-likelihood is therefore constant in θ, and the negative log-likelihood reduces to $\frac{1}{2}\mathcal{L}_\Sigma^{(\mathrm{gls})}$ up to that constant. But an estimator pooled over training windows, as in Section 4.2, targets the *second moment*

$$M_\theta := \mathbb{E}\big[ee^\top\big] = \underbrace{\mathbb{E}_L\big[\,\mathrm{Cov}(\eta \mid L)\big]}_{\bar{\Sigma}_\eta} + \underbrace{\mathbb{E}_L\big[b_\theta(L)\, b_\theta(L)^\top\big]}_{B_\theta \succeq 0}, \tag{9}$$

the cross term vanishing because $\mathbb{E}[\eta \mid L] = 0$. So M_θ exceeds $\bar{\Sigma}_\eta$ by the second moment of the model's bias, and unlike $\bar{\Sigma}_\eta$ it does depend on θ.

Remark 3 (Whitening by the second moment) *Whitening by M_θ rather than $\bar{\Sigma}_\eta$ gives relatively less weight to the directions in which the model is currently most wrong. When that error is irreducible noise, generalised least squares prescribes this. When it is reducible bias, it is the opposite of what one wants, since those directions are still worth fitting. Three observations bound the damage, and each one motivated a design choice in Section 4.2. First, for a well-specified model $B_\theta \to 0$ as training converges, so $M_\theta \to \bar{\Sigma}_\eta$ and the metric approaches the correct one. Second, B_θ is largest at initialisation, so we start the estimate at the identity and let it move only as evidence accumulates. Third, retaining a share $1 - \alpha$ of the unwhitened objective, and shrinking by γ, both bound below the weight any direction can receive. For a* misspecified *model B_θ does not vanish, and whether permanently downweighting the directions the model cannot fit concentrates capacity usefully or abandons real structure is decided empirically, and Appendix E measures it for a closed-form channel-independent linear forecaster.*

C.5 The Laplace Scale of an ℓ_1 Frequency Loss

Remark 4 (The Laplace scale) *FreDF measures its frequency term under ℓ_1, which is not unitarily invariant, so Corollary 1 does not apply to it. $\sum_k |F_k|$ is the negative log-likelihood of independent Laplace frequency coordinates with a common scale. Independence is strictly stronger than the decorrelation that FreDF's theorem delivers, and the Laplace model has a scale per coordinate, $\sum_k |F_k|/b_k$, so taking b_k constant reinstates the flat-spectrum assumption that Theorem 1 isolates. Under either likelihood, Gaussian or Laplace, the per-frequency weight is what an equal-weight loss omits, entering as λ_k^{-1} or as b_k^{-1}. The frequency coordinates are complex, so a Laplace model on their moduli must also specify the radial density and the conjugate symmetry; b_k^{-1} is therefore a coordinate-scale weighting suggested by that model, not a maximum-likelihood weight for a specified complex density. Supplying b_k^{-1} does not improve the ℓ_1 criterion at any shrinkage tried (Appendix F.1).*

C.6 Two Further Departures from the Circulant Model

Remark 5 (Two further departures from the circulant model) *Beyond flatness, the covariance model of Theorem 1 fails on benchmark data in two ways. First, real series exhibit regime shifts, trends and event shocks, so residuals are not wide-sense stationary and Σ_t is not circulant at the finite horizons $\mathrm{T} \in \{96, 192, 336, 720\}$ that benchmarks use. Even a stationary process has a Toeplitz, not circulant, finite-window covariance, and the asymptotic argument requires $\mathrm{T} \to \infty$. Second, the choice $\Sigma = \mathbf{I}_\mathrm{D} \otimes \Sigma_t$ assumes residuals are uncorrelated across variables and identically distributed across them, which is untenable on datasets whose variables are 321 clients (ECL) or 862 road sensors (Traffic) driven by common exogenous factors.*

C.7 A Constrained Example in Which the True Covariance Raises Test MSE

Aitken's theorem orders estimator covariances. The test metric of Section 5 is the plain MSE. The following example, which involves no estimation error in the covariance, separates the two.

Proposition 4 (Two outputs, one shared parameter) *Let $Y_1 = \mu + \varepsilon_1$ and $Y_2 = \mu + \delta + \varepsilon_2$ with independent noises of variances $a, b > 0$, observed on n independent training samples, and constrain the predictor to forecast both outputs by one scalar θ. Let $\hat{\theta}_{\mathrm{ols}}$ minimise the summed squared error and $\hat{\theta}_{\mathrm{gls}}$ the error weighted by the true precisions $1/a$ and $1/b$. For the test risk $R(\hat{\theta}) = \frac{1}{2}\,\mathbb{E}\big[(Y_1 - \hat{\theta})^2 + (Y_2 - \hat{\theta})^2\big]$ on an independent sample, averaged over the training sample,*

$$R(\hat{\theta}_{\mathrm{gls}}) - R(\hat{\theta}_{\mathrm{ols}}) = \frac{(a-b)^2}{4(a+b)^2}\Big(\delta^2 - \frac{a+b}{n}\Big). \tag{10}$$

Proof. The estimators are $\hat{\theta}_{\mathrm{ols}} = (\bar{Y}_1 + \bar{Y}_2)/2$ and $\hat{\theta}_{\mathrm{gls}} = (b\bar{Y}_1 + a\bar{Y}_2)/(a+b)$. The population test risk is minimised at $\theta^\star = \mu + \delta/2$, and for any estimator $R(\hat{\theta}) = \frac{1}{2}(a+b) + \frac{1}{4}\delta^2 + \mathbb{E}(\hat{\theta} - \theta^\star)^2$. The OLS estimator is unbiased for $\theta^\star$ with variance $(a+b)/(4n)$. The GLS estimator has bias $\delta(a-b)/(2(a+b))$ and variance $ab/((a+b)n)$. Subtracting and using $ab/(a+b) - (a+b)/4 = -(a-b)^2/(4(a+b))$ gives (10). □

The GLS estimator has the smaller variance, as Theorem 2 promises, and the larger test MSE whenever $\delta^2 > (a+b)/n$. Within a class that cannot fit both outputs, the weighted objective selects the member that favours the low-noise output, which is not the member the unweighted test metric prefers. The effect is the mismatch between training metric and test metric under misspecification, and the covariance in it is known exactly. With a separate parameter per output, the unconstrained case of Corollary 3, both estimators recover the means and the gap vanishes. The residual second moment of Section 4.2 folds the misfit δ into its weights (Appendix C.4), so an online estimate does not escape the trade-off. The example shows that a correct covariance is not sufficient for a gain in test MSE, which is consistent with the failures of the variable factor on ETTh1 (Section 4.3) and of whitening on ECL (Section 5.3). Equation (10) is checked numerically (Appendix K).

C.8 What the Detached Whitened Update Descends

Consider the temporal factor alone, the pure whitened term ($\alpha = 1$), and, in place of the EMA, the full-training-set second moment $M_\theta = \frac{1}{ND}\sum_n E_n E_n^\top \succ 0$ recomputed at every step, without shrinkage or trace normalisation. The whitened loss is $\mathrm{tr}(PM_\theta)$ with $P = M_\theta^{-1}$ held fixed by the stop-gradient, and by Jacobi's formula

$$\nabla_\theta\ \mathrm{tr}(PM_\theta)\big|_{P = M_\theta^{-1}\ \text{fixed}} = \mathrm{tr}\left(M_\theta^{-1}\nabla_\theta M_\theta\right) = \nabla_\theta \log\det M_\theta. \tag{11}$$

Each detached update is therefore a step of gradient descent on $\log\det M_\theta$, the log-volume of the residual second-moment ellipsoid, where DF descends $\mathrm{tr}\, M_\theta$: the former lowers the geometric mean of the eigenvalues of M_θ, the latter the arithmetic mean, and a relative reduction in a low-power direction counts as much as one in a high-power direction. The mixed objective descends $\alpha \log\det M_\theta + (1-\alpha)\,\mathrm{tr}\, M_\theta$ up to the positive scalar of the trace normalisation, and with shrinkage $\Sigma_\gamma = (1-\gamma)M_\theta + \gamma(\mathrm{tr}\, M_\theta/\mathrm{T})\mathbf{I}$ the identity becomes $\nabla \log\det \Sigma_\gamma = (1-\gamma)\, g_{\mathrm{wht}} + \gamma\,(\mathrm{tr}\,\Sigma_\gamma^{-1}/\mathrm{T})\, g_{\mathrm{mse}}$, with $g_{\mathrm{wht}} = \mathrm{tr}(\Sigma_\gamma^{-1}\nabla M_\theta)$ and $g_{\mathrm{mse}} = \nabla\,\mathrm{tr}\, M_\theta$: the shrunk whitened gradient and the MSE gradient are the two components of a log-det descent on the shrunk matrix. Both identities are checked numerically (Appendix K). Read this way the detached update is a form of iterated feasible GLS: $\log\det M_\theta$ is, up to constants, the Gaussian negative log-likelihood with the covariance profiled out. For the online procedure this reading is a hypothesis: the EMA replaces M_θ by a lagged minibatch average, and the variable factor is estimated from the unwhitened residual, so the two-factor identity would need the alternating step of Appendix D.1. Because this M_θ contains the model's bias (Remark 3), whether the low-power directions the update favours contain reducible error decides the sign of the effect (Appendix C.4, Proposition 4), and the reading says how to test it: a control that keeps the spectrum of the estimated factor and randomises its eigenvectors separates the direction of the weighting from its spectrum, and Appendix F.10 runs it.

D ALGORITHM AND COST

For a training minibatch $\mathcal{B}$ with residual matrices $\{E_b\}$, the two factors of Section 4.2 are updated as

$$\hat{S}_t \leftarrow \beta\,\hat{S}_t + (1-\beta)\,\frac{1}{|\mathcal{B}|\,\mathrm{D}}\sum_{b\in\mathcal{B}} E_b E_b^\top, \tag{12}$$

$$\hat{S}_c \leftarrow \beta\,\hat{S}_c + (1-\beta)\,\frac{1}{|\mathcal{B}|\,\mathrm{T}}\sum_{b\in\mathcal{B}} E_b^\top E_b, \tag{13}$$

initialised at the identity. Algorithm 1 gives the training loop of Section 4.2.

Algorithm 1 Training with WHITEDF. Additions to a standard DF trainer are lines 5–9.

Require: training set $\mathcal{D}$; forecaster g_θ; fusion weight α; shrinkage γ; EMA rate β; refresh period R

1: $\hat{S}_t \leftarrow \mathbf{I}_\mathrm{T},\ \hat{S}_c \leftarrow \mathbf{I}_\mathrm{D},\ W_t \leftarrow \mathbf{I}_\mathrm{T},\ W_c \leftarrow \mathbf{I}_\mathrm{D}$
2: **for** step $s = 1, 2, \ldots$ **do**
3: sample minibatch $\{(L_b, Y_b)\}_{b\in\mathcal{B}}$ from $\mathcal{D}$
4: $E_b \leftarrow Y_b - g_\theta(L_b)$ for all $b \in \mathcal{B}$
5: update $\hat{S}_t, \hat{S}_c$ by (12)–(13) $\{\mathcal{O}(|\mathcal{B}|(\mathrm{T}^2\mathrm{D} + \mathrm{T}\mathrm{D}^2))\}$
6: **if** $s \equiv 0 \pmod R$ **then**
7: shrink and trace-normalise by (4)
8: $W_t \leftarrow (\Sigma_t^{(\gamma)})^{-1/2},\ W_c \leftarrow (\Sigma_c^{(\gamma)})^{-1/2}$ via symmetric eigendecomposition $\{\mathcal{O}(\mathrm{T}^3 + \mathrm{D}^3)$, amortised over R steps$\}$
9: **end if**
10: $\mathcal{L} \leftarrow \frac{1}{|\mathcal{B}|}\sum_b \left[\alpha\|W_t E_b W_c\|_F^2 + (1-\alpha)\|E_b\|_F^2\right]$ $\{W_t, W_c$ detached$\}$
11: update θ by backpropagation on $\mathcal{L}$
12: **end for**

D.1 DESIGN NOTES

Section 4.2 states the design. This section records why each part is there. The full covariance is out of reach. On Traffic with $\mathrm{T} = 720$ and $\mathrm{D} = 862$, Σ would hold 3.9×10^{11} entries, and the Kronecker form is a far weaker restriction than either isotropy ($\Sigma \propto \mathbf{I}$, DF) or temporal-only circularity ($\Sigma = \mathbf{I}_\mathrm{D} \otimes \Sigma_t$ circulant, frequency alignment). The two factors have distinct jobs: left-multiplication by $\Sigma_t^{-1/2}$ reweights error across forecast steps, down-weighting the directions along the horizon in which the residuals are already large and mutually predictable, long-horizon drift being the obvious instance, relative to the directions in which the data pins the forecast down. Right-multiplication by $\Sigma_c^{-1/2}$ reweights error across variables, the factor Corollary 2 singles out. The moving average serves two purposes at once: it supplies the sample size a single minibatch cannot, and it makes the metric track the model, so the reweighting evolves over training instead of being fixed a priori. Trace normalisation removes one arbitrary scale from each factor, so that the whitened term does not drift with the overall residual variance and interact with the learning rate. Because γ already interpolates to the identity metric, α and γ are partially redundant by design. Both are kept because they control different things, γ the conditioning of the estimate and α the trust placed in it, and Appendix I reports their joint sensitivity.

Alternating (flip-flop) estimation. Equations (12)–(13) estimate each factor while treating the other as isotropic. The maximum-likelihood estimator of the matrix normal model (Dutilleul, 1999) instead alternates $\hat{S}_t \propto \sum_b E_b \hat{S}_c^{-1} E_b^\top$ and $\hat{S}_c \propto \sum_b E_b^\top \hat{S}_t^{-1} E_b$, and the alternation of Oberhofer & Kmenta (1974), which re-estimates Σ from the current residuals and refits under it, converges to the joint maximum-likelihood estimate. Our updates are one step of each with the other factor at the identity, and every result in the paper uses them. With detachment, shrinkage, trace normalisation and finite-step neural updates, their relation to the joint estimate is left open.

D.2 COST

Per training step WHITEDF adds two matrix products against each residual, at $\mathcal{O}(\mathrm{T}^2\mathrm{D} + \mathrm{T}\mathrm{D}^2)$ per window, and two EMA updates of the same order. The eigendecompositions cost $\mathcal{O}(\mathrm{T}^3 + \mathrm{D}^3)$ and are refreshed every $R = 100$ steps. Memory is $\mathcal{O}(\mathrm{T}^2 + \mathrm{D}^2)$: at $\mathrm{T} = 720$ and $\mathrm{D} = 862$ the two factors and their inverse square roots occupy about 10 MB, and the measured peak allocation of the whitening step at batch 32 is 0.63 GB. Inference is untouched. At that shape the whitening term costs roughly 3.1×10^{10} multiply-accumulates per step, by operation count a few percent of a Transformer forward-backward pass (not timed). Against DLinear (Zeng et al., 2023) it is not small: its two linear maps (seasonal and trend) cost about 3.8×10^9, so the loss term is about eight times as expensive as the model. Measured on DLinear on a 128-core CPU node at 12 threads, an ETTh1 epoch takes 1.23 s under DF against 1.43 s (+16%) with the temporal factor and 1.53 s (+24%) with both at $\mathrm{T} = 96$, and 1.67 s against 2.53 s (+51%) for either variant at $\mathrm{T} = 720$. On ECL at $\mathrm{T} = 192$ ($\mathrm{D} = 321$) the two-factor variant costs 25.0 s per epoch against 13.1 s (+91%). A reader porting WHITEDF to a lightweight forecaster should size the whitening against their own backbone.

E MEASURED RESIDUAL STRUCTURE (LINEAR CASE)

The theory of Section 4.1 is stated for linear predictors, and a linear forecaster can be fitted in closed form, so the claims can be checked exactly where they apply and the quantities the analysis invokes can simply be measured. We fit a channel-independent weight-sharing ridge forecaster (the DLinear parameterisation) at $96 \to 96$ under the standard chronological splits, estimate Σ_t and Σ_c from training residuals with $\gamma = 0.1$, and evaluate on the test split. No GPU and no training loop is involved. Figure 1 shows what these residuals look like on one dataset.

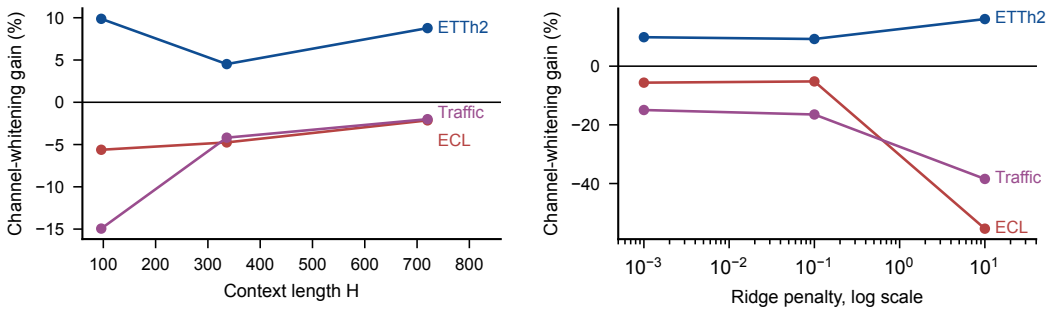
Table 2: Measured residual structure and the effect of channel whitening, closed-form linear forecaster. "Dyn. range" is $\lambda_{\max}/\lambda_{\min}$ of the spectrum of the best circulant approximation to the shrunk ($\gamma = 0.1$) Σ_t; "Circ. gap" is $\|\Sigma_t - C^\star(\Sigma_t)\|_F / \|\Sigma_t\|_F$ against the best circulant approximation; "Drift" is the largest relative difference between per-block Σ_t over four horizon blocks. $\mathrm{D}/(n\mathrm{T})$ uses the n windows the closed form was fitted on (8449 on ETTh, 6874 on ETTm, about 6100 elsewhere), not the full training split of Table 14. Bold: a change larger than 8% either way.

Dataset	D	Off-diag Σ_c	Dyn. range	Circ. gap	Drift	$\mathrm{D}/(n\mathrm{T})$	GLS vs OLS
ETTh1	7	0.600	153.9	0.316	0.420	8.6e-6	–0.3%
ETTh2	7	0.733	284.5	0.426	0.548	8.6e-6	**+10.2%**
ETTm1	7	0.575	223.9	0.353	0.474	1.1e-5	+1.1%
ETTm2	7	0.730	329.9	0.436	0.546	1.1e-5	**+8.3%**
Weather	21	0.590	119.6	0.377	0.477	3.6e-5	+2.8%
ECL	321	0.948	127.5	0.310	0.322	5.5e-4	–5.7%
Traffic	862	0.992	41.8	0.324	0.193	1.5e-3	**–15.3%**

Three of the paper's claims are confirmed and one of its predictions is not (Table 2, Figure 7).

Confirmed. The spectrum of the shrunk estimate spans one and a half to two and a half orders of magnitude on every dataset, and the raw residual spectrum two to three (Figure 8), so the flat-spectrum assumption isolated in Theorem 1 is false by a wide margin. Σ_t sits 31–44% away from its best circulant approximation, and the per-block drift of 19–55% confirms the non-stationarity that Remark 5 asserts. Whitening the targets by $\Sigma_t^{-1/2}$ and inverting the transform returns the OLS solution to a relative error of 10^{-13}, so temporal whitening cannot move this estimator, the cancellation of Remark 2.

Separability, measured. The nearest Kronecker product to the full $\mathrm{TD} \times \mathrm{TD}$ covariance of $\mathrm{vec}(E)$ in Frobenius norm, the rank-one term of its Van Loan–Pitsianis rearrangement, captures 0.927 of its squared Frobenius norm on ETTh1, 0.986 on ETTh2, 0.974 on ETTm1 and 0.996 on ETTm2 (relative error 0.27, 0.12, 0.16 and 0.06), 0.852 on Weather, and on a fixed random subset of 21 variables 0.839 on ECL and 0.702 on Traffic. Separability is a good approximation on ETT. The subsets compare datasets, not dimensions, and Traffic, where the deep model gains, is the least separable of the seven, so separability does not order the outcomes either.



(a) Context length

(b) Ridge penalty

Figure 7: Bias controls for the closed-form channel-whitening gain at a fixed window count of 5,000: (a) raising the context length H (more capacity, less bias) and (b) raising the ridge penalty (more bias). On ECL and Traffic all four curves move as the bias account predicts. On ETTh2 the ridge control runs the other way. Deterministic fits, no seed spread.

The Ledoit–Wolf intensity does not explain the failures. The analytic shrinkage intensity of Ledoit & Wolf (2004), computed on the same residual rows, is below 0.1 on every benchmark for both factors (for Σ_c: 0.004 on ETTh1, 0.09 on Weather, 0.02 on ECL, 0.01 on Traffic), which argues against iid sampling noise as the mechanism of the channel factor's failure on ECL and Traffic. The intensity cannot see the small eigenvalues whitening inverts and assumes independent rows where the residuals of overlapping windows are correlated, so it narrows the account of Section 5.3 without closing it. That a feasible GLS can lose to OLS in small samples is classical (Zellner, 1963; Kmenta & Gilbert, 1968; Binkley & Nelson, 1988).

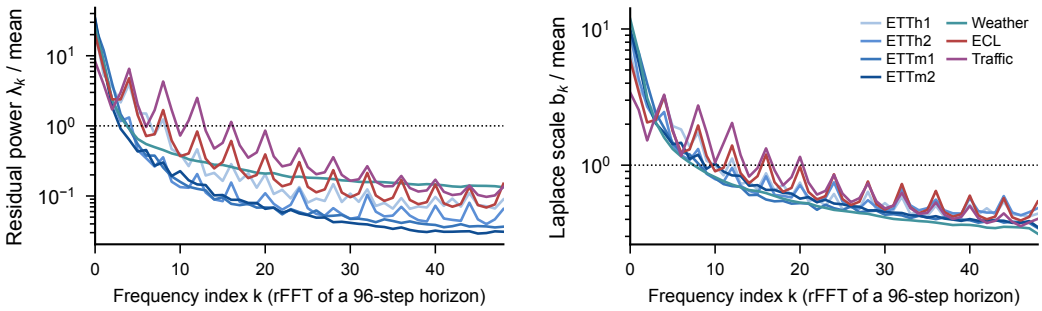
Not confirmed. The naive reading of Corollary 2, that the gain should track the off-diagonal mass of Σ_c, is contradicted: Traffic has the largest off-diagonal mass and the worst outcome. A sweep of $\gamma \in [0, 1]$ on ECL and Traffic is monotone with its optimum at $\gamma = 1$, the identity metric.

Two candidate mechanisms. Two mechanisms could produce the failures: the bias term B_θ of (9) dominating M_θ, so that whitening downweights reducible directions, or Σ_c being too poorly estimated at $\mathrm{D} = 862$ to be worth inverting. Monotonicity in γ does not discriminate between them. Two controls at a fixed window count (Figure 7) move the gain the way the bias account predicts on both failures. On ECL raising capacity improves it ($-5.6\% \to -2.1\%$ as the context grows from 96 to 720) and raising the ridge penalty degrades it ($-5.6\% \to -55.4\%$); on Traffic the two knobs move it from -14.9% to -2.0% and to -38.4%. On ETTh2, where whitening succeeds, the ridge control runs the other way ($+9.9\% \to +16.1\%$). Each knob moves the fit, the spectrum, the bias and the estimation error together, so the controls do not isolate the bias account, and with a ridge penalty the temporal cancellation of Remark 2 no longer holds, so no temporal claim rests on them. The linear forecaster is also badly misspecified on ECL and Traffic (test MSE 0.65 on Traffic against about 0.39 for a trained Transformer), and the trained iTransformer is below DF on Traffic at all four settings tried, beyond noise at one (Table 19), so the ordering by $\mathrm{D}/(n\mathrm{T})$ is a statement about the linear case.

F Additional Experiments

F.1 The ℓ_1 Norm Against the Basis, and Per-Frequency Weighting of the ℓ_1 Criterion

The norm against the basis. Corollary 1 settles the squared case. For the ℓ_1 case the missing control is the same norm without the transform, $\alpha \,\mathrm{mean}|E| + (1-\alpha)\,\mathrm{mean}\,E^2$ with FreDF's $\alpha = 0.5$ (for white residuals the two ℓ_1 terms enter at comparable scale, 0.80σ against 0.89σ). On the development cell it recovers about six sevenths of FreDF's gain (Section 3.2, Table 3), is behind FreDF on every seed, and is not separated from WhiteDF_t at the fixed setting (-0.00013, $t = -0.4$). One untested reading of the remainder is that an ℓ_1 term on the spectrum weights each frequency by the inverse magnitude of its own residual coefficient, sample by sample, a per-frequency weighting of



(a) Residual power λ_k

(b) Laplace scale b_k

Figure 8: Per-frequency residual power λ_k (a) and Laplace scale b_k (b) of the closed-form linear channel-independent forecaster at $\mathrm{T} = 96$ on all seven benchmarks, each divided by its mean over the 49 rFFT frequencies so that a flat spectrum is the dotted line at 1. λ_k spans about two to three orders of magnitude and b_k one to one and a half on every benchmark. The ridges every four bins on ECL and Traffic are the daily period. Their inverses are the weights the Gaussian and Laplace likelihoods prescribe, and neither is close to constant. Deterministic fits, no seed spread.

the kind Theorem 1 prescribes, which would also explain why an explicit b_k^{-1} on top of it does not help (below).

Wang et al. (2025a) motivate the ℓ_1 norm by the wide dynamic range of the spectrum. Experiment decides whether a robust norm or a weight handles that range better. Remark 4 says that FreDF's ℓ_1 frequency term is missing a per-frequency scale b_k^{-1} (Figure 8). Supplying it, as an EMA of the residual magnitude at each frequency, normalised and detached, on the development cell and its seeds: the weighted criterion still beats DF, by 0.0032 ($t = -12.6$), but it is worse than plain FreDF by 0.0011 ($t = +8.2$), losing a quarter of FreDF's gain. Shrinking the weight towards flat, as the squared-norm criterion needs, does not rescue it either. The deficit against FreDF falls monotonically with the shrinkage, $+0.0011$, $+0.0004$ and $+0.0001$ at $\gamma = 0$, 0.5 and 0.8 ($t > 7.8$ in each case, five seeds), and reaches zero only at $\gamma = 1$, which is FreDF itself. **Adding back what the likelihood omits does not help at any shrinkage we tried**, so the remark's prescription is used only with the squared norm.

F.2 PatchTST and DLinear at the Fixed Setting

Every online fixed-setting DLinear and PatchTST number in the paper comes from the rerun with the guarded loss module under the common checkpoint rule (Appendix I.1); the earlier run's tables are supplementary Tables S3 and S4.

Tables 21 and 22 have every cell. DLinear online is reported with the two-stage estimator in Appendix F.11.

PatchTST. Section 5.2 summarises the cells. At ETTh1 $\mathrm{T} = 96$ the order at $\mathrm{T} = 720$ reverses. The norm is the best criterion there, and $\mathrm{WhiteDF_t}$ is behind it by 0.0032 ($[0.0024, 0.0040]$). On ETTh2 $\mathrm{WhiteDF_t}$ is not separated from FreDF or from the norm at any horizon. Against DF, ETTm2 is not separated anywhere. Against the temporal factor, the variable factor helps on ETTh2 at $\mathrm{T} = 96$ (-0.0018, $t = -9.0$), the dataset where the closed form of Appendix E puts the largest channel gain, and is worse on ETTh1 at $\mathrm{T} \leq 336$ and on ETTm1 at three horizons, where it puts none or a small one. The hourly ETT sets have $8{,}545$ training windows against $34{,}465$ and $36{,}792$ for ETTm and Weather, the pattern the efficiency mechanism of Section 4.1 would produce, but iTransformer gains on ETTm1 as well, so sample size is not the whole account. On ETTh1 the criterion-independent validation MSE is worse for the whitened run in all four cells ($+0.004$ to $+0.011$ for the temporal factor) while test MSE improves by up to 0.028, whereas on ETTh2 the two splits agree at three of four horizons (Figure 10).

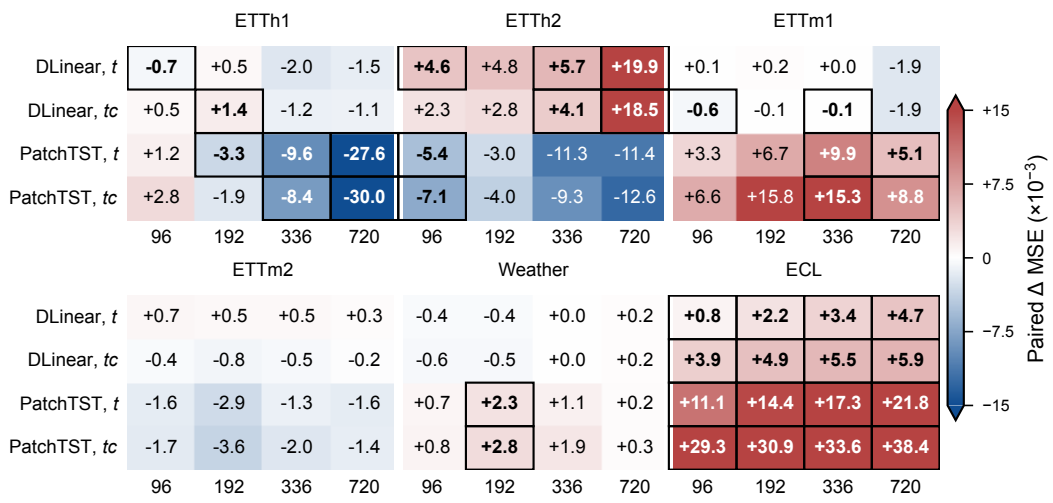


Figure 9: The fixed setting on two channel-independent backbones, every cell, guarded module and common checkpoint rule: mean paired difference in test MSE against DF ($\times 10^{-3}$, three seeds; negative is better). A black frame marks $p < 0.05$ (Table 16).

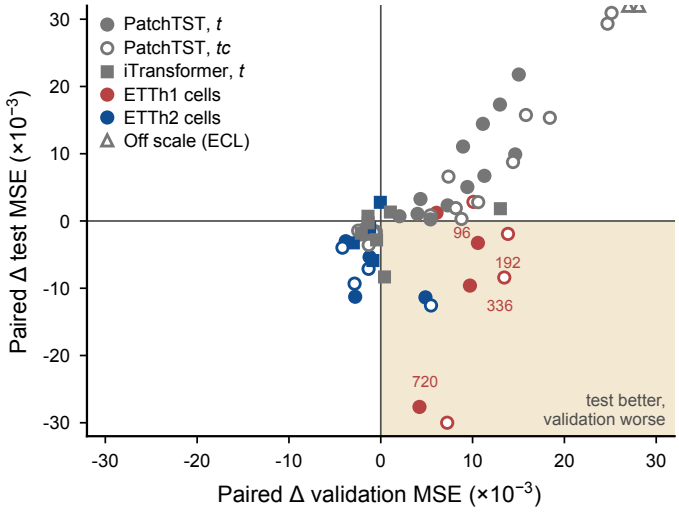


Figure 10: Validation against test, cell by cell. Each point is one dataset×horizon cell at the fixed setting: mean paired difference against DF in validation MSE (horizontal) and in test MSE (vertical), three seeds; PatchTST from the rerun (both factors, 24 cells each), iTransformer from the reference runs (11 cells). ETTh2 cells at $\mathrm{T} \leq 336$ sit where both splits agree that the criterion helps; ETTh1 cells at $\mathrm{T} \geq 192$ sit in the shaded quadrant where test improves and validation worsens, the disagreement described below. The off-scale ECL cells are worse on both splits.

F.3 Which Basis, and at What Strength?

Section 3.2 compares an ℓ_1 term in the frequency domain with the same norm in the time domain at equal α. This subsection asks whether equal α means equal strength, and whether the transform's contribution belongs to the Fourier basis in particular, on the development cell (iTransformer, ETTh1, $\mathrm{T} = 96$), five seeds per arm, reference protocol. Six bases are compared under the same criterion $\alpha \,\mathrm{mean}|B^\top e| + (1-\alpha)\,\mathrm{mean}\, e^2$: the identity (the time-domain control), FreDF's half-spectrum rFFT modulus, the orthonormal DCT-II, the real Fourier basis (cosine and sine columns, so the same subspace as the rFFT but with a per-coordinate ℓ_1 instead of a complex modulus), and two fixed random orthogonal matrices.

At equal α the gradient strength is already equal, by an identity. For orthonormal B, the auxiliary term's gradient with respect to the residual is $B\,\mathrm{sign}(B^\top e)/N$ over N entries, of norm $1/\sqrt{N}$ *whatever B is*. The training logs give a median of 0.006819 for the time domain, the DCT, the real Fourier basis and both random bases alike, against an analytic $1/\sqrt{32 \cdot 96 \cdot 7} = 0.0068193$. FreDF's half-spectrum modulus is not of this form, and its measured norm, 0.006886, is one percent higher. What differs at equal α is the term's *value* relative to the plain one (median 1.17 in the time domain, 0.98 for FreDF, 0.86 for the DCT and the real Fourier basis, 1.21 for the random bases), and that runs against the ordering below: the random bases are applied most strongly and do worst.

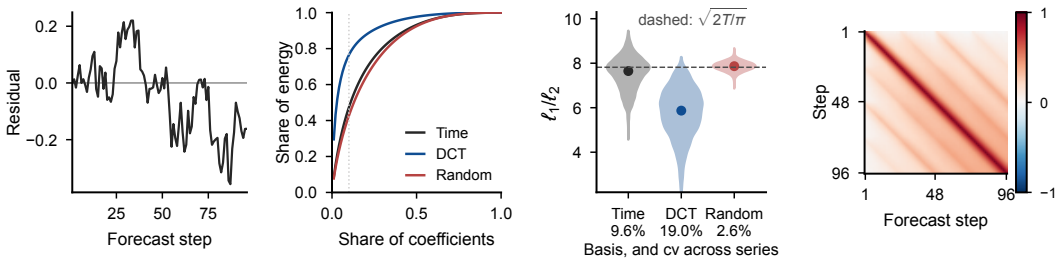
Calibrated arms. We also calibrate explicitly in two ways, using the training batch only. The first multiplies the auxiliary term at every step by $\mathrm{sg}(\|\partial\mathcal{L}^{(\mathrm{tmp})}/\partial e\|/\|\partial\mathcal{L}^{(\mathrm{aux})}/\partial e\|)$, so the two residual gradients have equal norm (the primary rule). The second multiplies it by $\mathrm{sg}(\mathcal{L}^{(\mathrm{tmp})}/\mathcal{L}^{(\mathrm{aux})})$, so the two values agree (the sensitivity check). The realised ratios in Table 3 confirm both act as intended. Neither changes the result. Against the time domain, FreDF is ahead by 0.00059 at equal α ($p = 0.047$), by 0.00074 under gradient calibration ($p = 0.041$) and by 0.00082 under value calibration ($p = 0.011$). Matching the strength makes the frequency term's advantage slightly larger. The advantage therefore does not come from the $\sqrt{\mathrm{T}}$ normalisation or from unequal effective mixing.

Table 3: Norm and basis on the development cell (iTransformer, ETTh1, $\mathrm{T} = 96$), five seeds per arm. Mean paired difference in test MSE ($\times 10^{-3}$) against DF and against the time-domain arm at the same calibration, with the paired t; bold: $p < 0.05$, uncorrected. "Realised ratio" is the median over training of the auxiliary term's value, and of its residual-gradient norm, relative to the plain term's.

Basis	– DF		– time domain		realised ratio	
	Δ	t	Δ	t	value	grad.
equal α (no calibration)						
time domain	**–3.64**	–48.4	–	–	1.17	0.85
rFFT modulus (FreDF)	**–4.22**	–26.9	**–0.59**	–2.8	0.98	0.85
DCT	**–4.07**	–21.2	–0.43	–1.8	0.86	0.85
real Fourier	**–3.98**	–20.6	–0.34	–1.4	0.86	0.85
random orthogonal 1	**–2.43**	–18.4	**+1.21**	+6.8	1.21	0.85
random orthogonal 2	**–2.63**	–28.9	**+1.00**	+8.8	1.21	0.85
gradient-norm calibrated						
time domain	**–3.72**	–49.0	–	–	1.38	1.00
rFFT modulus (FreDF)	**–4.46**	–20.6	**–0.74**	–3.0	1.16	1.00
DCT	**–4.38**	–18.9	–0.67	–2.6	1.01	1.00
real Fourier	**–4.25**	–17.0	–0.54	–1.9	1.01	1.00
random orthogonal 1	**–2.47**	–16.4	**+1.25**	+6.6	1.42	1.00
value calibrated						
time domain	**–3.38**	–35.8	–	–	1.00	0.72
rFFT modulus (FreDF)	**–4.20**	–23.7	**–0.82**	–4.5	1.00	0.86

Which basis. Every arm beats DF, and the transform's contribution is small and not Fourier-specific. Measured as a share of FreDF's own gain of 0.0042, the time domain alone recovers 86%, the DCT 96% and the real Fourier basis 94%. Against the time domain the DCT and the real Fourier basis are ahead by 0.00043 and 0.00034, neither separated from zero at equal α, and the DCT is not separated from FreDF either ($+0.00015$, $p = 0.054$). The exception is the random bases. They recover only 57% and 62%, and are behind the time domain by 0.00121 and 0.00100, beyond noise at equal α, and basis 1 also under gradient-norm calibration (0.00125). An arbitrary rotation is therefore worse than no transform at all, while the structured bases are interchangeable with each other to within the resolution of five seeds. **The ℓ_1 norm carries the effect, and the choice among structured bases makes no measurable difference.** Why a random basis falls behind has a simple account. A random rotation makes the coordinates of any residual nearly Gaussian, so $\|Qe\|_1$ concentrates at $\sqrt{2\mathrm{T}/\pi}\,\|e\|_2$ and the ℓ_1 term becomes a scaled Euclidean norm. On AR(1) paths with $\mathrm{T} = 96$ and ρ from 0 to 0.99 the ratio $\|Qe\|_1/\|e\|_2$ has a coefficient of variation of 2.3–2.7%, against 11–23% for the DCT at $\rho \geq 0.9$, where a correlated residual is sparse. On such Gaussian paths the time-domain

ratio is also nearly constant, so the time domain's advantage over a random basis has to come from residuals that are not Gaussian. Those of a trained model are not (Figure 11): on the test residuals of a DF iTransformer on ETTh1 at $\mathrm{T} = 96$ (2,785 windows, seven variables) the ratio varies by 9.6% across series in the time domain and 19.0% in the DCT against 2.6% in a random basis, and the excess kurtosis of the standardised coefficients is 0.62, 12.4 and -0.12. The ℓ_1 term therefore keeps a shape of its own in the time domain and the DCT and loses it in a random basis, which is consistent with the ordering of the arms without showing that the shape is what lowers test error.



(a) One test residual (b) Energy compaction (c) ℓ_1/ℓ_2 per series (d) Correlation of $\hat{S}_t$

Figure 11: **The same residuals in three bases, and the factor WHITEDF inverts.** A DF iTransformer on ETTh1 at $\mathrm{T} = 96$, seed 42, the setting of Section 2.3. (a) The residual of the first test window, OT variable. (b) Share of the residual energy held by the largest coefficients, averaged over all test series, in the time domain, the orthonormal DCT and a fixed random orthogonal basis. (c) ℓ_1/ℓ_2 ratio of every test series per basis, with its coefficient of variation. A random basis concentrates the ratio at $\sqrt{2\mathrm{T}/\pi}$, so an ℓ_1 term there acts as a Euclidean norm. (d) Correlation matrix of the training-residual second moment $\hat{S}_t$, whose condition number is 321.

F.4 THE WLS RUNG: DIAGONAL FACTORS ONLY

Between OLS and GLS sits weighted least squares, which reweights by marginal variance and ignores correlation. Whether WHITEDF's gain comes from that reweighting or from the correlation structure can be checked cheaply. We keep every ingredient of Section 4.2 (EMA, shrinkage, trace normalisation, stop-gradient) and zero the off-diagonals of the estimated factors before inverting them, so the temporal factor becomes a per-step variance and the variable factor a per-variable variance. On the development cell at $(0.5, 0.5)$, five seeds, paired against the same DF runs. The diagonal temporal factor improves on DF by 0.00036 ($t = -19.8$) and the diagonal pair by 0.00038 ($t = -23.6$), against 0.00351 for the full temporal factor. The diagonal rungs are worse than the full factor by 0.0031 ($t = +12.4$ and $+12.7$) on every seed. **Reweighting by marginal variance recovers a tenth of the gain.** The rest goes with the off-diagonal structure, which whitening uses and WLS does not. Zeroing the off-diagonals also changes the eigenvalues and the scale of the inverse, so the difference is not assigned to correlation alone; the matched-spectrum control of Appendix F.10, run on PatchTST, separates direction from spectrum.

F.5 SEARCHING FOR A STRONGER CRITERION ON THE DEVELOPMENT CELL

After Table 17 showed WHITEDF not exceeding FreDF, we asked whether a criterion built from the same estimate could. Candidates, each at the mixing weight $\alpha = 0.5$ unless stated: the ℓ_1 norm of the whitened residual, $\mathrm{mean}\,|\hat{\Sigma}_t^{-1/2}E|$, with the temporal factor at $\gamma = 0.5$ and 0.9 and with both factors; FreDF's frequency ℓ_1 term applied to the whitened residual; a three-way mixture of the squared, the ℓ_1 and the whitened squared terms; and the squared criterion at $\alpha = 0.7$ and 1. All were run on the development cell (ETTh1, $\mathrm{T} = 96$) with five seeds, and DF, FreDF and the WLS rungs of Appendix F.4 were rerun to record validation MSE. The selection rule was fixed before the runs: the candidate with the lowest mean validation MSE over five seeds, provided it beats DF's, ties within 10^{-4} going to the simpler criterion. The test columns play no part in it. Table 5 lists every arm.

By validation the winner is the whitened ℓ_1 criterion with the temporal factor at $\gamma = 0.9$ (0.6739 against DF's 0.6765). On test it improves on DF by 0.0038 and is worse than FreDF by 0.0005

Table 4: The two candidates of Table 5 and the norm-only control (time-domain ℓ_1, $\alpha = 0.5$) transferred unchanged to the nineteen held-out iTransformer cells, three seeds (42–44, paired on seed against the stored DF and FreDF runs of those seeds): mean paired difference in test MSE ($\times 10^{-3}$) with the paired t; bold $p < 0.05$. The last column pairs the whitened ℓ_1 criterion against the time-domain one on the same seeds.

Dataset	T	whitened ℓ_1 (0.5, 0.9) – DF	whitened ℓ_1 (0.5, 0.9) – FreDF	whitened FreDF ℓ_1 (0.5, 0.5) – DF	whitened FreDF ℓ_1 (0.5, 0.5) – FreDF	time-domain ℓ_1 (0.5) – DF	time-domain ℓ_1 (0.5) – FreDF	whitened ℓ_1 – time ℓ_1
ETTh1	192	**−3.9 (−10.9)**	+0.6 (+1.3)	**−4.4 (−4.7)**	+0.0 (+0.1)	**−3.4 (−13.2)**	+1.0 (+1.5)	−0.4 (−1.9)
ETTh1	336	**−4.4 (−5.6)**	+0.1 (+0.1)	−4.1 (−1.8)	+0.4 (+0.9)	**−4.3 (−9.3)**	+0.2 (+0.2)	−0.2 (−0.5)
ETTh1	720	−6.2 (−3.1)	+1.8 (+0.9)	−7.5 (−2.6)	+0.5 (+0.6)	−4.3 (−3.7)	+3.7 (+1.1)	−1.8 (−0.9)
ETTh2	96	+0.4 (+0.7)	+0.3 (+3.0)	**−4.4 (−12.3)**	**−4.5 (−8.6)**	−2.3 (−3.4)	−2.3 (−4.1)	**+2.6 (+5.3)**
ETTh2	192	**−2.6 (−4.8)**	+0.1 (+0.3)	−0.4 (−0.2)	+2.3 (+1.2)	**−2.7 (−8.6)**	+0.0 (+0.0)	+0.1 (+0.3)
ETTh2	336	−4.6 (−1.4)	+1.6 (+0.4)	−3.8 (−4.1)	+2.4 (+0.5)	**−2.3 (−4.4)**	+3.9 (+0.7)	−2.2 (−0.8)
ETTh2	720	+1.9 (+0.2)	+7.0 (+0.9)	**−6.1 (−21.8)**	−0.9 (−1.2)	+4.8 (+0.7)	+10.0 (+1.4)	−2.9 (−1.7)
ETTm1	96	**−11.5 (−20.2)**	**−5.8 (−18.5)**	**−5.7 (−5.3)**	+0.1 (+0.1)	**−10.5 (−6.7)**	−4.7 (−2.3)	−1.0 (−0.5)
ETTm1	192	**−8.7 (−76.1)**	**−4.5 (−5.5)**	**−3.2 (−5.6)**	+0.9 (+3.0)	**−8.4 (−18.8)**	**−4.3 (−4.4)**	−0.3 (−0.8)
ETTm1	336	**−10.9 (−11.3)**	−3.1 (−3.4)	**−6.5 (−5.0)**	+1.3 (+2.3)	**−10.7 (−6.6)**	−2.9 (−2.7)	−0.2 (−0.2)
ETTm1	720	−4.7 (−3.8)	+2.7 (+1.7)	**−7.4 (−9.7)**	+0.0 (+0.1)	**−8.3 (−5.6)**	−0.9 (−0.5)	+3.6 (+3.9)
ETTm2	96	**−3.0 (−4.5)**	**−2.6 (−4.7)**	+1.1 (+1.6)	+1.6 (+2.9)	**−3.9 (−13.8)**	**−3.4 (−24.6)**	+0.8 (+2.0)
ETTm2	192	−2.3 (−2.1)	−0.6 (−0.5)	−0.7 (−0.7)	+1.0 (+3.9)	**−5.0 (−5.0)**	**−3.2 (−21.6)**	+2.7 (+2.6)
ETTm2	336	−0.3 (−0.3)	−1.5 (−0.5)	+1.8 (+0.7)	+0.6 (+1.8)	**−3.1 (−5.8)**	−4.3 (−1.7)	**+2.8 (+5.0)**
ETTm2	720	**−1.9 (−4.8)**	−1.4 (−1.2)	+1.2 (+1.2)	+1.7 (+1.0)	−3.3 (−3.2)	**−2.8 (−6.4)**	+1.4 (+1.6)
Weather	96	**−1.5 (−14.6)**	−0.7 (−1.8)	−0.1 (−0.1)	+0.8 (+1.7)	−2.2 (−4.1)	**−1.4 (−4.5)**	+0.7 (+1.6)
Weather	192	**−0.9 (−6.3)**	−1.5 (−2.8)	+1.0 (+1.1)	+0.3 (+1.3)	**−1.8 (−16.8)**	−2.5 (−3.7)	**+1.0 (+6.9)**
Weather	336	−0.5 (−1.6)	−1.7 (−4.2)	**+1.6 (+12.1)**	+0.4 (+1.0)	**−1.3 (−6.1)**	**−2.4 (−6.2)**	**+0.8 (+8.0)**
Weather	720	−1.2 (−4.0)	**−2.5 (−5.6)**	+1.6 (+3.9)	+0.3 (+1.2)	−1.3 (−4.2)	−2.6 (−3.5)	+0.1 (+0.4)

($t = 3.7$). The arm with the lowest test error, FreDF's ℓ_1 term on the whitened residual, beats FreDF by 0.0003 ($t = -3.1$) and WHITEDF$_t$ by 0.0010 ($t = -7.5$) and has the worst validation error of the $\alpha = 0.5$ candidates (0.6791), the disagreement between ETTh1's two splits once more (Appendix F.11). The differences are small. The whitened ℓ_1, FreDF, the time-domain ℓ_1 and WHITEDF$_t$ lie within 0.0007 of one another on test, and raising α for the squared criterion only worsens validation.

Transfer to the held-out cells. Both picks and the norm-only control (the time-domain ℓ_1 criterion of Appendix F.1, second by validation on the development cell) were run unchanged on the nineteen held-out cells with seeds 42–44 (Table 4). The validation pick beats DF beyond noise on ten cells and is worse on none; against FreDF it is ahead on four, behind on none and not separated over the nineteen (Wilcoxon $p = 0.28$). The test pick, FreDF's ℓ_1 term on the whitened residual, is above FreDF in mean on seventeen of nineteen (Wilcoxon $p = 0.007$): the arm that looked best on the development cell's test error does not transfer, and the pre-registered validation rule chose one that does. The validation pick is behind the plain norm on four cells, ahead on none and not separated over the nineteen (Wilcoxon $p = 0.42$), and it sits at $\gamma = 0.9$, nine tenths of the way to the identity metric, so **the search converged on the norm**. The four cells on which it is ahead of FreDF are all cells on which the plain norm is ahead of FreDF with five seeds.

The norm-only control on five seeds. The main text leans on this control, so seeds 45 and 46 were added for it on all nineteen cells, paired against the stored DF, FreDF and WHITEDF$_t$ runs of those seeds (the two candidates stay on three seeds). The resulting counts are in Section 3.2, Figure 3b and Appendix I.2; against WHITEDF$_t$ the norm is ahead on fourteen cells and behind on none. The three-seed verdicts of Table 4 were not reversed on any cell: the added seeds moved cells from "not separated" to "ahead", and the cells where the norm is not ahead of FreDF are all on ETTh1 and ETTh2 at T $\geq$ 192 and ETTm1 at T = 720.

F.6 Comparison with QDF in Its Authors' Repository

QDF (Wang et al., 2026a) learns the same T $\times$ T object as our temporal factor by bilevel optimisation (Appendix A.5), and its released code trains a different backbone under a different protocol from ours, so the comparison was run inside that repository rather than in ours: TQNet backbone, the authors' ETTh1 hyperparameters per horizon (learning rates, meta learning rate, inner steps, task count; 30 epochs, patience 5, batch 32, RevIN, dropout 0.5), three seeds (42–44) for every arm, the

Table 5: Development cell (ETTh1, T = 96), five seeds per arm, ranked by mean validation MSE, the pre-registered selection score. Test MSE and paired test differences against DF and FreDF (t in parentheses; bold $p < 0.05$). *: fewer than five seeds carry a validation record (three or four); the test difference is paired over the seeds available (five for the fixed setting, from the stored runs).

Criterion	val. MSE	test MSE	− DF	− FreDF
whitened ℓ_1, Σ_t (0.5, 0.9)	0.6739	0.3836	**−0.0038 (−30.45)**	**+0.0005 (+3.68)**
time-domain ℓ_1 ($\alpha = 0.5$)	0.6747	0.3837	**−0.0036 (−48.39)**	**+0.0006 (+2.85)**
whitened ℓ_1, both (0.5, 0.5)	0.6754	0.3836	**−0.0038 (−36.93)**	**+0.0004 (+2.96)**
diag. Σ_t, Σ_c (WLS) (0.5, 0.5)	0.6757	0.3870	**−0.0004 (−23.59)**	**+0.0038 (+24.67)**
whitened ℓ_1, Σ_t (0.5, 0.5)	0.6758	0.3832	**−0.0042 (−36.00)**	+0.0000 (+0.13)
diag. Σ_t (WLS) (0.5, 0.5)	0.6762	0.3870	**−0.0004 (−19.80)**	**+0.0039 (+26.13)**
DF	0.6765	0.3874	–	**+0.0042 (+26.90)**
FreDF (ℓ_1, $\alpha = 0.5$)	0.6780	0.3832	**−0.0042 (−26.90)**	–
three-way mix, both (0.5, 0.5)	0.6780	0.3837	**−0.0036 (−20.86)**	**+0.0006 (+4.47)**
three-way mix, Σ_t (0.5, 0.5)	0.6784	0.3834	**−0.0040 (−21.21)**	+0.0003 (+1.95)
whitened FreDF ℓ_1, Σ_t (0.5, 0.5)	0.6791	0.3829	**−0.0045 (−21.42)**	**−0.0003 (−3.08)**
WHITEDF$_t$ (0.5, 0.5)*	0.6792	0.3839	**−0.0035 (−13.95)**	**+0.0007 (+3.42)**
WHITEDF$_t$ (0.7, 0.5)	0.6875	0.3833	**−0.0041 (−13.08)**	+0.0002 (+0.47)
WHITEDF$_t$ (1, 0.5)*	0.7063	0.3842	**−0.0034 (−12.69)**	+0.0008 (+1.91)

checkpoint chosen on validation MSE for every arm by that code's own rule. Arms: the repository's DF; FreDF as the repository implements it (rfft ℓ_1 with weight 0.5, MSE 0.5); WHITEDF$_t$ and WHITEDF$_{t+c}$ as an auxiliary term with the same $0.5/0.5$ mixing and $\gamma = 0.5$, untuned, and QDF itself, with no algorithm touched. Table 6 gives the paired differences. Over three seeds no arm separates from DF at $p < 0.05$ except FreDF at T = 96: QDF's largest gain, -0.0081 at T = 720 ($t = -2.6$), does not reach it, and its seed spread at T = 96 is 0.371–0.378 against DF's 0.376–0.377; WHITEDF$_t$ is below DF at T = 96, 192 and 336 (-0.0023, -0.0028, -0.0016) without reaching it either. Head to head, WHITEDF$_t$ is not separated from QDF at T = 96, 336 and 720 and better at 192 (-0.0025, $t = -5.7$). FreDF is the strongest arm on this backbone, better than DF by 0.0089 at T = 96 ($t = -5.5$) and lowest in mean of the five arms above at three of four horizons (-0.0174 at T = 336, $t = -3.5$), as in Section 3.2. **On this data the ℓ_1 norm is worth more than either estimate of Σ_t.**

ECL in the same repository. With the authors' ECL settings (cycle 168, no dropout, batch 16, RevIN, their learning rates), three seeds and the checkpoint chosen on validation MSE for every arm (Table 6, first row), every arm is below DF beyond noise: FreDF by 0.0007 ($t = -14.5$), WHITEDF$_t$ by 0.0008 ($t = -18.7$), WHITEDF$_{t+c}$ by 0.0010 ($t = -44.4$) and QDF by 0.0004 ($t = -10.3$), on a DF of 0.1353 whose three seeds span 0.00014. Head to head, WHITEDF$_{t+c}$ is better than QDF (-0.0006, $t = -12.1$) and not separated from FreDF (-0.0003, $t = -4.1$, $p = 0.06$), and the variable factor's increment over the temporal one (-0.0002, $p = 0.07$) is not separated from zero. The temporal factor, which never crosses zero on ECL for the three backbones of Section 5, does so on this one, by half a percent of DF's error. On iTransformer the common checkpoint rule does not reverse the ECL failure (Table 10), so the difference is not the checkpoint rule; backbone, instance normalisation and repository protocol all differ between the two settings. The reproduced margins are also smaller than the published ones: QDF's Table 6 reports 0.365 against DF's 0.372 on ETTh1 at T = 96 where we reproduce 0.3735 against 0.3765, and its published ECL number at T = 96 (0.135 against DF's 0.143) equals the DF we reproduce (0.1353). What the run establishes is that **the temporal factor's failure on ECL belongs to a backbone and protocol, not to the dataset**, as Traffic already showed for the deep model (Section 5.3).

The norm-only control inside the same repository. The FreDF arm above is the authors' own implementation, whose unnormalised transform makes its frequency term $\sqrt{\mathrm{T}}$ heavier than ours at the same weight (Appendix I). Two time-domain arms were added on ETTh1, three seeds each: the ℓ_1 residual norm at the FreDF arm's nominal weights (0.5 and 0.5) and the same norm weighted 0.9 against 0.1 MSE (Table 6, columns "time ℓ_1"). Neither separates from DF at any horizon. Against the authors' FreDF the nominally weighted norm is behind at T = 96 ($+0.0072$, $t = 12.6$) and not separated at the other three horizons, and the reweighted norm is not separated at any. On this backbone, with the released implementation, the basis keeps part of frequency alignment's gain at

Table 6: Inside the QDF repository: TQNet on ETTh1 (four horizons) and ECL (T = 96) with the authors' hyperparameters, three seeds, every arm's checkpoint chosen on validation MSE. Entries: mean paired difference in test MSE against the repository's DF ($\times 10^{-3}$), with the paired t; bold $p < 0.05$. QDF runs with its tuned meta hyperparameters, WHITEDF with the untuned $(0.5, 0.5)$; "time ℓ_1" is the time-domain norm at the FreDF arm's weights and "(0.9)" the same norm weighted 0.9 (ETTh1 only).

Dataset	T	DF	FreDF – DF	time ℓ_1 – DF	time ℓ_1 (0.9) – DF	WHITEDF$_t$ – DF	WHITEDF$_{t+c}$ – DF	QDF – DF
ECL	96	0.135	**−0.7 (−14.5)**	–	–	**−0.8 (−18.7)**	**−1.0 (−44.4)**	**−0.4 (−10.3)**
ETTh1	96	0.377	**−8.9 (−5.5)**	−1.8 (−1.5)	−4.7 (−3.1)	−2.3 (−3.5)	−2.8 (−2.9)	−3.1 (−1.3)
ETTh1	192	0.438	−3.7 (−0.7)	−5.6 (−2.3)	−6.3 (−2.7)	−2.8 (−3.3)	−1.5 (−0.9)	−0.3 (−0.3)
ETTh1	336	0.490	−17.4 (−3.5)	−8.9 (−1.7)	−11.4 (−1.8)	−1.6 (−0.2)	−1.7 (−0.2)	+3.4 (+0.5)
ETTh1	720	0.497	−5.5 (−1.2)	−5.5 (−2.1)	−3.4 (−0.3)	+0.5 (+0.1)	−1.8 (−0.4)	−8.1 (−2.6)

the shortest horizon and none that separates from noise at the other three, so the norm-only reading of Section 3.2 holds only in part here.

F.7 FREDF AT ITS RELEASED OPERATING POINT

The FreDF arm of Section 3.2 mixes an orthonormal-rFFT ℓ_1 term with MSE in equal parts. The released iTransformer scripts of Wang et al. (2025a) use the unnormalised transform, whose coefficients are $\sqrt{\mathrm{T}}$ larger, with per-dataset weights (Appendix I; the α column of Table 7). This subsection runs FreDF at that operating point on the twenty iTransformer cells.

Arms. DF; FreDF as released, $(1-\alpha)\,\mathrm{mean}\,E^2 + \alpha\,\mathrm{mean}|\mathrm{rfft}(E)|$ with the unnormalised transform and α the released weight, which reproduces the loss of the released training code to rounding; and the time-domain ℓ_1 at the same weight and matched strength, $(1-\alpha)\,\mathrm{mean}\,E^2 + \alpha\sqrt{\mathrm{T}}\,\mathrm{mean}|E|$. The factor $\sqrt{\mathrm{T}}$ puts the two auxiliary terms at the same residual-gradient norm, to within 1% at T = 96 and 720 on random inputs (Appendix F.3 gives the identity for orthonormal bases). Architecture, optimiser and checkpoint rule are the reference setting of Section 2.3, not the hyperparameters of FreDF's own scripts (which use, for instance, $d_{\mathrm{model}} = 128$ and a learning rate of 10^{-3} on ETTh1), so that DF and every arm share one configuration. Seeds 42–46, twenty cells, 300 runs at the released weights and 400 more for the weight grid described below.

Hardware. DF and every released-weight run were trained on one rented RTX PRO 6000, so those pairs share hardware; 218 runs of the weight grid ran on a rented RTX 3090 with the same software and loss module (M4, Appendix I.1) and are paired with the first card's DF, which the second card reproduces to six decimals and which matches the A6000 records of the same seeds to four decimals.

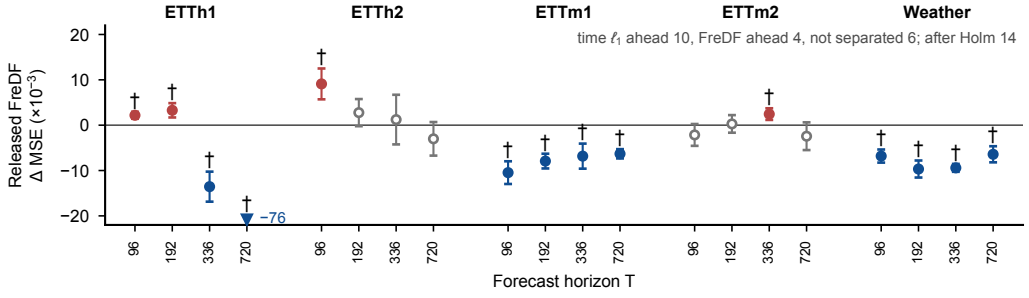


Figure 12: **The time-domain norm stays ahead of FreDF on most cells when FreDF runs at its own operating point.** Time-domain ℓ_1 at matched strength minus FreDF at its released weights on iTransformer, test MSE (Table 7; the same contrast against our equal-mix FreDF is Figure 3b). Five seeds, 95% intervals. Blue: norm ahead at $p < 0.05$; red: FreDF ahead; hollow: not separated; dagger: survives Holm over the twenty cells. The ETTh1 T = 720 cell lies off scale at −76.

Result. Against DF, FreDF at its released weights is better on ten cells and worse on five, and the time-domain ℓ_1 at matched strength is better on nineteen and worse on none. Head to head, the norm is ahead on ten cells and behind on four, not separated on six, and every one of these fourteen verdicts

Table 7: FreDF at its released weight α on iTransformer, five seeds per cell, DF retrained on the same GPU. Mean paired difference in test MSE ($\times 10^{-3}$) with the paired t. Bold marks $p < 0.05$, and † marks a cell that also survives Holm within its column over the twenty cells. The last column is the time-domain ℓ_1 at matched strength minus FreDF.

Dataset	α	T	DF MSE	FreDF (released) − DF	time ℓ_1 ($\sqrt{\text{T}}$) − DF	time ℓ_1 − FreDF
ETTh1	0.8	96	0.387	**−6.4 (−19.3)**†	**−4.2 (−31.0)**†	**+2.2 (+7.9)**†
		192	0.442	**−7.2 (−17.7)**†	**−4.0 (−4.8)**†	**+3.3 (+5.8)**†
		336	0.484	**+6.4 (+3.8)**	**−7.2 (−7.8)**†	**−13.6 (−11.4)**†
		720	0.508	**+58.1 (+22.2)**†	**−17.9 (−7.8)**†	**−75.9 (−24.8)**†
ETTh2	0.9	96	0.301	**−14.9 (−18.0)**†	**−5.8 (−4.8)**†	**+9.1 (+7.4)**†
		192	0.381	**−11.0 (−9.6)**†	**−8.2 (−5.6)**†	+2.8 (+2.6)
		336	0.426	−1.4 (−0.9)	−0.2 (−0.1)	+1.2 (+0.6)
		720	0.429	−0.8 (−0.7)	**−3.8 (−3.8)**†	−3.0 (−2.3)
ETTm1	0.8	96	0.339	**−11.3 (−5.9)**†	**−21.8 (−16.0)**†	**−10.5 (−11.6)**†
		192	0.381	−2.4 (−2.0)	**−10.3 (−9.6)**†	**−7.9 (−13.5)**†
		336	0.422	**−7.7 (−4.5)**	**−14.5 (−14.3)**†	**−6.8 (−6.9)**†
		720	0.492	**−7.5 (−6.2)**†	**−13.8 (−9.1)**†	**−6.3 (−17.1)**†
ETTm2	0.9	96	0.183	**−5.6 (−4.7)**	**−7.8 (−20.9)**†	−2.2 (−2.5)
		192	0.252	**−8.9 (−15.6)**†	**−8.6 (−7.3)**†	+0.3 (+0.4)
		336	0.314	**−10.2 (−7.0)**†	**−7.8 (−7.1)**†	**+2.4 (+5.3)**†
		720	0.411	−2.5 (−1.8)	**−5.0 (−8.3)**†	−2.4 (−2.2)
Weather	0.6	96	0.180	+1.1 (+1.9)	**−5.7 (−12.3)**†	**−6.8 (−13.1)**†
		192	0.228	**+5.0 (+10.0)**†	**−4.6 (−8.6)**†	**−9.7 (−14.3)**†
		336	0.284	**+4.6 (+11.5)**†	**−4.8 (−8.7)**†	**−9.4 (−30.0)**†
		720	0.360	**+3.5 (+5.9)**†	**−2.9 (−8.9)**†	**−6.4 (−10.0)**†

survives Holm. FreDF's edge is at ETTh1 $\text{T} = 96$ ($+0.0022$) and 192 ($+0.0033$), ETTh2 $\text{T} = 96$ ($+0.0091$) and ETTm2 $\text{T} = 336$ ($+0.0024$). At the shortest hourly horizons the transform therefore adds something the norm does not, as in QDF's repository (Appendix F.6). The norm's advantage is on all four ETTm1 horizons, all four Weather horizons and ETTh1 at $\text{T} = 336$ and 720. Part of it comes from cells where the released weights make FreDF worse than DF: ETTh1 at $\text{T} = 720$ by 0.058, with every seed and the validation split agreeing, ETTh1 at $\text{T} = 336$ and Weather at $\text{T} = 192$–720. Choosing FreDF's weight on validation does not remove this harm (below). Over the twenty cells the mean difference is -0.0070 for the norm against FreDF. **The reading of Section 3.2 therefore holds at FreDF's released operating point on most cells**, and its exception, the shortest hourly horizons, is the one the QDF-repository run had already shown. Tuning both weights on validation keeps that picture and widens FreDF's edge on the hourly datasets.

Weights chosen on validation. The released weights were tuned together with FreDF's own hyperparameters, so we also let each arm choose its weight. For every cell, FreDF and the time-domain norm were run at the released weight and at two others (0.5 and 0.9 on ETTh1, ETTm1 and Weather, 0.5 and 0.8 on ETTh2 and ETTm2), five seeds each, and each arm keeps the weight with the lowest mean validation MSE (Table 8). FreDF then picks 0.5 on the hourly datasets, on Weather and on ETTm1 beyond $\text{T} = 96$, and 0.8 or 0.9 elsewhere. Against DF, the tuned FreDF is better on thirteen cells and worse on four, and the tuned norm better on nineteen and worse on none. Head to head, the norm is ahead on ten cells and behind on six, not separated on four, and fifteen of the sixteen verdicts survive Holm. FreDF's edge now covers ETTh1 at $\text{T} = 96$ and 192, ETTh2 at $\text{T} = 96$, 192 and 336 and ETTm2 at $\text{T} = 336$, and the norm's advantage stays on ETTh1 at $\text{T} = 336$ and 720 and on every ETTm1 and Weather horizon. The harm of FreDF does not disappear with tuning: at its best weight in the grid it is still worse than DF by 0.055 on ETTh1 at $\text{T} = 720$ and by 0.003 to 0.005 on Weather at $\text{T} = 192$–720. The harm is therefore not a matter of the released weight transferring badly, and the scale of the unnormalised transform is a candidate cause. The grid has three values per arm, and weights and verdicts use the validation and test splits of the same five seeds.

Table 8: FreDF and the time-domain ℓ_1 at matched strength on iTransformer, each with its weight chosen on validation from three values. Picked α per arm, and mean paired differences in test MSE ($\times 10^{-3}$) with the paired t, five seeds. Bold marks $p < 0.05$, and † marks a cell that also survives Holm within its column over the twenty cells. *: weight chosen on validation.

		picked α		FreDF*	time ℓ_1^*	time ℓ_1^*
Dataset	T	FreDF	time ℓ_1	− DF	− DF	− FreDF*
ETTh1	96	0.5	0.5	**−6.6 (−20.0)**†	**−4.5 (−33.6)**†	**+2.0 (+6.6)**†
	192	0.5	0.5	**−7.7 (−16.1)**†	**−4.1 (−5.0)**†	**+3.6 (+6.8)**†
	336	0.5	0.5	+0.9 (+0.6)	**−6.9 (−8.8)**†	**−7.8 (−5.6)**†
	720	0.5	0.9	**+54.7 (+24.8)**†	**−18.0 (−7.8)**†	**−72.8 (−24.5)**†
ETTh2	96	0.5	0.8	**−11.5 (−11.0)**†	**−5.6 (−4.6)**†	**+5.9 (+3.2)**
	192	0.5	0.5	**−12.5 (−11.0)**†	**−7.6 (−4.9)**†	**+5.0 (+5.1)**†
	336	0.5	0.5	**−10.5 (−6.7)**†	−2.9 (−1.8)	**+7.5 (+6.6)**†
	720	0.5	0.5	**−3.8 (−5.7)**†	**−4.0 (−2.8)**	−0.3 (−0.2)
ETTm1	96	0.9	0.9	**−10.6 (−5.9)**†	**−22.2 (−19.6)**†	**−11.7 (−12.6)**†
	192	0.5	0.5	**−5.0 (−3.4)**	**−11.3 (−10.9)**†	**−6.3 (−9.9)**†
	336	0.5	0.5	**−9.1 (−6.2)**†	**−15.0 (−14.4)**†	**−5.9 (−6.9)**†
	720	0.5	0.5	**−9.3 (−5.9)**†	**−13.9 (−8.5)**†	**−4.7 (−9.3)**†
ETTm2	96	0.9	0.9	**−5.6 (−4.7)**	**−7.8 (−20.9)**†	−2.2 (−2.5)
	192	0.9	0.9	**−8.9 (−15.6)**†	**−8.6 (−7.3)**†	+0.3 (+0.4)
	336	0.9	0.9	**−10.2 (−7.0)**†	**−7.8 (−7.1)**†	**+2.4 (+5.3)**†
	720	0.8	0.9	−3.6 (−2.3)	**−5.0 (−8.3)**†	−1.3 (−1.1)
Weather	96	0.5	0.9	+0.4 (+1.0)	**−7.1 (−14.1)**†	**−7.5 (−11.1)**†
	192	0.5	0.9	**+4.9 (+12.8)**†	**−4.9 (−9.2)**†	**−9.9 (−17.6)**†
	336	0.5	0.9	**+4.0 (+6.5)**†	**−5.0 (−7.1)**†	**−9.0 (−16.8)**†
	720	0.5	0.5	**+3.3 (+4.8)**	**−3.0 (−10.7)**†	**−6.4 (−10.3)**†

F.8 Results Under a Common Checkpoint Rule

The reference training loop stops on the validation value of the criterion it trains with, so a DF run and a WhiteDF run are selected by different rules, and part of any paired difference can be the stopping rule rather than the criterion. The runs below choose the checkpoint on validation MSE for every criterion.

PatchTST on hourly ETT. With the checkpoint chosen on validation MSE and the loss module that updates its statistics on training batches only, the eight hourly-ETT cells of PatchTST give the picture of the earlier reference-rule runs: no cell is worse under either rule. The temporal factor lowers test MSE on every cell under the reference rule and on seven of eight under the common one (ETTh1 at $\mathrm{T} = 96$: $+0.0012$, $t = 1.1$), beyond noise on three cells and on four, worse on none in either. The two-factor variant is better beyond noise on two and three, worse on none. The ETTh2 $\mathrm{T} = 96$ gain halves, from 0.0099 to 0.0054, and crosses the threshold only under the common rule ($t = -5.0$). The ETTh1 $\mathrm{T} = 720$ gain is 0.028 ($t = -8.9$) under the common rule against 0.026 ($t = -3.9$) under the reference one. The reference cells were trained on a CPU node and the reruns on a GPU, and DF itself differs between the two by up to 0.009, the size of the paired differences, so only within-protocol paired differences are compared (side by side in supplementary Table S2).

ECL under the common rule. iTransformer on ECL at $\mathrm{T} = 96$ was rerun with the checkpoint chosen on validation MSE for every criterion, three seeds, $\gamma \in \{0.5, 0.9\}$ for both factors (Table 10); DF's three runs reproduce the reference ones to the last digit. The failure survives. At $\gamma = 0.5$ the temporal factor is worse than DF by $+0.0008$ ($t = 4.4$) and both factors by $+0.0014$ ($t = 9.2$), within 0.0001 of the numbers under the reference rule. At $\gamma = 0.9$ both lie above zero without clearing the threshold ($+0.0001$ and $+0.0004$). **On this backbone the ECL boundary is not a stopping-rule artefact**, so the gain inside QDF's repository (Appendix F.6) cannot be attributed to the checkpoint rule alone.

Table 9: PatchTST on hourly ETT under the common checkpoint rule, with the guarded loss module and three seeds, paired on seed: all six contrasts. Entries are mean paired differences in test MSE ($\times 10^{-3}$) with the paired t. Bold marks $p < 0.05$, uncorrected, and † marks a cell that also survives Holm within its column. The first four contrast columns are those of Table 1.

		DF	Minus DF				WHITEDF$_t$ minus	
Dataset	T	MSE	FreDF	Time ℓ_1	WHITEDF$_t$	WHITEDF$_{t+c}$	Time ℓ_1	FreDF
ETTh1	96	0.379	+0.5 (+0.4)	−2.0 (−1.9)	+1.2 (+1.1)	+2.8 (+2.3)	**+3.2 (+17.6)**†	**+0.7 (+4.3)**
	192	0.431	**−3.2 (−10.7)**	**−3.6 (−17.2)**†	**−3.3 (−7.7)**	−1.9 (−4.1)	+0.4 (+1.2)	−0.0 (−0.1)
	336	0.474	**−8.8 (−5.8)**	**−8.8 (−20.6)**†	**−9.6 (−13.2)**†	**−8.4 (−13.5)**†	−0.8 (−2.4)	−0.8 (−1.0)
	720	0.521	**−18.4 (−8.4)**	−2.4 (−0.3)	**−27.6 (−8.9)**	**−30.0 (−11.5)**	**−25.3 (−6.0)**	**−9.3 (−10.0)**
ETTh2	96	0.300	**−4.6 (−4.7)**	**−5.4 (−4.4)**	**−5.4 (−5.0)**	**−7.1 (−6.0)**	+0.0 (+0.0)	−0.8 (−0.6)
	192	0.377	−1.9 (−1.6)	−3.4 (−1.4)	−3.0 (−0.7)	−4.0 (−1.0)	+0.4 (+0.2)	−1.1 (−0.3)
	336	0.433	−9.0 (−3.0)	−3.9 (−3.9)	−11.3 (−2.0)	−9.3 (−3.3)	−7.4 (−1.1)	−2.3 (−0.8)
	720	0.442	−6.6 (−1.7)	−3.2 (−0.7)	−11.4 (−3.5)	−12.6 (−4.1)	−8.1 (−3.5)	−4.8 (−3.5)

Table 10: iTransformer on ECL, T = 96, three seeds: mean paired difference in test MSE against DF with the paired t (bold $p < 0.05$) under the reference protocol (each criterion stops on its own validation loss) and under a common rule (every criterion stops on validation MSE). DF is identical under the two rules. Cells not run are marked –.

Criterion	γ	reference rule	common rule
WHITEDF$_t$	0.1	**+0.0018 (+11.93)**	–
WHITEDF$_t$	0.5	**+0.0008 (+4.58)**	**+0.0008 (+4.40)**
WHITEDF$_t$	0.9	–	+0.0001 (+1.20)
WHITEDF$_{t+c}$	0.1	**+0.0027 (+13.68)**	–
WHITEDF$_{t+c}$	0.5	**+0.0015 (+10.95)**	**+0.0014 (+9.18)**
WHITEDF$_{t+c}$	0.9	+0.0004 (+3.16)	+0.0004 (+3.16)

The ℓ_1 family under the common rule. Every ℓ_1-family arm of Appendices F.1 and F.5 was trained under the reference protocol, that is, stopped on its own validation loss. Rerunning FreDF and the time-domain ℓ_1 criterion on the nineteen held-out iTransformer cells with the checkpoint chosen on validation MSE (three seeds) leaves the verdicts in place. Cell by cell the common-rule runs are not separated from the reference-rule runs for either criterion (nineteen of nineteen, mean shifts $+0.0001$ and $+0.0002$), the time-domain norm beats DF on eleven cells and is worse on none, FreDF on six and none, and the norm is ahead of FreDF on eight cells, behind on none, and below it in mean on thirteen of nineteen (pooled Wilcoxon $p = 0.06$, descriptive). **The stopping rule is not what separates the two criteria.**

F.9 THE GLOBAL SCALE OF THE WHITENED TERM

Trace normalisation fixes $\operatorname{tr}\Sigma_t^{(\gamma)}$ and $\operatorname{tr}\Sigma_c^{(\gamma)}$, not the value of the whitened term, so at $\alpha = 0.5$ the two terms of Eq. (5) need not carry equal weight, and a gain could in principle come from the global scale of the whitened term rather than from its direction weights. We measured both. A copy of the loss module logs, every 50 training steps, the plain term $\mathcal{L}^{(\mathrm{tmp})}$, the whitened term $\mathcal{L}^{(\mathrm{wht})}$, the norms of the two gradients with respect to the residual, and the mean precisions $\operatorname{tr}\Sigma_t^{-1}/\mathrm{T}$ and $\operatorname{tr}\Sigma_c^{-1}/\mathrm{D}$, and a scale-matched arm multiplies the whitened term at every step by $\mathrm{sg}(\mathcal{L}^{(\mathrm{tmp})}/\mathcal{L}^{(\mathrm{wht})})$, so that its value, and the global scale of its gradient, equal the plain term's while its direction weights are unchanged. Three cells, three seeds each, guarded module, common checkpoint rule.

Measured scale. Over training the whitened term is 0.40–0.43 of the plain MSE on PatchTST ETTh1 at T = 720, 0.56–0.58 at T = 96, and 0.52–0.54 on iTransformer ETTh1 at T = 96 (temporal factor; 0.33–0.48 with both factors), and the whitened gradient is 0.64–0.79 of the plain one. The mean precision $\operatorname{tr}\Sigma_t^{-1}/\mathrm{T}$ sits at 1.45–1.60. So at $\alpha = 0.5$ the whitened direction receives well under half the weight, by an amount that varies with dataset, horizon and factor: α is not the effective mixing weight.

Scale-matched arm. Matching the scale enlarges each effect. On PatchTST ETTh1 at T = 720 the temporal factor lowers test MSE by 0.0276 against DF and the scale-matched arm by 0.0454

([0.0343, 0.0565], 8.7%), ahead of the unmatched arm by 0.0178 ($t = -29.5$). With both factors, 0.0300 becomes 0.0385. On iTransformer ETTh1 at T = 96, 0.0035 becomes 0.0049 ($t = -16$ against the unmatched arm), and 0.0028 becomes 0.0048 with both factors. On PatchTST ETTh1 at T = 96, where the unmatched temporal factor is not separated from DF (+0.0012), the scale-matched arm is worse than it by 0.0011 ($t = 7.7$), and with both factors worse than DF beyond noise (+0.0082). **The direction weights therefore set the sign of the effect and the global scale its size, on these cells.** The scale-matched arm is a diagnostic: it removes the interpolation to DF that the shrinkage provides.

F.10 Direction or Spectrum? A Same-Spectrum Rotation Control

Appendix F.9 shows that the whitened term's global scale sets the size of an effect and not its sign, and Appendix F.4 that zeroing the off-diagonals costs most of the gain while changing the spectrum too. Neither separates *which directions* the metric weights from *how* it weights them. This control does, at two cells of PatchTST on ETTh1 under the common checkpoint rule, with the guarded module and three seeds. One is T = 720, the cell with the paper's largest effect. The other is T = 336, the one cell of that family whose gain survives Holm.

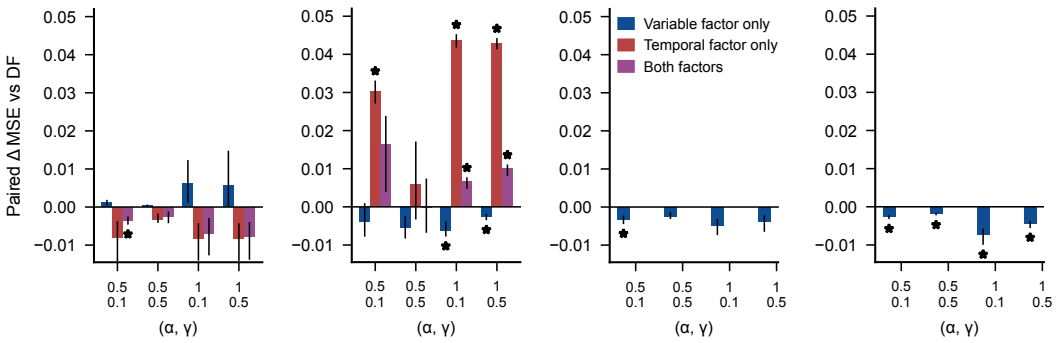
For each seed we train DF, then run the trained model over its own training split and record the residual second moment $\hat{\Sigma}_t$ (720 × 720 from 7,825 windows at T = 720). Three fixed random orthogonal matrices Q_1, Q_2, Q_3 give the rotated statistics $Q_j \hat{\Sigma}_t Q_j^\top$, and WhiteDF$_t$ is then trained with the temporal factor frozen at $\hat{\Sigma}_t$ or at one of the rotations. A rotation changes only the eigenvectors: the eigenvalues agree to 2×10^{-8} (4×10^{-8} at T = 336), so the shrunk and trace-normalised factor, its condition number and its mean precision are identical across the four frozen arms of a cell, while the overlap between the original and rotated eigenvectors is at most 0.18 (0.27). Whatever a rotated arm loses is therefore not scale, conditioning or spectrum.

Table 11: Same-spectrum rotation control on PatchTST, ETTh1, T = 720 and T = 336, three seeds. Mean paired difference in test MSE ($\times 10^{-3}$) against DF, and against the unrotated frozen factor. Bold marks $p < 0.05$ against DF. Within a cell, the four frozen arms share the eigenvalues, trace, condition number and mean precision of $\hat{\Sigma}_t$.

Arm	T	Δ vs DF	t	p	Δ vs frozen $\hat{\Sigma}_t$ (t)
online (paper)	720	**−27.6**	−8.9	0.012	–
frozen $\hat{\Sigma}_t$ (DF training residuals)	720	**−29.0**	−11.2	0.008	–
frozen $Q_1 \hat{\Sigma}_t Q_1^\top$	720	−1.4	−0.8	0.522	+27.6 (+7.0)
frozen $Q_2 \hat{\Sigma}_t Q_2^\top$	720	−1.6	−1.3	0.316	+27.4 (+8.3)
frozen $Q_3 \hat{\Sigma}_t Q_3^\top$	720	+0.3	+0.4	0.717	+29.3 (+15.3)
online (paper)	336	**−9.6**	−13.2	0.006	–
frozen $\hat{\Sigma}_t$ (DF training residuals)	336	**−10.7**	−7.7	0.016	–
frozen $Q_1 \hat{\Sigma}_t Q_1^\top$	336	+0.6	+0.9	0.452	+11.3 (+5.7)
frozen $Q_2 \hat{\Sigma}_t Q_2^\top$	336	+0.3	+0.6	0.634	+11.1 (+8.1)
frozen $Q_3 \hat{\Sigma}_t Q_3^\top$	336	−0.2	−0.4	0.744	+10.6 (+5.8)

The effect does not need the online estimate. Freezing $\hat{\Sigma}_t$ at the DF model's own training residuals reproduces it, −0.0290 against DF ($t = -11.2$) where the online criterion gives −0.0276. It is also specific to the directions of that estimate. Every rotation loses it: −0.0014, −0.0016 and +0.0003 against DF, none separated from zero, and +0.0276, +0.0274 and +0.0293 against the unrotated factor, all three beyond noise, with a mean over rotations of +0.0281 ([0.0153, 0.0409]). A metric with the same spectrum pointed in arbitrary directions gains nothing here. **At this cell the gain therefore depends on the alignment between the weighting and the residual's own structure.** Strength and conditioning are identical in the four frozen arms and cannot account for it.

The pattern repeats at T = 336. The frozen factor gives −0.0107 ($t = -7.7$) where the online criterion gives −0.0096. The rotations give +0.0006, +0.0003 and −0.0002, none separated from DF, and each is worse than the unrotated factor beyond noise (+0.0113, +0.0111 and +0.0106,



(a) ETTh1 (−0.3%) (b) ETTh2 (+10.2%) (c) ETTm2 (+8.3%) (d) Weather (+2.8%)

Figure 13: DLinear under a fixed 40-epoch schedule with frozen two-stage factors at T = 96, four settings (α, γ), three seeds. Bars: mean paired difference against DF. Whiskers: per-seed range. Star: $p < 0.05$. The percentage after each dataset is the closed-form channel-whitening gain. The variable factor alone helps on the three datasets for which the closed form predicts a channel gain and on none of ETTh1's four settings. The frozen temporal factor is a preconditioner the optimiser cannot absorb on ETTh2. Only the variable factor was run on ETTm2 and Weather.

$t = 5.7$, 8.1 and 5.8). The DF and online runs of this control reproduce the paper's runs of that cell seed by seed.

F.11 DLinear and the Two-Stage Estimator

Figure 13 draws the two-stage results. The online DLinear numbers come from the guarded rerun of Appendix F.2. The two-stage factors are estimated once from training residuals and never updated, so the module version does not matter for them.

DLinear online: no usable gain, and validation agrees. Over twenty-four cells, the temporal factor is better beyond noise in one (ETTh1 at T = 96, by 0.0007), worse in seven and not separated in sixteen. The two-factor variant is better in two (ETTm1 at T = 96 and 336, by at most 0.0006), worse in seven and not separated in fifteen. The harm concentrates on ETTh2, where it reaches +0.020 at T = 720, and on ECL, where every cell is worse, as Section 5.3 reports for D = 321 (for the temporal factor this is outside the closed-form account). The criterion-independent validation MSE favours the whitened run in four of the twenty-four cells for the temporal factor and in nine for the two-factor variant, so validation-based selection would have kept DF on most of this backbone. The reference protocol barely trains the model. DLinear starts at the window-mean initialisation and trains at a learning rate of 10^{-4} halved every epoch, and on ETTh2 at T = 96 it stops at 0.345, where the closed-form least-squares solution of the same function class reaches 0.271 and the closed-form channel-whitened solution 0.243 (Appendix E). A criterion chooses among the solutions the optimiser can reach. When the optimiser stays near initialisation there is nothing to choose. The temporal factor's harm on ETTh2 is itself evidence for Remark 2. That factor cannot change the estimator, so what it changes is the trajectory, and on an undertrained linear model the trajectory is all there is.

On ETTh1 the two splits disagree, in both directions. For the frozen variable factor, test is at or above DF at all four settings (up to +0.0062; the largest t, +2.8, is at +0.0012) while validation prefers the whitened run at all four (−0.003 to −0.009, t from −4.5 to −8.2). For the frozen temporal factor at $(1, 0.5)$, test improves by 0.0082 ($t = -2.7$) while validation worsens by 0.0097, neither beyond noise. Elsewhere the splits agree (ETTh2 and ETTm2 on every setting; Weather on test with validation flat). Throughout, test decides the verdict of a cell, validation is reported beside it, and a cell where the two disagree is flagged. By that rule ETTh1 is "no gain", flagged.

Under a longer fixed schedule, the picture does not improve. As a diagnostic outside the reference protocol (learning rate 5×10^{-3}, cosine schedule, 40 epochs, no early stopping), DF improves to 0.390 and 0.297 on ETTh1 and ETTh2 at T = 96. Whitening then helps in one of four dataset–factor pairs, ETTh1 at T = 96 (−0.0032, $t = -4.4$, temporal factor), and is not separated elsewhere.

The $+10\%$ that the closed-form channel factor delivers on ETTh2 is not realised by the online estimate at the fixed setting ($\Delta = -0.00002$), and weaker shrinkage reverses it monotonically, to $+0.024$ ($t = +4.2$) at $\gamma = 0.1$ for the two-factor variant and $+0.078$ with the fusion removed. The closed form estimates Σ_c once, from a converged fit; the online estimate whitens by the residual structure of a model still in motion.

Estimating once, then whitening, separates the two. The obvious repair is the two-stage estimator. It takes Σ_t and Σ_c from the residuals of the closed-form fit, freezes them, and trains under the fixed metric (three seeds, same 40-epoch protocol, $\alpha \in \{0.5, 1\}$, $\gamma \in \{0.1, 0.5\}$, at $\mathrm{T} = 96$). The variable factor *alone* now helps wherever the closed form says it should and nowhere else. On ETTh2, ETTm2 and Weather, for which Appendix E predicts channel gains of 10.2%, 8.3% and 2.8%, it lowers test MSE at all twelve settings, by 0.0018 to 0.0073 (t from -1.5 to -8.3), beyond noise at seven (the largest, at $\alpha = 1$, $\gamma = 0.1$, are -0.0062, -0.0048 on ETTm2, not separated, and -0.0073), with validation agreeing in sign at all but one. On ETTh1, for which it predicts -0.3%, it helps at none of four ($+0.0004$ to $+0.006$). The realised gains are a fifth to a third of the closed-form ones on ETT, larger than predicted on Weather, and come from the same estimator through an optimiser rather than a solve. The temporal factor alone, which Remark 2 says cannot move the solution, costs $+0.030$ to $+0.044$ ($t > 17$) on ETTh2 at every setting but $(0.5, 0.5)$. The closed-form Σ_t has a dynamic range above 10^3, and whitening by it is a preconditioner that first-order optimisation cannot absorb in 40 epochs. Adding it to the variable factor erases the variable factor's gain. On ETTh1 the temporal factor alone improves test MSE by 0.003 to 0.008 at every setting while worsening validation MSE at every setting (neither beyond noise), the same disagreement between the two ETTh1 splits that PatchTST shows (Appendix F.2). On a linear backbone the variable factor therefore does what Corollary 2 says and no more, the temporal factor is an optimiser effect in both directions, and the online estimate at the fixed setting captures neither.

The same two-stage estimator on iTransformer says the variable factor is a channel-independence effect. Freezing Σ_t and Σ_c taken from the trained DF model of the same seed (its training-split residuals) and retraining iTransformer under the fixed metric, at $\mathrm{T} = 96$ with the reference protocol, gives the complement of the DLinear result. On ETTh2, where the closed form and PatchTST both put a channel gain, the frozen variable factor alone does nothing on this channel-mixing backbone ($+0.0005$ and -0.0001 at the two settings, neither separated from DF). Corollary 2 leads one to expect this for a model that already conditions every output on every variable. On ETTh1 the variable factor gives -0.0005 and $+0.0006$. The frozen two-factor criterion at $(0.5, 0.5)$ reproduces the online gain on ETTh1 (-0.0033, $t = -8.4$, against -0.0035 for the online estimate on the same seeds), so that gain does not depend on the estimate following the model during training, and at $(1, 0.1)$ it costs $+0.014$ on ETTh1 ($t = +14.8$) and $+0.002$ on ETTh2, the same conditioning failure that the frozen temporal factor produced on DLinear.

F.12 How Much of the Frozen Factor's Cost Is Finite Training?

The paragraphs above call the frozen temporal factor's cost an optimiser effect. This subsection measures it, on the class for which Remark 2 makes the exact statement: the shared-weight channel-independent linear map, $\hat{Y}_n = WX_n$, at $\mathrm{H} = \mathrm{T} = 96$. Three quantities are compared: the closed-form unregularised minimiser $W^\star$, the training objective each trained arm reaches, and the distance between the two. The optimiser, budget and schedule are those of the two-stage runs above (Adam, batch 32, learning rate 5×10^{-3}, cosine, 40 epochs, no early stopping, three seeds), and the frozen factor is the second moment of the closed-form training residuals, shrunk and trace-normalised exactly as in Section 4.2. No regularisation is used anywhere, so the setting matches the remark's unregularised optimum rather than the ridge of Appendix E.

The cancellation itself holds numerically. Substituting $V = \Sigma_t^{-1/2} W$ turns the whitened objective into an ordinary least-squares problem whose solution is $\Sigma_t^{-1/2} W^\star$, and the two agree to 10^{-13} at both shrinkages on both datasets. So every arm in Table 12 has the same exact minimiser, and what separates them is how close 40 epochs come to it.

Under the plain squared loss the optimiser effectively reaches the minimiser, within 0.02 in parameter distance. Under the pure whitened loss it does not: at $\alpha = 1$, $\gamma = 0.1$ the distance is 0.149 on ETTh1 and 0.372 on ETTh2, the training gap is fifty and two hundred times the plain one, and both order with the condition number of the unshrunk second moment (406 and 1054). Raising

Table 12: Finite training under a frozen temporal factor, shared-weight linear class, three seeds. "gap": training MSE minus the closed-form minimum ($\times 10^{-3}$); "dist.": $\|W - W^\star\|_F / \|W^\star\|_F$; "$\Delta$ test": test MSE minus the closed form's ($\times 10^{-3}$). All arms share the exact minimiser $W^\star$. "cond." is the condition number of the unshrunk second moment $\hat{S}_t$ from which the frozen factor is built.

	ETTh1 (cond. 406)			ETTh2 (cond. 1054)		
Training criterion	gap	dist.	Δ test	gap	dist.	Δ test
plain squared loss	0.005	0.009	+0.01	0.012	0.017	+0.43
whitened, $\alpha = 0.5$, $\gamma = 0.5$	0.006	0.009	+0.00	0.014	0.016	+0.23
whitened, $\alpha = 0.5$, $\gamma = 0.1$	0.006	0.008	−0.02	0.014	0.018	+0.09
whitened, $\alpha = 1$, $\gamma = 0.5$	0.025	0.053	−0.06	0.386	0.196	+0.02
whitened, $\alpha = 1$, $\gamma = 0.1$	0.251	0.149	−0.10	2.394	0.372	+2.39

the shrinkage to $\gamma = 0.5$ cuts the distance to 0.053 and 0.196, and **keeping half the plain term ($\boldsymbol{\alpha = 0.5}$) returns both quantities to the plain-loss values**. The effect on test MSE is small in this class and separates only in that one arm on ETTh2 ($+0.0024$).

F.13 Rolling-Origin Folds on ETTh1

Every number elsewhere in this paper comes from one chronological split per dataset, so the uncertainty we report is seed variation at a fixed test period. This subsection repeats the PatchTST ETTh1 comparison on later periods. The reference protocol takes the first 12 months for training, the next 4 for validation and the next 4 for test. Here every border is moved forward by 720, 1,440, 2,160 or 2,880 hours (one to four months). At four months this gives training rows 2,880–11,520, validation 11,424–14,400 and test 14,304–17,280 of the 17,420 available, so that test period lies almost entirely beyond the reference protocol's last row. The test period of a fold shifted by k months shares $4 - k$ of its four months with the reference one, so the folds overlap and are not independent. The standardisation and every statistic are fitted on each shifted training segment alone, and the loss module, checkpoint rule (validation MSE), hyperparameters and seeds are otherwise unchanged. Horizons 96 and 720, three seeds, arms DF, time-domain ℓ_1, FreDF and WhiteDF$_t$ at the paper's fixed setting.

Table 13: PatchTST on ETTh1 under the reference split and under folds shifted forward by one to four months (720 to 2,880 hours). DF is that fold's mean test MSE, and the other columns are mean paired differences against that fold's own DF ($\times 10^{-3}$, three seeds) with the paired t. Bold marks $p < 0.05$, uncorrected. Later periods are harder for every arm.

Shift	T	DF	time ℓ_1 − DF		FreDF − DF		WhiteDF$_t$ − DF	
		MSE	Δ	t	Δ	t	Δ	t
reference	96	0.379	−2.0	−1.9	+0.5	+0.4	+1.2	+1.1
+720 h	96	0.417	**−3.3**	−4.9	−1.7	−1.3	−1.6	−1.4
+1440 h	96	0.467	**−3.1**	−5.9	+0.6	+0.5	+1.6	+1.1
+2160 h	96	0.476	+0.5	+1.5	+3.3	+3.9	**+5.9**	+6.8
+2880 h	96	0.468	+2.1	+2.9	+5.8	+3.6	**+9.1**	+14.3
reference	720	0.521	−2.4	−0.3	**−18.4**	−8.4	**−27.6**	−8.9
+720 h	720	0.595	**−14.9**	−53.2	**−20.7**	−29.7	**−27.5**	−35.9
+1440 h	720	0.727	−5.3	−4.2	**−7.9**	−5.1	−8.7	−3.3
+2160 h	720	0.810	−11.0	−4.2	−1.2	−0.5	+1.2	+0.6
+2880 h	720	0.816	−2.1	−3.4	−8.6	−0.8	+0.8	+0.1

The largest effect this paper reports fades with the shift. At T = 720 WhiteDF$_t$ is 0.0276 below DF on the reference split, 0.0275 below on the fold shifted by one month ($t = -35.9$), 0.0087 below at two months ($t = -3.3$, not separated), and 0.0012 and 0.0008 *above* at three and four months. **The gain is therefore concentrated in the reference test period and absent from the months after it.** Against the time-domain norm, WhiteDF$_t$ goes from 0.0253 ahead on the reference split, beyond noise (Table 9), to 0.0029 behind at four months, not separated. FreDF's gain at the same cell follows a similar path (-0.0184, -0.0207, -0.0079, -0.0012 and -0.0086), and the time-

domain norm shows no consistent trend. At $\mathrm{T} = 96$ WHITEDF$_\mathrm{t}$ is not separated from DF on the reference split and at one and two months, and is worse at three and four months ($+0.0059$, $t = 6.8$, and $+0.0091$, $t = 14.3$). At four months it is also worse than the norm, by 0.0070. The ordering at $\mathrm{T} = 96$ that the paper already reports, the norm ahead of WHITEDF$_\mathrm{t}$, holds on the reference and the four-month fold.

Selecting on validation MSE, with DF among the candidates, gives the same reading from the other side. On the reference split validation picks the time-domain norm at $\mathrm{T} = 96$, 192 and 336 and DF at $\mathrm{T} = 720$. It never selects WHITEDF$_\mathrm{t}$, including at the cell where WHITEDF$_\mathrm{t}$ wins the test by 5.3%. On the four-month fold validation does pick WHITEDF$_\mathrm{t}$ at $\mathrm{T} = 720$, and the test difference it collects is $+0.0008$. **A practitioner following that rule would therefore never have collected the gain**, which is consistent with the disagreement between the two ETTh1 splits reported in Appendix F.2.

Later periods are harder for every arm because their data are harder: DF's test MSE at $\mathrm{T} = 720$ rises from 0.521 on the reference split to 0.816 at four months, and the four-month test window, standardised on its own training segment, has variance 1.32 against 1.11. Every arm is compared with the DF of its own fold, so this raises all four MSEs without biasing the differences.

G REPRODUCIBILITY

The criteria of Sections 3–5 are implemented in one module (`losses.py`), whose WHITEDF class maps one-to-one onto the equations of Section 4.2, and WHITEDF's factors are registered as buffers, so a checkpoint carries them. Baselines are the public reference implementations, and the protocol follows Liu et al. (2024a) and Qiu et al. (2024) without modification in input length, horizons, chronological splits, metrics and the disabled drop-last trick at test time. A version-controlled, idempotent patch script makes three changes to the reference code: it reads the seed from the command line (the upstream seed is a hard-coded constant and would hide the noise floor), makes the criterion selectable at the one existing hook, and writes test metrics to JSON. The training loop, data splits, normalisation, learning-rate schedule and early stopping are untouched, so early stopping uses each criterion's own validation loss, and Appendix F.8 measures that difference under a common rule. Appendix I.1 describes the versions of the loss module, and two scripts check the analytical claims and exit non-zero on failure (Appendix K). The supplementary material holds the code, every version of the loss module, the raw per-run records, a table that gives the module, code path, checkpoint rule and record file of every result family (Table S1), and one script (`code/regen_all.sh`) that rewrites every result table and data figure of this paper from those records (Figure 1 also reads the public ETTh2 file). The tables it writes are byte-identical to the ones printed here.

H DATASETS

Table 14 lists the benchmarks with the split sizes of Liu et al. (2024a), which we use unchanged, and the dimension-to-sample ratio of Section 5.3 at the full training split with $\mathrm{T} = 96$. Table 2 uses the strided subsample of the closed-form fits. The two orderings differ among ETT and Weather, and both put ECL and Traffic far above the rest.

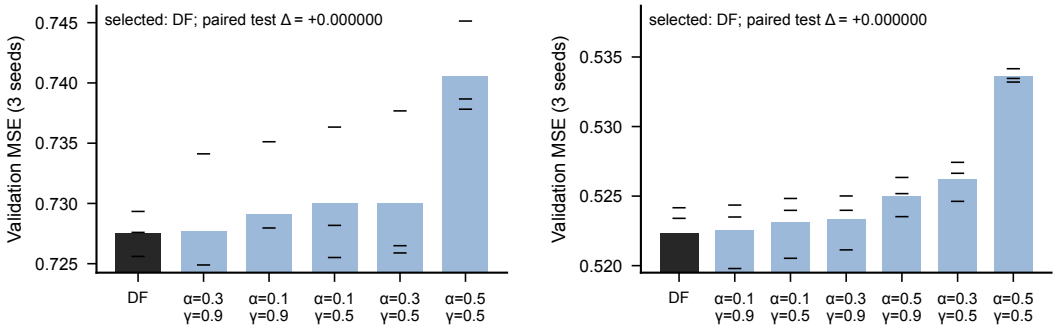
I IMPLEMENTATION DETAILS

Backbones. The main backbone is iTransformer (Liu et al., 2024a), with the reference hyperparameters of its public benchmark, unchanged. WHITEDF is a criterion, not an architecture, so every comparison is within-backbone and against a DF baseline retrained under identical seeds and code paths. Because Corollary 2 is a claim about channel independence, Section 5.2 adds two channel-independent backbones, DLinear and PatchTST, from the Time-Series-Library reference scripts at the same fixed setting, and Section 4.3 trains DLinear with frozen factors on a fixed schedule. The claims are about a criterion, so the backbones are chosen to contrast channel mixing (iTransformer) with channel independence (DLinear, PatchTST).

The FreDF arm. Our FreDF criterion is a re-implementation: the orthonormal rFFT of the residual along time, the mean absolute value of its coefficients, mixed with MSE at $\alpha = 0.5$. The released

Table 14: Benchmark datasets. Splits are chronological, following Liu et al. (2024a); the counts are the reference protocol's convention, the rows of each split minus 95 (windows of input length 96, not reduced by the horizon; the data loader of Wu et al. (2023) yields T fewer complete forecast windows at horizon T, e.g. 8449 ETTh training windows at $T = 96$, the number the closed-form fits of Appendix E use). $D/(nT)$ uses the unreduced n. Overlapping windows are dependent, so n is not an effective sample size.

Dataset	D	Split	Frequency	$D/(nT)$
ETTh1, ETTh2	7	8545 / 2881 / 2881	hourly	8.5×10^{-6}
ETTm1, ETTm2	7	34465 / 11521 / 11521	15 min	2.1×10^{-6}
Weather	21	36792 / 5271 / 10540	10 min	5.9×10^{-6}
ECL	321	18317 / 2633 / 5261	hourly	1.8×10^{-4}
Traffic	862	12185 / 1757 / 3509	hourly	7.4×10^{-4}



(a) iTransformer, Weather, $T = 720$

(b) PatchTST, ETTm1, $T = 192$

Figure 14: Validation-selected shrinkage on the two followed-up cells. Bars: mean validation MSE over three seeds for DF and every whitened candidate of the grid, ranked. Ticks: per-seed values. DF ranks first on both, so the selected criterion is DF and the paired test difference is exactly zero.

FreDF code uses the unnormalised transform, whose coefficients are $\sqrt{T}$ larger, so at the same α the reference code weights its frequency term 9.8 to 26.8 times more than ours over the tested horizons. Its released iTransformer scripts also tune the mix per dataset, with frequency weight 0.8 against 0.2 for MSE on ETTh1 and ETTm1, 0.9 against 0.1 on ETTh2, ETTm2 and ECL, and 0.6 against 0.4 on Weather. Relative to the MSE term the released frequency term is therefore $1.5\sqrt{T}$ to $9\sqrt{T}$ times heavier than ours, and our arm tests FreDF's basis and norm at an equal mix, not at its tuned operating point. Appendix F.7 runs the released weighting on iTransformer, and Appendix F.3 calibrates the time-domain ℓ_1 control against the frequency term.

Optimisation. Every trained number inherits the optimiser, learning rate, schedule, batch size, epoch budget and early-stopping rule of the backbone's public reference script, except where a section says otherwise: the fixed 40-epoch schedule of Section 4.3 and Appendices F.11 and F.12, the common checkpoint rule, and QDF's repository (Appendix F.6). Nothing is re-tuned for WHITEDF, and the DF baseline in each cell is trained by the same script with the criterion switched off. For iTransformer these are the settings of Liu et al. (2024a). For the channel-independent backbones they are the per-dataset scripts of the Time-Series-Library (Wu et al., 2023), whose PatchTST configuration varies by dataset and horizon and is used as found.

Selecting on validation. One rule is to choose the shrinkage on validation MSE, DF among the candidates. On two cells where the fixed setting was above DF (iTransformer Weather $T = 720$, beyond noise, and PatchTST ETTm1 $T = 192$, in mean) validation ranks DF first over the grid (Figure 14), so the selected criterion is DF and the test difference is zero: the harm is removed and nothing is gained. Both cells were chosen after the harm was seen, so they illustrate the rule without validating it (Appendix F.13 applies it to PatchTST on ETTh1).

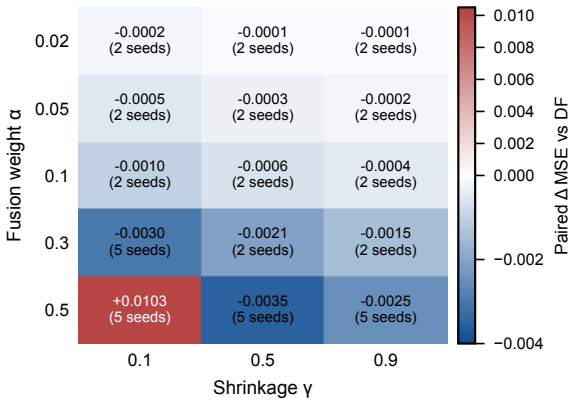


Figure 15: Sensitivity of the criterion to (α, γ) on the development cell (iTransformer, ETTh1, $\mathrm{T} = 96$). Each cell shows the mean paired difference against DF and the number of seeds behind it (two from the grid search, five from the confirmation run). At $\alpha = 0.5$ the $\gamma = 0.1$ column is the single worst cell and the $\gamma = 0.5$ column the optimum.

Criterion hyperparameters. The grid searched on the development cell (iTransformer, ETTh1, $\mathrm{T} = 96$) was $\alpha \in \{0.02, 0.05, 0.1, 0.3, 0.5\}$ and $\gamma \in \{0.1, 0.5, 0.9\}$, two seeds per cell, with the three best settings and a control arm at $(0.5, 0.1)$ re-run at five seeds. The control had to reproduce its known harm, and did, to five decimal places against the independent earlier run. The selected setting is $(\alpha, \gamma) = (0.5, 0.5)$ (Figure 15). The setting $\alpha = 0.5$, $\gamma = 0.1$, our default before the search, is the single worst cell of the grid. The EMA rate $\beta = 0.99$ and the refresh interval $R = 100$ steps were fixed throughout and never searched. Every other cell in the paper uses the selected setting unchanged, including the cells on backbones it was never tuned on. A single setting transferred to held-out cells distinguishes a method that works from a grid wide enough that some cell always wins. Per-cell tuning, the customary protocol, would report better numbers and cannot make that distinction.

I.1 Loss-Module Versions, Hardware and Seeds

Two training code paths and five versions of the loss module were used. On the iTransformer path the reference loop evaluates on CPU tensors while the criterion's buffers live on the GPU, and every version skips its statistics update when the input is not on the buffers' device, so the running second moments saw training batches only. We verified this directly: five evaluation calls on CPU tensors leave the buffers bit-identical, and five same-device calls under `no_grad` change them. On the Time-Series-Library path the reference loop evaluates on the training device, so the earliest version (M0) also updated its statistics on the per-epoch validation and test passes. The later versions update only when gradients are enabled (M1, and M3 and M4, which add per-step logging, the basis controls of Appendix F.3 and the two criteria of Appendix F.7), or carry the device guard only, on paths where it suffices (M2, used on iTransformer for the candidate criteria and the common-rule reruns, and inside QDF's repository, where the whitened term is called in the training step only). Every iTransformer number was produced on the guarded path. The result families that rested on M0 with statistics updated on held-out batches, the online DLinear and PatchTST grids and the ETTm1 $\mathrm{T} = 192$ validation-selection grid on PatchTST, were rerun in full with M1 under the common checkpoint rule, with a DF baseline trained by the same script (447 training runs). The rerun agrees with the earlier verdict on 81 of 96 cells (Table 16), and the validation grid again ranks DF first (Appendix I). Supplementary Table S1 gives, for every result family, the code path, the module version and its hash, the batches that updated the statistics, the checkpoint rule and the record file.

Hardware and seeds. iTransformer runs used one NVIDIA RTX A6000 (48 GB) per run. The two-stage DLinear runs used a 128-core CPU node, the guarded DLinear and PatchTST reruns ran on a GPU, and the superseded earlier-module grids were split between the two by dataset, with each cell's DF baseline trained on the same machine as its WhiteDF runs so that no pairing crosses

hardware. The runs of Appendix F.7 used rented GPUs as described there, and the DF model of Figure 11 was trained on the same rented RTX PRO 6000; its test MSE equals that of the A6000 run with the same seed to four decimals. Seeds are $\{42, 43, 44\}$ for three-seed cells and $\{42, \ldots, 46\}$ for five-seed cells, and the seed controls the Python, NumPy, and PyTorch CPU and CUDA generators. Every comparison is paired on seed because WHITEDF's own seed standard deviation on ETTh1 is 0.00094 over five seeds, $1.5\times$ DF's 0.00062, so judging it against the baseline's spread alone would understate the uncertainty.

I.2 HOLM-CORRECTED VERDICTS

Every per-cell verdict in the paper is an uncorrected two-sided paired t-test at $p < 0.05$. Table 15 repeats the same tests with a Holm step-down correction applied within each family, a family being one backbone, one protocol and one contrast over all of its cells: the nineteen held-out iTransformer cells under the reference rule (five seeds for the DF, FreDF, WHITEDF$_t$ and time-domain-norm arms, three when a candidate of Appendix F.5 is involved), the eight hourly-ETT PatchTST cells under the common rule, and the twenty-four-cell reruns of both channel-independent backbones. Holm controls the family-wise error rate under arbitrary dependence between the cells, so it needs no independence between the horizons of one dataset, and with three to five seeds its power is low. The families were defined after the results were seen, so the table is a sensitivity analysis. The last column averages the four horizons of each dataset and then the datasets with equal weight. It is descriptive.

Few verdicts survive. On iTransformer, WHITEDF$_t$ keeps three of its six gains against DF (ETTh1 at $\mathrm{T} = 192$ and 336, ETTm1 at $\mathrm{T} = 192$) and loses its one harm. The time-domain norm keeps fifteen of its eighteen gains against DF and four of its twelve against FreDF (ETTm1 at $\mathrm{T} = 192$, ETTm2 at 96, Weather at 192 and 720). Nothing survives in the WHITEDF$_t$-versus-FreDF family, and against the norm seven of WHITEDF$_t$'s fourteen losses remain. On the eight PatchTST cells the temporal factor keeps one of four gains (ETTh1 at $\mathrm{T} = 336$, $p_{\mathrm{Holm}} = 0.046$), the two-factor variant one of three (the same cell), the norm two of three, and FreDF none of four. The $\mathrm{T} = 720$ gain of Section 5.2 has $p_{\mathrm{Holm}} = 0.087$ against DF, 0.078 against FreDF and 0.188 against the norm, while the norm's advantage over WHITEDF$_t$ at $\mathrm{T} = 96$ survives (0.026). Over the twenty-four-cell reruns only the four ECL harms survive for either factor on either backbone. No sign reverses, and the correction marks which uncorrected verdicts are strong enough to stand alone.

J FULL RESULTS

Global test. The pooled tests in this paragraph are descriptive, since the four horizons of one dataset are not independent replicates; the verdicts rest on the per-cell tests and their Holm correction (Appendix I.2). Over the nineteen held-out mean paired differences of Table 20 (five seeds each), a Wilcoxon signed-rank test gives $p = 0.08$ and the sign test $11/19$ negative ($p = 0.65$), median -0.0008. The fixed setting does not shift the benchmark set as a whole. The same test in the other direction: against FreDF on the same nineteen cells WHITEDF$_t$ is behind as a group (Wilcoxon $p = 0.003$, five of nineteen negative), and FreDF itself is ahead of DF as a group ($p = 0.001$, fifteen of nineteen negative). The whitened ℓ_1 criterion of Appendix F.5 is ahead of DF ($p < 0.001$) and level with FreDF ($p = 0.28$). Over the twenty-four cells of each channel-independent grid the group test is indifferent or adverse: DLinear's temporal factor is worse as a group ($p = 0.03$), its two-factor variant ($p = 0.25$) and PatchTST's two factors ($p = 0.81$ and 0.44) are not separated from zero. On these backbones every cell has three seeds and the per-cell test needs $|t| > 4.30$, so Sections 4.3 and 5 draw group conclusions from cross-cell counts and the two-stage and common-rule reruns rather than from single cells.

Tables 20–22 list every trained cell at the fixed setting $(\alpha, \gamma) = (0.5, 0.5)$: mean MSE and MAE over the seeds paired with DF, and the paired MSE difference with its t-statistic, bold where $p < 0.05$ (two-sided paired t). The table bodies are generated from the raw per-run records by the released scripts. On iTransformer, ETTh1 at $\mathrm{T} = 96$ is the development cell on which the setting was chosen and is listed for completeness.

Table 15: Nominal and Holm-corrected counts of better / worse / not separated for the first arm against the second, per family; $\bar{\Delta}$: mean paired difference in test MSE, horizons averaged within each dataset and datasets weighted equally.

Contrast	cells	nominal	Holm	$\bar{\Delta}$ ($\times 10^{-3}$)
iTransformer, reference rule, 19 held-out cells				
WHITEDF$_t$ – DF	19	6/1/12	3/0/16	–2.10
FreDF – DF	19	9/1/9	3/0/16	–2.98
WHITEDF$_t$ – FreDF	19	1/4/14	0/0/19	+0.88
time ℓ_1 – DF	19	18/0/1	15/0/4	–4.17
time ℓ_1 – FreDF	19	12/0/7	4/0/15	–1.19
WHITEDF$_t$ – time ℓ_1	19	0/14/5	0/7/12	+2.07
whitened ℓ_1 – DF	19	10/0/9	2/0/17	–3.58
whitened ℓ_1 – time ℓ_1	19	0/4/15	0/0/19	+0.34
whitened FreDF ℓ_1 – DF	19	7/1/11	1/0/18	–2.56
whitened FreDF ℓ_1 – time ℓ_1	19	1/5/13	0/1/18	+1.36
PatchTST, guarded module, common rule, 8 hourly-ETT cells				
FreDF – DF	8	4/0/4	0/0/8	–6.50
time ℓ_1 – DF	8	3/0/5	2/0/6	–4.09
WHITEDF$_t$ – DF	8	4/0/4	1/0/7	–8.78
WHITEDF$_{t+c}$ – DF	8	3/0/5	1/0/7	–8.81
WHITEDF$_t$ – time ℓ_1	8	1/1/6	0/1/7	–4.69
WHITEDF$_t$ – FreDF	8	1/1/6	0/0/8	–2.29
PatchTST, guarded module, common rule, 24 cells				
WHITEDF$_t$ – DF	24	4/7/13	0/4/20	+0.67
WHITEDF$_{t+c}$ – DF	24	3/7/14	0/4/20	+4.39
DLinear, guarded module, common rule, 24 cells				
WHITEDF$_t$ – DF	24	1/7/16	0/4/20	+1.76
WHITEDF$_{t+c}$ – DF	24	2/7/15	0/4/20	+1.75

Table 16: Channel-independent backbones at the fixed setting $(\alpha, \gamma) = (0.5, 0.5)$: rerun with the guarded loss module under the common checkpoint rule (Appendix I.1). Cells are dataset×horizon pairs judged by the two-sided paired t over three seeds at $p < 0.05$; "$\bar{\Delta}$" is the mean of the cell means over the four ETT sets, over Weather and over ECL; "as before" counts the cells whose verdict agrees with the earlier-module run (Appendix F.2).

Backbone	Factor	better	worse	n.s.	cells	$\bar{\Delta}$ ETT	$\bar{\Delta}$ Weather	$\bar{\Delta}$ ECL	as before
DLinear	WHITEDF$_t$	1	7	16	24	+0.0020	–0.0002	+0.0028	21 / 24
DLinear	WHITEDF$_{t+c}$	2	7	15	24	+0.0014	–0.0002	+0.0051	21 / 24
PatchTST	WHITEDF$_t$	4	7	13	24	–0.0033	+0.0011	+0.0161	19 / 24
PatchTST	WHITEDF$_{t+c}$	3	7	14	24	–0.0020	+0.0015	+0.0331	20 / 24

K NUMERICAL VERIFICATION OF THE ANALYTICAL CLAIMS

A script released with the code checks the analytical results numerically on random instances: Parseval equality for the unweighted squared frequency loss and its failure under the ℓ_1 norm (Corollary 1); the circulant-GLS identity (6) (Theorem 1); the separable whitening identity (Proposition 2); the coincidence of OLS and GLS under shared regressors (Corollary 3); their divergence under channel independence with non-diagonal Σ_c and their re-coincidence once Σ_c is diagonalised (Corollary 2), and the positive semi-definiteness of $\mathrm{Cov}(\hat{\beta}_{\mathrm{ols}}) - \mathrm{Cov}(\hat{\beta}_{\mathrm{gls}})$ (Theorem 2); the two-output risk gap (10) (Proposition 4); the detached-update identity (11) with its shrinkage form (Appendix C.8); and the noise-plus-bias decomposition (9) of the residual second moment, checked by Monte Carlo with the relative Frobenius error decaying as $n^{-1/2}$ (Remark 3). All nineteen checks pass.

Those checks exercise the algebra on random matrices. A second, stronger confirmation is run through an actual differentiable model. At fixed weights and a fixed batch, the equal-weight squared frequency loss and the DF loss agree to zero measured relative difference, and their gradients with respect to all model parameters agree with cosine similarity 1.000000 and relative ℓ_2 difference

Table 17: FreDF against WHITEDF$_t$, both paired on seed against the same DF runs, on every iTransformer cell where both were trained (FreDF at its default $\alpha = 0.5$; five seeds on every cell; † marks the development cell). Bold: $p < 0.05$. FreDF is better than WHITEDF$_t$ beyond noise on five cells (the development cell among them) and worse on one (ETTm1, T = 192); the squared-norm criterion does not exceed the ℓ_1 criterion overall.

Dataset	T	DF	FreDF	WHITEDF$_t$	FreDF – DF	WHITEDF$_t$ – DF	WHITEDF$_t$ – FreDF
ETTh1†	96	0.387	0.383	0.384	**–0.0042 (–26.90)**	**–0.0035 (–13.95)**	**+0.0007 (+3.42)**
ETTh1	192	0.442	0.438	0.438	**–0.0041 (–7.61)**	**–0.0042 (–12.79)**	–0.0001 (–0.38)
ETTh1	336	0.484	0.478	0.478	**–0.0061 (–4.44)**	**–0.0058 (–6.99)**	+0.0003 (+0.35)
ETTh1	720	0.508	0.502	0.502	**–0.0062 (–3.15)**	–0.0060 (–2.46)	+0.0002 (+0.09)
ETTh2	96	0.301	0.300	0.300	–0.0006 (–0.78)	–0.0008 (–2.13)	–0.0002 (–0.18)
ETTh2	192	0.381	0.377	0.382	**–0.0039 (–3.89)**	+0.0011 (+0.81)	**+0.0050 (+3.75)**
ETTh2	336	0.426	0.421	0.422	–0.0057 (–1.76)	–0.0044 (–1.24)	+0.0014 (+1.14)
ETTh2	720	0.429	0.427	0.430	–0.0020 (–0.73)	+0.0011 (+0.29)	+0.0030 (+2.36)
ETTm1	96	0.339	0.333	0.333	**–0.0058 (–6.86)**	**–0.0060 (–3.23)**	–0.0002 (–0.17)
ETTm1	192	0.381	0.377	0.376	**–0.0046 (–4.87)**	**–0.0054 (–6.99)**	**–0.0008 (–2.81)**
ETTm1	336	0.422	0.416	0.416	**–0.0069 (–4.19)**	**–0.0067 (–4.22)**	+0.0002 (+0.23)
ETTm1	720	0.492	0.484	0.488	**–0.0076 (–11.52)**	**–0.0042 (–3.04)**	+0.0033 (+1.88)
ETTm2	96	0.183	0.183	0.183	–0.0008 (–2.52)	–0.0005 (–1.45)	+0.0003 (+0.83)
ETTm2	192	0.252	0.250	0.250	**–0.0019 (–3.70)**	–0.0017 (–2.60)	+0.0002 (+0.36)
ETTm2	336	0.314	0.315	0.314	+0.0005 (+0.30)	+0.0003 (+0.24)	–0.0002 (–0.12)
ETTm2	720	0.411	0.410	0.412	–0.0009 (–0.92)	+0.0008 (+0.87)	+0.0018 (+1.51)
Weather	96	0.180	0.179	0.181	–0.0006 (–1.36)	+0.0009 (+1.97)	**+0.0015 (+5.41)**
Weather	192	0.228	0.229	0.230	+0.0008 (+1.77)	+0.0016 (+2.75)	**+0.0008 (+3.21)**
Weather	336	0.284	0.284	0.285	+0.0007 (+1.54)	+0.0010 (+2.34)	+0.0003 (+0.80)
Weather	720	0.360	0.362	0.362	**+0.0016 (+3.89)**	**+0.0022 (+5.72)**	**+0.0006 (+3.35)**

Table 18: The AR(1) whitening that stands in for Sun et al. (2021) (three seeds, five on the development cell; a residual whitened by a jointly estimated AR(1) coefficient, applied to the multi-output head) against WHITEDF$_t$, both at the untuned default $(\alpha, \gamma) = (0.5, 0.1)$, the only setting at which AR(1) was run and the worst cell of WHITEDF's own grid. Each is worse than DF on six of twelve cells; WHITEDF$_t$ is the less harmed on nine. The comparison is mainly one of harms; WHITEDF$_t$ is better than DF on two ETTm1 cells and AR(1) on none.

Dataset	T	DF	AR(1)	WHITEDF$_t$	AR(1) – DF	WHITEDF$_t$ – DF
ETTh1	96	0.388	0.398	0.397	**+0.0098 (+25.20)**	**+0.0096 (+23.06)**
ETTh1	192	0.442	0.452	0.453	**+0.0102 (+16.30)**	**+0.0104 (+18.25)**
ETTh1	336	0.482	0.498	0.501	**+0.0162 (+24.58)**	**+0.0183 (+25.86)**
ETTh1	720	0.507	0.519	0.525	+0.0120 (+2.45)	+0.0178 (+4.03)
ETTm1	96	0.338	0.339	0.332	+0.0014 (+0.89)	**–0.0053 (–6.71)**
ETTm1	192	0.380	0.380	0.375	+0.0000 (+0.05)	**–0.0049 (–9.94)**
ETTm1	336	0.423	0.425	0.416	+0.0018 (+1.08)	–0.0072 (–3.16)
ETTm1	720	0.491	0.493	0.490	+0.0016 (+0.61)	–0.0008 (–0.47)
Weather	96	0.180	0.183	0.182	+0.0028 (+3.89)	**+0.0014 (+12.00)**
Weather	192	0.229	0.233	0.232	**+0.0048 (+5.52)**	**+0.0030 (+4.97)**
Weather	336	0.283	0.287	0.286	**+0.0037 (+4.56)**	+0.0030 (+4.11)
Weather	720	0.360	0.365	0.363	**+0.0042 (+4.66)**	**+0.0027 (+7.24)**

1.8×10^{-7}, which is float32 rounding. The control matters as much as the test: FreDF's ℓ_1 frequency gradient has cosine similarity only 0.912 with the DF gradient under the same conditions. So the two criteria that differ only in *basis* are empirically the same criterion, while the one that differs in *norm* is not. That is the empirical content of Corollary 1 and Remark 4.

This is a fixed-point test by design. The corollary concerns the criterion, whereas end-to-end training can amplify a float-level gradient difference, so the end-to-end comparison is read against a

Table 19: iTransformer on Traffic (D = 862), T = 96, reference configuration, $\alpha = 0.5$, three seeds; each arm is paired on seed against DF retrained in the same sweep (bold: the lower MSE of the pair and $p < 0.05$). The last column is the paired difference in validation MSE, a criterion-independent score.

Criterion	γ	DF	WHITEDF	ΔMSE (t)	Δ val. MSE
WHITEDF$_{\mathrm{t}}$	0.5	0.393	**0.392**	–0.0008 (–1.92)	–0.0008
WHITEDF$_{\mathrm{t}}$	0.9	0.393	**0.392**	–0.0009 (–2.10)	–0.0008
WHITEDF$_{\mathrm{t+c}}$	0.5	0.393	**0.391**	–0.0014 (–3.24)	–0.0007
WHITEDF$_{\mathrm{t+c}}$	0.9	0.393	**0.390**	**–0.0026 (–7.63)**	–0.0025

Table 20: iTransformer, fixed setting, five seeds per cell.

		DF		WHITEDF$_{\mathrm{t}}$		
Dataset	T	MSE	MAE	MSE	MAE	ΔMSE (t)
ETTh1	96	0.387	0.405	**0.384**	**0.400**	**–0.0035 (–13.95)**
	192	0.442	0.436	**0.438**	**0.431**	**–0.0042 (–12.79)**
	336	0.484	0.459	**0.478**	**0.453**	**–0.0058 (–6.99)**
	720	0.508	0.494	**0.502**	**0.491**	–0.0060 (–2.46)
ETTh2	96	0.301	0.350	**0.300**	**0.349**	–0.0008 (–2.13)
	192	**0.381**	**0.400**	0.382	**0.400**	+0.0011 (+0.81)
	336	0.426	0.434	**0.422**	**0.432**	–0.0044 (–1.24)
	720	**0.429**	**0.448**	0.430	**0.447**	+0.0011 (+0.29)
ETTm1	96	0.339	0.375	**0.333**	**0.370**	**–0.0060 (–3.23)**
	192	0.381	0.395	**0.376**	**0.391**	**–0.0054 (–6.99)**
	336	0.422	0.422	**0.416**	**0.417**	**–0.0067 (–4.22)**
	720	0.492	0.460	**0.488**	**0.457**	**–0.0042 (–3.04)**
ETTm2	96	0.183	0.266	**0.183**	**0.265**	–0.0005 (–1.45)
	192	0.252	0.311	**0.250**	**0.310**	–0.0017 (–2.60)
	336	**0.314**	**0.351**	**0.314**	**0.351**	+0.0003 (+0.24)
	720	**0.411**	**0.406**	0.412	**0.406**	+0.0008 (+0.87)
Weather	96	**0.180**	**0.221**	0.181	0.222	+0.0009 (+1.97)
	192	**0.228**	**0.261**	0.230	0.262	+0.0016 (+2.75)
	336	**0.284**	**0.301**	0.285	0.302	+0.0010 (+2.34)
	720	**0.360**	**0.350**	0.362	0.352	**+0.0022 (+5.72)**

measured noise floor (a DF seed range of 0.001525 over five seeds on ETTh1). That the two trained runs turned out identical rather than merely within that floor is stronger than the corollary requires.

L FIGURE NOTES

Details that the main-text captions omit for space, one entry per figure.

Figure 1. Residuals of one fitted model: the closed-form channel-independent shared-weight ridge forecaster on ETTh2 (H = T = 96, D = 7, 8,449 training and 2,855 test windows), the model class to which Corollary 2 and Remark 2 apply exactly. It reproduces its stored test MSE, 0.2705868. (a) Ground truth, forecast and residual of the standardised OT variable over the first test window: the residual drifts for hours instead of resetting at every step. (b) Residual power per frequency under the orthonormal rFFT, averaged over training windows and variables and divided by its mean. It spans nearly three orders of magnitude (a ratio of 824), while an equal-weight frequency loss weights every frequency alike (dotted line at 1). (c) For each of the $19{,}985$ test residual series ($2{,}855$ windows $\times$ seven variables), the loss after the full unitary DFT divided by the same loss in the time domain. For the squared loss the ratio is 1 for every series, to a relative error of 6×10^{-16} (Corollary 1). For the ℓ_1 loss it has median 0.77 and a 5–95% range of 0.41–1.11 (full range 0.18–1.87; the axis holds 98.7% of the series). A change of basis therefore moves an ℓ_1 loss and leaves a squared one

Table 21: DLinear, fixed setting, three seeds: rerun with the guarded loss module under the common checkpoint rule (Appendix F.2).

		DF		WHITEDF$_t$			WHITEDF$_{t+c}$		
Dataset	T	MSE	MAE	MSE	MAE	ΔMSE (t)	MSE	MAE	ΔMSE (t)
ETTh1	96	0.396	0.411	**0.396**	**0.410**	**−0.0007 (−4.35)**	0.397	0.412	+0.0005 (+2.52)
	192	**0.446**	**0.442**	0.447	**0.442**	+0.0005 (+3.46)	0.447	0.443	**+0.0014 (+8.15)**
	336	0.492	0.470	**0.490**	**0.468**	−0.0020 (−0.62)	0.490	0.468	−0.0012 (−0.38)
	720	0.516	0.513	**0.514**	**0.511**	−0.0015 (−0.36)	**0.515**	**0.512**	−0.0011 (−0.26)
ETTh2	96	**0.345**	**0.399**	0.350	0.402	**+0.0046 (+5.94)**	0.348	0.400	+0.0023 (+2.80)
	192	**0.476**	**0.476**	0.481	0.479	+0.0048 (+3.48)	0.479	0.477	+0.0028 (+2.08)
	336	**0.597**	**0.543**	0.603	0.546	**+0.0057 (+10.05)**	0.601	0.545	**+0.0041 (+8.22)**
	720	**0.829**	**0.657**	0.849	0.665	**+0.0199 (+5.07)**	0.847	0.664	**+0.0185 (+4.64)**
ETTm1	96	0.346	0.372	0.346	0.372	+0.0001 (+0.44)	**0.345**	**0.371**	**−0.0006 (−5.07)**
	192	**0.382**	**0.390**	**0.382**	**0.390**	+0.0002 (+4.30)	**0.381**	**0.390**	−0.0001 (−1.65)
	336	**0.413**	**0.413**	**0.413**	**0.413**	+0.0000 (+2.40)	**0.413**	**0.413**	**−0.0001 (−8.72)**
	720	0.474	0.451	**0.472**	**0.449**	−0.0019 (−0.96)	**0.472**	**0.449**	−0.0019 (−0.98)
ETTm2	96	**0.192**	0.290	0.193	0.291	+0.0007 (+2.06)	**0.192**	**0.290**	−0.0004 (−1.85)
	192	0.282	0.359	0.282	0.359	+0.0005 (+0.86)	**0.281**	**0.358**	−0.0008 (−1.86)
	336	0.372	0.420	0.372	0.421	+0.0005 (+0.59)	**0.371**	**0.420**	−0.0005 (−1.27)
	720	**0.549**	**0.521**	0.549	**0.521**	+0.0003 (+0.36)	**0.549**	**0.520**	−0.0002 (−0.28)
Weather	96	0.196	0.257	**0.196**	**0.256**	−0.0004 (−0.61)	**0.196**	**0.256**	−0.0006 (−0.93)
	192	0.237	0.295	**0.236**	**0.294**	−0.0004 (−0.81)	**0.236**	**0.294**	−0.0005 (−0.94)
	336	**0.282**	**0.331**	**0.282**	**0.331**	+0.0000 (+0.08)	**0.282**	**0.332**	+0.0000 (+2.93)
	720	**0.346**	**0.383**	**0.346**	**0.383**	+0.0002 (+2.41)	**0.346**	**0.383**	+0.0002 (+2.47)
ECL	96	**0.210**	**0.302**	0.211	0.302	**+0.0008 (+159.98)**	0.214	0.306	**+0.0039 (+320.81)**
	192	**0.210**	**0.305**	0.212	0.308	**+0.0022 (+1888.22)**	0.215	0.311	**+0.0049 (+1889.99)**
	336	**0.223**	**0.319**	0.226	0.323	**+0.0034 (+248.58)**	0.229	0.326	**+0.0055 (+371.16)**
	720	**0.258**	**0.350**	0.262	0.355	**+0.0047 (+421.09)**	0.264	0.356	**+0.0059 (+401.17)**

untouched. (d) Uncentred normalised second moment $S_{ij}/\sqrt{S_{ii}S_{jj}}$ of the training residuals across the seven variables.

Figure 4. Whiskers span the per-seed paired differences and are truncated at the axis floor where marked. DF's own seed range varies from 0.0012 to 0.0148 across these cells. Every cell is in Table 20.

Figure 6. (a) The $\gamma = 0.1$ bar is off scale at $+10.3$ and drawn broken. (b) On ECL the temporal factor at $\gamma = 0.9$ was run only under the common checkpoint rule (Table 10), not in this sweep. (c) Positive is better, unlike panels a and b. The sign flips between Weather and ECL, and the off-diagonal mass of Σ_c is largest where the gain is most negative and does not order the other outcomes.

M LIMITATIONS AND EXTENSIONS

Four observations remain unexplained: the ETTm1 harm on PatchTST, the ECL failure on iTransformer but not on TQNet, the Traffic gain, and validation missing the ETTh1 test gains. Three further limitations concern the estimate itself. First, separability, $\Sigma \approx \Sigma_c \otimes \Sigma_t$, is adopted for tractability and holds to within 6–27% on ETT (Appendix E). A low-rank correction to the Kronecker form is the natural next step. Second, the estimate is global over the training distribution, whereas residual correlation plausibly varies by regime. A regime-conditional metric would fit the non-stationarity measured in Appendix E more closely. Third, the multi-resolution forecasting heads of models such as Time-MoE (Shi et al., 2025b) average independent per-horizon losses, leaving cross-head error correlation unmodelled. The framework extends there directly, and verifying it needs pre-training scale. The disagreement between the validation and test splits on ETTh1 (Appendix F.2) is the limitation we would most want addressed: a practitioner selecting on validation would discard the largest gain we report, and which split better guides deployment is open. The uncertainty we re-

Table 22: PatchTST, fixed setting, three seeds: rerun with the guarded loss module under the common checkpoint rule (Appendix F.2).

		DF		WHITEDF$_t$			WHITEDF$_{t+c}$		
Dataset	T	MSE	MAE	MSE	MAE	ΔMSE (t)	MSE	MAE	ΔMSE (t)
ETTh1	96	**0.379**	**0.398**	0.380	**0.399**	+0.0012 (+1.09)	0.382	0.400	+0.0028 (+2.34)
	192	0.431	0.434	**0.428**	**0.430**	**−0.0033 (−7.69)**	0.429	0.431	−0.0019 (−4.07)
	336	0.474	0.459	**0.464**	**0.453**	**−0.0096 (−13.17)**	0.465	0.454	**−0.0084 (−13.52)**
	720	0.521	0.505	0.493	**0.496**	**−0.0276 (−8.87)**	**0.491**	0.496	**−0.0300 (−11.45)**
ETTh2	96	0.300	0.350	0.294	0.346	**−0.0054 (−4.96)**	**0.293**	**0.344**	**−0.0071 (−6.05)**
	192	0.377	0.399	0.374	0.397	−0.0030 (−0.68)	**0.373**	**0.396**	−0.0040 (−0.97)
	336	0.433	0.440	**0.421**	**0.432**	−0.0113 (−1.95)	0.423	0.434	−0.0093 (−3.29)
	720	0.442	0.458	0.431	0.451	−0.0114 (−3.50)	**0.429**	**0.450**	−0.0126 (−4.05)
ETTm1	96	**0.331**	**0.368**	0.334	0.370	+0.0033 (+1.49)	0.337	0.373	+0.0066 (+2.40)
	192	**0.369**	**0.390**	0.376	0.395	+0.0067 (+1.92)	0.385	0.401	+0.0158 (+2.66)
	336	**0.401**	**0.410**	0.411	0.415	**+0.0099 (+7.66)**	0.417	0.416	**+0.0153 (+16.91)**
	720	**0.459**	**0.446**	0.464	0.447	**+0.0051 (+6.33)**	0.467	0.449	**+0.0088 (+10.39)**
ETTm2	96	0.184	0.266	**0.182**	0.266	−0.0016 (−0.85)	**0.182**	**0.265**	−0.0017 (−0.79)
	192	0.250	0.311	0.247	0.308	−0.0029 (−1.49)	**0.246**	**0.308**	−0.0036 (−1.85)
	336	0.310	0.348	0.308	0.347	−0.0013 (−1.60)	**0.308**	**0.346**	−0.0020 (−4.24)
	720	0.412	**0.407**	**0.411**	0.408	−0.0016 (−1.04)	**0.411**	0.408	−0.0014 (−0.59)
Weather	96	**0.175**	**0.217**	0.176	0.219	+0.0007 (+1.50)	0.176	0.220	+0.0008 (+1.59)
	192	**0.221**	**0.257**	0.224	0.260	**+0.0023 (+9.26)**	0.224	0.261	**+0.0028 (+7.63)**
	336	**0.281**	**0.299**	0.282	0.301	+0.0011 (+1.86)	0.283	0.302	+0.0019 (+2.62)
	720	**0.356**	**0.348**	**0.356**	0.349	+0.0002 (+0.94)	**0.356**	0.350	+0.0003 (+0.44)
ECL	96	**0.180**	**0.273**	0.191	0.291	**+0.0111 (+47.63)**	0.209	0.319	**+0.0293 (+86.37)**
	192	**0.187**	**0.279**	0.202	0.303	**+0.0144 (+33.01)**	0.218	0.327	**+0.0309 (+191.67)**
	336	**0.204**	**0.296**	0.222	0.322	**+0.0173 (+133.88)**	0.238	0.343	**+0.0336 (+123.24)**
	720	**0.246**	**0.329**	0.268	0.358	**+0.0218 (+46.45)**	0.284	0.379	**+0.0384 (+88.95)**

port is seed variation on one chronological split of each benchmark. Four overlapping rolling-origin folds test that on PatchTST ETTh1 at $\mathrm{T} = 96$ and $\mathrm{T} = 720$: the $\mathrm{T} = 720$ gain holds one month later and is gone by three (Appendix F.13). Overlapping folds on one dataset are not a distribution over periods, so how often a gain of this kind replicates is still unmeasured, and no other dataset has been checked this way. The iTransformer setting was chosen on the test error of ETTh1 at $\mathrm{T} = 96$, and all six held-out cells it improves share that cell's station and test months, so those gains are not independent of the choice. The per-cell verdicts are uncorrected, and under a Holm correction few survive (Appendix I.2). Within the equal mix the norm-versus-basis comparison is calibrated two ways (Appendix F.3) and the ordering does not move, but the basis comparison is a single cell on a single backbone. FreDF at its released weights, and with its weight chosen on validation from three values, is compared with the norm on iTransformer only (Appendix F.7). The transfer grid of Appendix F.5 was not repeated for the bases of Appendix F.3.