ISOKOOP: THE CYCLE COMES FROM OUTSIDE THE WINDOW

Anonymous authors
Paper under double-blind review

ABSTRACT

In long-horizon multivariate forecasting, a look-back window shorter than the cycle it must forecast does not fix that cycle's length: two cycles of different length can agree on the whole window, so the length is absent from it, not hard to fit. Per-window normalisation turns that observation into an exact interface, one that binds published models too: moved in front of its normaliser, FreqCycle's cycle table costs that model up to 210.6% under a gain. A forecaster that standardises each window by its own mean and scale is scale-equivariant for every inner network, and a profile supplied from outside the window keeps that guarantee when subtracted in raw units only if it is a constant level, but at any shape once it enters in the normalised coordinates or is rescaled by the window's scale. The ratio of the dataset cycle to the look-back, fixed by the protocol, says what the window cannot hold, and IsoKoop supplies it. Against eight forecasters retrained in one pipeline on thirteen public benchmarks, four horizons and three seeds, IsoKoop attains the best mean rank in MSE and MAE (2.34 against 3.77 and 2.14 against 3.65) and the lowest MSE in 34 of the 52 settings (2.24 against 3.99 and 38 of 56 with our TrafficC panel). On the 883-variate network it uses a fourteenth of EMAformer's parameters, the runner-up in mean rank. IsoKoop's supply is a rank-R prototype read at the sample index, zero-initialised so that its profile, though not its period, is learned, and an optional level anchor. In our ablations, and in the baselines with no new run, the supply helps most where the cycle exceeds the look-back, on panels that here are also the sensor networks, so ratio and panel type are not separated; on the real shifts we tested, these readings lower the error only on Solar at $H{=}96$, less where the scale moved most, and raise it on PEMS04 and PEMS08. The same lens audits the benchmarks: on Exchange no model beats persistence, many variates change scale from training to test, and nine drifting sensors decide Traffic's ranking.

1 INTRODUCTION

Almost every competitive long-horizon forecaster of the last three years performs the same first and last step: it standardises each look-back window by its own mean and scale, lets a learned operator act, and inverts the standardisation on the forecast. Reversible instance normalisation (Kim et al., 2022) made this step a default, and the operators that follow it range from attention over variate tokens (Liu et al., 2024) and over patches (Nie et al., 2023) to frequency-domain and Koopman operators (Yue et al., 2025; Liu et al., 2023a). The literature treats the step as preprocessing.

We treat it as a *coordinate system*. The normaliser fixes the units of everything the operator sees, and so it decides what a model may add from outside the window and where.

Consider a forecaster that adds knowledge the window does not have. A look-back of L steps holds only part of a longer cycle's profile, and a window that does not close the cycle does not fix its length, so a model that must commit to a length has to take it from outside. Neither depth, nor width, nor attention reads the length off the window (Design remark 5). Recent models supply the cycle from the calendar, through phase-indexed cycle tables (Lin et al., 2024; Zhang et al., 2026a) or phase embeddings (Zhang et al., 2026b). A cycle table is fitted on the series, not the window. Which side of the normaliser it enters decides an exact symmetry (Figure 1).

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

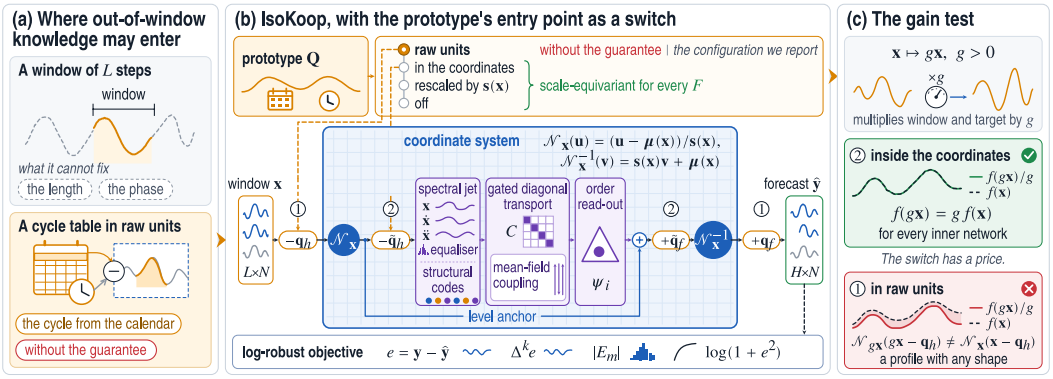


Figure 1: (a) What the window cannot fix, the length and phase of a long cycle, enters from outside. (b) In IsoKoop the entry point is a four-position switch, among them raw units (①) and inside the coordinates (②) of the per-window normaliser $\mathcal{N}_{\mathbf{x}}$; the reported configuration is in raw units, the side without the guarantee of Proposition 1. (c) The gain test scales window and target by g: ② gives $f(g\mathbf{x}) = g\, f(\mathbf{x})$ for every inner network, ① need not.

Proposition 1 makes this exact. A forecaster wrapped in a per-window normaliser with positively homogeneous statistics is scale-equivariant for *every* inner network. A fixed profile subtracted in raw units keeps that guarantee for every network if and only if it is a constant level; one that enters inside the normalised coordinates, or is rescaled by the window's scale, keeps it, and a gain applied to the whole window tests it on any trained model (Figure 1c). The law is the per-window-normalised form of the classical choice between additive and multiplicative seasonality (Winters, 1960; Hyndman & Athanasopoulos, 2021).

IsoKoop turns the ratio P/L of the cycle to the look-back, and the law, into an architecture. The name joins *iso-* (Greek *isos*, equal), for the equivariance law whose entry-point switch the model exposes, to the *Koop*man-style transport it runs in the coordinates (Figure 2). A rank-R prototype read at the sample's own index supplies the phase, an optional level anchor, an AR(1) forecast about the window mean, handles the level, and a spectral jet lifts the remainder (I). A gated diagonal transport with rank-one mean-field coupling advances it in the normalised coordinates (II), and a log-robust objective over the error, its differences and its spectrum trains it (III). The prototype's placement is the switch the proposition describes: the configuration we report puts it in raw units, the side without the guarantee, and Section 4.5 prices both sides.

Contributions.

- **A criterion, not a module.** The ratio of a dataset's cycle to the look-back says, before any run, when a window-only operator cannot fix the cycle's length ($P > L$) and when it can ($P \leq L/2$). We state the exact interface that makes the question well posed: an equivariance law for per-window-normalised forecasters (Proposition 1) that fixes where out-of-window knowledge may enter. The panels with $P > L$ are here also the sensor networks; in our ablations and on the baseline side the supply matters most on them.
- **A machine the criterion motivates.** Three stages answer the three questions the law leaves open: what to supply from outside the window, how to advance it once it is in the coordinates, and in which metric to train so that metric and coordinates agree. Each stage rests on a statement whose hypotheses and departures we give, and we read it out on the trained model and switch it off in a factorial. Where attention of the same depth and width trains, the transport beats it in eight of twelve single-seed cells.
- **An account, read through one lens.** On thirteen public benchmarks (fourteen with TrafficC), four horizons and three seeds IsoKoop has the best mean rank of nine retrained models in MSE and MAE, at a fourteenth of EMAformer's parameters on 883 variates. The lens audits the benchmarks: train-to-test scale shift, Exchange won by persistence, and nine drifting Traffic sensors.

2 Related Work

Normalisation and the operators inside it. RevIN (Kim et al., 2022) standardises each window and inverts the map on the forecast. Dish-TS, SAN and FAN learn the statistics or remove the window's dominant frequencies (Fan et al., 2023; Liu et al., 2023b; Ye et al., 2024), Non-stationary Transformers re-inject the discarded scale into attention (Liu et al., 2022), and TimeBridge removes each patch's trend within a variate and keeps it across variates (Liu et al., 2025). These works choose the coordinates, and part (i) of Proposition 1 is the scale equivariance that motivates them. None states which components a model may add in them and keep equivariance; that is part (ii). The operators inside vary: iTransformer attends over variate tokens (Liu et al., 2024), PatchTST over patches (Nie et al., 2023), SOFTS aggregates variates through a central series (Han et al., 2024), STID adds variate and time identity embeddings (Shao et al., 2022), TimeMixer mixes scales (Wang et al., 2024), and Koopman models estimate a linear operator on a lifted state (Lusch et al., 2018; Liu et al., 2023a; Wang et al., 2023), recently with Mamba-style selective recurrences (Gu & Dao, 2024; Li et al., 2025). All of them see only the window and its timestamps. When the window does not hold a full cycle, neither the cycle's length nor its level is there to be read.

Cycle and frequency methods. CycleNet and FreqCycle subtract a phase-indexed cycle table after the normaliser (Lin et al., 2024; Zhang et al., 2026a), AsyTO factors one at low rank (Lin et al., 2026), EMAformer embeds calendar phase (Zhang et al., 2026b), FreDF trains in the frequency domain (Wang et al., 2025), DERITS forecasts frequency derivatives (Fan et al., 2024), FreEformer attends over spectral tokens (Yue et al., 2025), and FITS and FilterNet learn frequency filters (Xu et al., 2024; Yi et al., 2024). Objectives vary too: robust, shape- and transformation-aware losses (Barron, 2019; Le Guen & Thome, 2019; Lee et al., 2022). Each fixes its side of the normaliser by design, and none says what that side decides; we give the exact consequence and price both sides.

3 Method

3.1 Per-window normalisation is a coordinate system

A look-back window is $\mathbf{x} \in \mathbb{R}^{L\times N}$ (L steps, N variates) and its continuation $\mathbf{y} \in \mathbb{R}^{H\times N}$. A per-window normaliser with statistics $\boldsymbol{\mu}(\mathbf{x}), \mathbf{s}(\mathbf{x}) \in \mathbb{R}^N$,

$$\mathcal{N}_{\mathbf{x}}(\mathbf{u}) = \frac{\mathbf{u} - \boldsymbol{\mu}(\mathbf{x})}{\mathbf{s}(\mathbf{x})}, \qquad \mathcal{N}_{\mathbf{x}}^{-1}(\mathbf{v}) = \mathbf{s}(\mathbf{x})\,\mathbf{v} + \boldsymbol{\mu}(\mathbf{x}), \tag{1}$$

defines a chart for each window, and a forecaster of the form $f = \mathcal{N}_{\mathbf{x}}^{-1} \circ F \circ \mathcal{N}_{\mathbf{x}}$ acts in it. Such an f is *coordinate-native*. A component is *in coordinates* if it reads $\mathcal{N}_{\mathbf{x}}(\mathbf{x})$ or is rescaled by $\mathbf{s}(\mathbf{x})$.

Proposition 1 (Where out-of-window knowledge may enter)**.** *(i) If $\boldsymbol{\mu}, \mathbf{s}$ are positively homogeneous and $\mathbf{s}(\mathbf{x}) > 0$, every coordinate-native forecaster is scale-equivariant, $f(g\mathbf{x}) = g\,f(\mathbf{x})$ for all $g > 0$, whatever the inner network F. (ii) Let $\boldsymbol{\mu}$ be the per-variate mean and $\mathbf{s}$ the per-variate standard deviation, and let a profile $\mathbf{q}_h \in \mathbb{R}^{L\times N}$ that does not depend on the window's values be subtracted in raw units before $\mathcal{N}$ and $\mathbf{q}_f \in \mathbb{R}^{H\times N}$ added after $\mathcal{N}^{-1}$. Then f is scale-equivariant for every F if and only if $\mathbf{q}_h = \mathbf{q}_f = \mathbf{1}c^\top$ is a constant level.*

Corollary 1. *Each of the other three placements IsoKoop implements, switching the profile off, subtracting it inside the coordinates ($\mathbf{v} \mapsto F(\mathbf{v} - \tilde{\mathbf{q}}_h) + \tilde{\mathbf{q}}_f$) and rescaling it by $\mathbf{s}(\mathbf{x})$ before $\mathcal{N}$ (exactly while its scale-ratio clamp is inactive at $\mathbf{x}$ and $g\mathbf{x}$ and its horizon blend is off), is scale-equivariant for every F. So is a published cycle table subtracted after its normaliser.*

Proof sketch. $G = \mathcal{N}_{\mathbf{u}}^{-1} \circ F \circ \mathcal{N}_{\mathbf{u}}$ is positively homogeneous and commutes with a per-variate level. The two directions then follow from two constant networks (Appendix C).

The proposition assumes nothing about the data or the network. Removing a seasonal profile from the raw series and restoring it on the forecast, the additive seasonal index of the Holt–Winters family, forfeits the guarantee as soon as the profile has any shape. The corollary is the multiplicative counterpart.

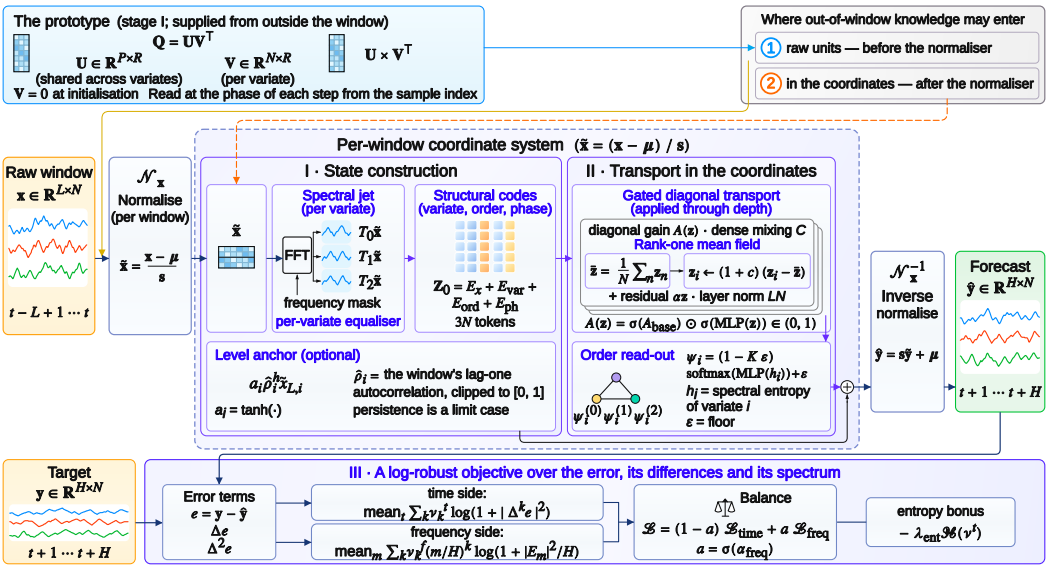


Figure 2: IsoKoop. Everything inside the dashed frame is the inner network F, acting in the coordinates of $\mathcal{N}_{\mathbf{x}}$; stage III is the objective, shown without its spectral-entropy matching term (λ_{spen}). The prototype enters at ①, in raw units (the model of Table 1), or at ②, in the coordinates; its rescaled and off placements, used only in Section 4.5, are not drawn here.

From the law to the machine. The law leaves three questions open, one per stage of Figure 2, and the answers form a chain. (1) A window shorter than the cycle does not fix its length (Design remark 5), and a phase within the cycle is defined only once the length is, so both come from outside. (2) Proposition 1 says where they may enter; this is stage I. (3) A spectral derivative errs by the periodic extension's jump (Statement 2), which removing the cycle shrinks, so the profile leaves before the derivative is formed. (4) What remains is advanced in the coordinates by a mean field, the only linear, permutation-equivariant coupling on zero row sums (Design remark 7); this is stage II. (5) Since the operator carries derivatives, the error is measured in the matching seminorm (Design remark 9); this is stage III. The statements behind (1) and (2) are exact, and the one behind (4) is exact among linear couplings; the implementation departs from them at the jet's span, the equaliser's premise and the objective's weights (Design remarks 6 and 9, Appendix I). The factorial prices what each stage buys (Section 4.4).

3.2 I · State construction: the phase, the level and the jet

The cycle the window cannot fix enters where Proposition 1 allows, its length from the protocol and its phase from the sample index, and only what remains is lifted.

A window shorter than a cycle contains only part of its profile and does not determine the cycle's period (Design remarks 4 and 5). IsoKoop fixes the period at P and supplies the profile as the **prototype**, a rank-R table $\mathbf{Q} = \mathbf{U}\mathbf{V}^\top$ over the P phases of the dataset cycle, with $\mathbf{U} \in \mathbb{R}^{P\times R}$ shared across variates and $\mathbf{V} \in \mathbb{R}^{N\times R}$ per variate ($\mathbf{V} = 0$ at initialisation, so the module is the identity at step 0), read at the phase of each step from the sample index. Its entry point is the *iso*-switch of Proposition 1: before $\mathcal{N}$ in raw units (①), after $\mathcal{N}$ in the coordinates (②), rescaled by $\mathbf{s}(\mathbf{x})$, or off. It is on for seven of the fourteen datasets (Appendix H), and the model of Table 1 places it in raw units. The **level anchor**, on for Exchange and Solar, $a_i\hat{\rho}_i^h\tilde{x}_{L,i}$ acts on the normalised window $\tilde{\mathbf{x}} = \mathcal{N}_{\mathbf{x}}(\mathbf{x})$, with $\hat{\rho}_i$ the window's lag-one least-squares coefficient clipped to $[0, 1]$ and $a_i = \tanh(\cdot)$ a learned gain. De-normalised it is an AR(1) forecast about the window mean, $\mu_i(\mathbf{x}) + a_i\hat{\rho}_i^h(x_{L,i} - \mu_i(\mathbf{x}))$ (Statement 1 under A1), and as $\hat{\rho}, a \to 1$ it tends to the last-value normalisation of NLinear: persistence is a limit case of the model.

Spectral jet. *The profile leaves before the derivative is formed.* The remainder is lifted to its derivatives $(\mathbf{x}, \dot{\mathbf{x}}, \ddot{\mathbf{x}})$, computed spectrally behind an equaliser,

$$(T_k\mathbf{x})_i = \mathrm{IFFT}\Big[(j\omega)^k\,\sigma\Big(\beta_k(\omega) + \mathbf{V}_i^{\mathrm{eq}\top} W_k\Big)\,\mathrm{FFT}(\mathbf{x}_i)\Big], \qquad k = 0, 1, 2, \tag{2}$$

with β_k a learned per-order frequency profile and $\mathbf{V}_i^{\mathrm{eq}}$ a per-variate fingerprint; the three channels are divided by the norm of channel 0. A spectral derivative is exact only on a window on which the signal is periodic. Otherwise its error is the jump of the periodic extension times a fixed profile that peaks at the window's edges, one where the forecast begins (Statement 2, on the split A2 that every even-length window admits). Removing the phase first shrinks that jump, so the derivative channels carry more of the signal and less of the window's boundary. In the coordinates the jet also sheds the level: for $k \geq 1$, T_k annihilates a constant, so its channels on $\tilde{\mathbf{x}}$ are the raw-unit ones divided by $s_i(\mathbf{x})$ (Appendix I). A banded variant, selected on PEMS03/04/07, Traffic and TrafficC, replaces $(j\omega)^k$ by a learned partition of unity into a low, a middle and a high band. Tokens embed whole series, timestamp features included, and **structural codes** mark each of the $3N$ tokens by variate, order and phase, $\mathbf{Z}_0 = E_x + E_{\mathrm{var}} + E_{\mathrm{ord}} + E_{\mathrm{ph}}$; the phase code is constant on PEMS and Solar, which carry no timestamps.

Under its hypothesis $(\mathbf{x}, \dot{\mathbf{x}})$ already spans the jet, and the third channel is a design choice (Design remark 6). The equaliser's statement makes a per-variate, per-order multiplier necessary and leaves the minimality of ours open (Statement 3); its power-law premise (A5) is measured in slope form (Appendix J.3) and the fingerprint outside the factorial (Appendix J.3). Under the equaliser, and in the five banded panels, the channels are reweighted spectra of the signal in place of its exact derivatives (Appendix I).

What stage I learns. Trained, the prototype is almost entirely a cycle on the four PEMS networks (0.87 to 1.00 of its RMS amplitude varies with phase, in 2.0 to 3.3 effective dimensions of R=8), mostly a per-sensor level on Traffic and TrafficC (0.35 to 0.44), and one-dimensional on Solar (1.14 to 1.16 of R=16; Figure 3b,c). Removing the prototype alone at H=96 (seed 2027, factorial runs) costs 3.7% on PEMS04, stays within 3% on PEMS03 and Traffic and lowers the error by 4.7% on Solar, and the structural codes carry more of its stage (21.3% and 11.9% on PEMS04 and PEMS03). Solar's anchor learns a gain of 0.054 at H=96, falling to 0.027 at H=720, and removing the anchor costs 4.9%. On Exchange the anchor runs its last-value mode (a_i=0.99, $\hat{\rho}_i$=1) and removing it costs 8.5%. In the sixteen-cell factorial (Table 2 prints fourteen), removing the whole stage (I) raises the test error at seed 2027 in twelve, sparing only Traffic, whose window holds a full cycle, and Exchange and PEMS03 at H=720.

3.3 II · Transport in the coordinates

That lifted state is advanced once in the coordinates: a bounded diagonal gain carries it through depth, and cross-variate information enters through one mean field.

Stage I hands over $\mathbf{z} = W_{\mathrm{in}}\mathbf{Z}_0$. Between $\mathcal{N}$ and $\mathcal{N}^{-1}$ the transport below, with the jet, the codes and the anchor, sees only $\tilde{\mathbf{x}}$ and calendar inputs (Figure 2), so it forms the inner network F of Proposition 1 for every parameter value; the prototype alone has a choice of side (Appendix I). Each layer applies

$$\mathbf{z} \;\leftarrow\; \mathrm{LN}(C\,(A(\mathbf{z}) \odot \mathbf{z}) + \alpha\mathbf{z}), \qquad A(\mathbf{z}) = \sigma(A_{\mathrm{base}}) \odot \sigma(\mathrm{MLP}(\mathbf{z})) \in (0, 1), \tag{3}$$

a selective diagonal gain followed by a dense mixing C: a recurrence through depth that is linear in $\mathbf{z}$ for a fixed gain, as in Koopman lifting. Cross-variate information enters as a rank-one mean field on $C(A(\mathbf{z}) \odot \mathbf{z})$ before LN, $\mathbf{z}_i \leftarrow (1 + c)(\mathbf{z}_i - \bar{\mathbf{z}})$. Among linear couplings on exchangeable tokens the restriction is forced: on zero row sums, centring is the whole of a permutation-equivariant one (Design remark 7). An **order read-out** $\psi_i = (1 - K\varepsilon)\,\mathrm{softmax}(\mathrm{MLP}(h_i)) + \varepsilon$, driven by the variate's spectral entropy h_i, recombines the K=3 orders. Its floor ε keeps a gradient path to every order at the price of a ceiling (Design remark 8). Trained, it leaves its step-0 weights in all 56 cells and moves weight off order 0 in 51. On Traffic and TrafficC at $H \leq 192$ order 1 overtakes order 0 (0.51 to 0.55 against 0.38 to 0.41; Figure 3a).

Removing the transport, the mean field and the read-out together (II) raises the test error in eleven of the sixteen cells of the factorial, by 62% to 315% in the six PEMS cells and by at most 26%

elsewhere. The mean field alone accounts for much of it: removing it at $H{=}96$ (seed 2027) costs 49.6% on PEMS04, 16.9% on PEMS03 and 6.3% on Electricity, and is a tie on ETTh1 and Solar. In a screen of an earlier configuration ($H{=}96$, seed 2027), attention, SOFTS-style aggregation or a general zero-sum coupling in its place cost 3% to 6% on Electricity and 13% to 22% on Weather, and none leads it by over 0.5% on the six panels screened. Attention of the same depth and width in place of the recurrence (seed 2027) collapses on Electricity to one validation error at both horizons, and those runs are left out of the counts (Table 7). Of the twelve cells where it trains it is ahead on Exchange and in the $H{=}720$ cells of Solar and Traffic (0.86 to 0.99) and raises the error in the other eight, by factors of 3.2 and 2.4 on PEMS04 ($H{=}96$ and 720) and 2.6 on PEMS03 at $H{=}96$.

3.4 III · A LOG-ROBUST OBJECTIVE OVER THE ERROR, ITS DIFFERENCES AND ITS SPECTRUM

Because the Koopman-style transport carries derivatives, its error is measured in the space it predicts in: the objective is modelled on one weighted H^2 seminorm of the error, written once per step and once per frequency.

For $e = \mathbf{y} - \hat{\mathbf{y}}$, forward differences Δ^k and DFT bins E_m,

$$\begin{aligned}\mathcal{L} = (1-a)\operatorname*{mean}_t \textstyle\sum_k \nu_k^{\mathrm{t}} \log\!\big(1 + |\Delta^k e|_t^2\big) \\ + a \operatorname*{mean}_m \textstyle\sum_k \nu_k^{\mathrm{f}}\, (m/H)^k \log\!\big(1 + |E_m|^2/H\big) - \lambda_{\mathrm{ent}} \mathcal{H}(\nu^{\mathrm{t}}) \\ + \lambda_{\mathrm{spen}} \operatorname*{mean}_i \big(\bar{h}_i(\hat{\mathbf{y}}) - \bar{h}_i(\mathbf{y})\big)^2,\end{aligned} \tag{4}$$

where $|\Delta^k e|_t^2$ is the squared k-th difference at step t averaged over variates, taken over the $H{-}2$ steps on which all three orders exist, and $\bar{h}_i$ is the normalised spectral entropy of variate i, averaged over the batch. The order weights $\nu^{\mathrm{t}} \geq \varepsilon$ come per step from a small gate network on the energies of the history's jet channels and the forecast's differences, ν^{f} is their mean over steps, and the entropy bonus keeps the gate off a single order. The gate and the balance $a = \sigma(\alpha_{\mathrm{freq}})$ (fixed on Weather) train with the model, so the loss can move weight toward its smaller terms. From its initial value of 0.54, the learned balance settles between 0.69 and 0.999, a per-dataset reading that barely moves with the horizon. Fixing it at 0.85 changes the error by -3.1% to $+0.8\%$ at $H{=}96$: a reports how much of the error the spectrum carries. Replacing the objective by MSE (III) raises the error in thirteen of the sixteen factorial cells, by 1.9% to 13.3%, and lowers it on Solar at both horizons and on Exchange at $H{=}720$. The kernel $\log(1 + e^2)$ is the Cauchy robust loss, applied to a frequency-domain term as well.

Why one seminorm, and why the log. By Parseval a weighted sum of squared circular differences of the error is one Sobolev seminorm in both domains (Design remark 9). Before the log, and with the remark's weights, the two halves of equation 4 are that one quantity written twice, weighting the error's velocity and acceleration, and the log makes each a robust surrogate of it. In the implementation the weights are to leading order $\nu^k \propto |e^{i\omega} - 1|^k$, with constant $(2\pi)^{-k}$, the square root of the exact symbol, and its time-side differences are truncated rather than circular (Appendix I). In the amplitude measure the two halves' gap grows with the error's scale, so a trainable balance would follow the scale rather than the data. In the log measure it converges to a finite limit, which makes a a dimensionless number the run can report, and the linear-amplitude variant diverges to a non-finite loss on six of seven panels (seed 2027, $H{=}96$, all but Exchange).

4 EXPERIMENTS

4.1 SETUP

Datasets and protocol. Fourteen datasets: ETTh1/h2/m1/m2, Weather, Solar, Electricity, Traffic, Exchange, the sensor networks PEMS03/04/07/08, and TrafficC (Appendix D), Traffic without the nine sensors whose level and scale shift after the training period. All models forecast every variate from a look-back of $L = 168$ at $H \in \{96, 192, 336, 720\}$ with seeds 2027/2028/2029 (Table 11). The eight comparison models of Table 1 keep their official repository configurations and are retrained in our repository with the same splits, three seeds and at most 20 epochs (input exceptions: Appendix H). IsoKoop's per-dataset settings are chosen on validation and ship with the code; every mechanism but the prototype and the anchor is on everywhere.

Table 1: Long-horizon forecasting on fourteen datasets: MSE and MAE averaged over $H \in \{96, 192, 336, 720\}$ and three seeds, look-back $L = 168$ (lower is better). Rows are ordered by P/L, the dataset cycle over the look-back. **Red**: best; **blue**: second best, at the printed precision (ties are all red). *1st Count*: datasets on which the model is best, all fourteen counted. Without TrafficC, IsoKoop leads ten of the thirteen public benchmarks in MSE and nine in MAE. TrafficC is Traffic without nine sensors identified by their train-to-test regime shift (statistics from the test period). Per-horizon results: Table 3; seed standard deviations: Table 4.

Dataset	P/L	**IsoKoop**		EMAformer		FreEformer		FreqCycle		SOFTS		FreDF		TimeMixer		iTransformer		PatchTST	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Exchange	0.03	0.389	0.415	**0.357**	**0.404**	**0.369**	0.414	0.381	0.415	**0.369**	**0.412**	0.439	0.460	0.393	0.417	0.395	0.428	0.397	0.425
ETTh1	0.14	**0.427**	**0.432**	0.436	0.437	0.445	0.448	0.436	**0.435**	0.443	0.444	**0.434**	0.438	0.462	0.448	0.455	0.452	0.475	0.463
ETTh2	0.14	**0.370**	**0.399**	0.375	0.402	0.378	0.405	**0.374**	**0.400**	**0.374**	0.404	0.376	0.403	0.380	0.411	0.386	0.411	0.388	0.417
ETTm1	0.57	**0.351**	**0.378**	0.364	0.387	0.374	0.394	**0.355**	**0.382**	0.375	0.393	0.367	0.387	0.432	0.418	0.374	0.396	0.357	0.385
ETTm2	0.57	**0.264**	**0.312**	0.276	0.324	0.272	0.321	**0.264**	**0.313**	0.280	0.328	0.271	0.318	**0.268**	0.319	0.284	0.333	0.281	0.333
Solar	0.86	**0.210**	0.262	0.216	0.260	**0.211**	0.256	0.226	0.266	0.214	**0.251**	**0.211**	**0.250**	0.244	0.265	0.217	0.257	**0.211**	0.265
Weather	0.86	**0.234**	**0.265**	0.237	**0.267**	**0.236**	0.268	**0.236**	0.268	0.251	0.278	0.248	0.274	**0.236**	0.268	0.252	0.278	0.245	0.272
Electricity	1.00	**0.152**	**0.246**	0.162	0.256	0.166	0.260	0.162	0.255	**0.161**	**0.253**	0.165	0.256	0.169	0.259	0.167	0.261	0.176	0.274
Traffic	1.00	0.425	0.274	0.427	0.288	0.424	0.278	0.437	0.286	**0.392**	**0.265**	**0.396**	**0.271**	0.435	0.277	0.397	0.274	0.424	0.286
TrafficC	1.00	**0.300**	**0.261**	0.322	0.275	**0.310**	0.265	0.322	0.273	**0.310**	**0.256**	0.313	0.262	0.323	0.267	0.314	0.265	0.334	0.277
PEMS03	1.71	0.196	0.281	**0.165**	**0.260**	**0.168**	**0.264**	0.210	0.305	0.196	0.290	0.200	0.295	0.215	0.308	0.234	0.328	0.250	0.346
PEMS04	1.71	**0.116**	**0.216**	0.147	0.243	**0.129**	**0.232**	0.216	0.316	0.192	0.287	0.217	0.309	0.186	0.293	0.214	0.310	0.303	0.387
PEMS07	1.71	**0.103**	**0.191**	0.111	**0.206**	**0.108**	0.208	0.201	0.301	0.138	0.237	0.320	0.353	0.158	0.263	0.292	0.381	0.272	0.360
PEMS08	1.71	**0.275**	**0.251**	0.277	**0.253**	**0.276**	**0.253**	0.387	0.350	0.290	0.272	0.306	0.305	0.368	0.341	0.468	0.439	0.424	0.392
1st Count		**11**	**9**	2	2	0	0	1	0	1	2	0	1	0	0	0	0	0	0

A diverged run (seed mean above 20 times the median over models in that cell) is left out of its average: TimeMixer on ETTm1 (1 horizon).

4.2 Main results

Across the (dataset, horizon) settings of the thirteen public benchmarks, IsoKoop has the best mean rank of the nine models: 2.34 against 3.77 for the runner-up EMAformer in MSE and 2.14 against 3.65 in MAE (ties averaged at the printed precision), and the lowest MSE in 34 of the 52. Adding the TrafficC panel gives mean ranks of 2.24 against 3.99 and 2.23 against 3.96; IsoKoop has the lowest MSE in 38 of the 56 settings, 22 of them by more than the larger of the two models' seed standard deviations, and the lowest MAE in 34. On the averages of all fourteen datasets it is first on eleven in MSE and nine in MAE, with a mean MSE rank of 1.86 against 3.79 for FreEformer.

Its four largest leads are on high-dimensional panels. On PEMS04 IsoKoop is 10.1% below the best comparison model (FreEformer), on Electricity 5.6% (SOFTS), on the 883-sensor PEMS07 4.6% (FreEformer) and on the 853-sensor TrafficC 3.2% (SOFTS and FreEformer, tied). On seven more, the four ETT panels, Solar, Weather and PEMS08, it is first in MSE by smaller margins; the other three go to EMAformer on Exchange, to SOFTS on Traffic, where nine drifting sensors decide the order (Section 4.6), and to EMAformer on PEMS03, whose level leaves its training range for 91.3% of its variates.

Cost. On PEMS07 IsoKoop uses 20.12M parameters against 278.97M for EMAformer and 6.52GB of peak memory against 7.21 to 24.25GB for the five attention models; only PatchTST is slower per sample (Appendix H.1).

4.3 Readings of the criterion

A window-only operator against the full model. With all three stages off, $\emptyset$ is per-window normalisation followed by a linear map of the window (Li et al., 2026). Its test error is 3.8 to 6.1 times the full model's at H=96 (1.8 to 3.8 at H=720) on the three panels whose cycle is 1.7 times the look-back, 0.98 to 1.4 times on the four whose cycle is at most as long (Traffic's in Appendix B), and 0.99 and 0.87 on Exchange, where persistence already wins (Table 2). The I×II interaction is positive at both horizons only where $P \geq L$ (Figure 5): there stage I and the transport substitute for one another.

Figure 3 reads the trained model along the same axis, with no test error used. The readings follow the ratio where the window lacks the cycle: at P/L=1.71 the prototype is a low-dimensional cycle and removing stage II costs most. At $P/L \leq 1$, where validation also turned the prototype on, what

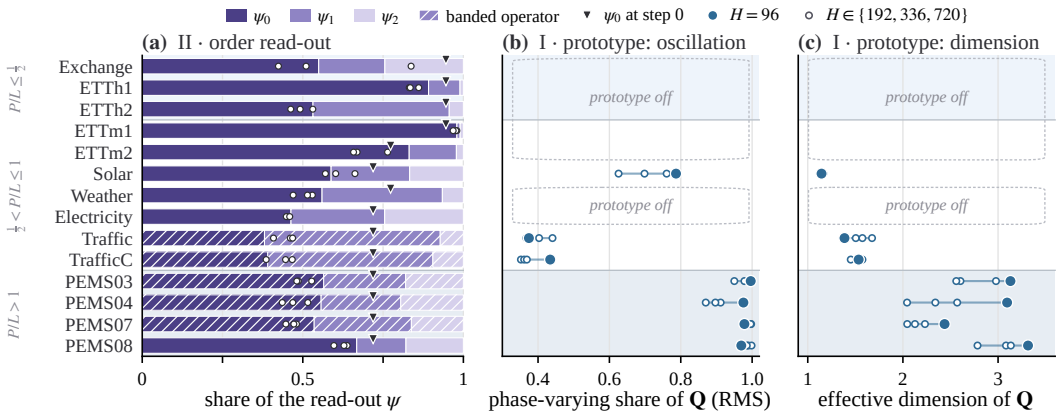


Figure 3: What the trained model reads. **(a)** Order read-out ψ (test mean), bars at H=96, ▼ at step 0, hatched: banded operator. **(b, c)** Prototype $\mathbf{Q}$: RMS share of its phase-varying part; effective dimension (rank R=8, 16 on Solar). Seed 2027, the runs of Table 1; read-outs only, no test error is used.

Table 2: Ablation: the 2^3 factorial over the three mechanisms (I state construction; II transport, mean field and read-out; III objective). Test MSE of every subset, seed 2027, at $H \in \{96, 720\}$, with all three pairwise interactions, $\mathrm{MSE}(A{+}B){-}\mathrm{MSE}(A){-}\mathrm{MSE}(B){+}\mathrm{MSE}(\emptyset)$ in % of the full model (negative = complementary, positive = substitutes). **Red**: best subset; shaded: the complete model. The eighth factorial dataset, Traffic, is omitted.

		mechanisms kept (test MSE)								interaction (%)		
Dataset	H	I II III	I II	I III	II III	I	II	III	$\emptyset$	I×II	I×III	II×III
ETTh1	96	**0.365**	0.372	**0.365**	0.375	0.372	0.391	0.379	0.380	−3.2	−1.7	−4.3
	720	0.466	0.477	0.462	0.518	**0.453**	0.509	0.467	0.455	−6.6	−0.7	−0.8
Solar	96	0.186	**0.177**	0.218	0.217	0.214	0.205	0.233	0.229	−7.1	−0.4	+4.0
	720	0.235	**0.218**	0.296	0.246	0.298	0.258	0.322	0.323	−6.3	−0.2	−4.0
Electricity	96	**0.123**	0.129	0.149	0.128	0.149	0.133	0.155	0.155	+1.7	−0.1	−4.3
	720	**0.185**	0.192	0.218	0.188	0.217	0.201	0.224	0.224	−1.0	+0.6	−7.2
Exchange	96	0.084	0.091	0.084	0.085	**0.083**	0.088	0.084	**0.083**	+3.4	−0.4	−5.3
	720	0.962	0.888	0.835	0.886	0.865	0.915	0.846	**0.834**	−6.0	−4.4	−4.4
PEMS03	96	**0.145**	0.154	0.418	0.173	0.370	0.178	0.581	0.554	+110.5	+14.5	−22.2
	720	0.245	0.254	0.398	**0.242**	0.376	0.267	0.461	0.440	+21.0	+0.7	−18.6
PEMS04	96	**0.100**	0.105	0.415	0.171	0.367	0.185	0.625	0.609	+162.0	+31.4	−30.5
	720	**0.132**	0.144	0.398	0.211	0.387	0.228	0.505	0.495	+18.1	+0.8	−20.8
PEMS08	96	**0.180**	0.187	0.476	0.222	0.480	0.228	0.707	0.695	+96.5	−9.4	−10.7
	720	**0.340**	0.348	0.569	0.379	0.571	0.393	0.691	0.683	+19.5	−3.3	−6.7

it learns differs across panels; the P/L=1.71 group is also the sensor networks and every run uses $L = 168$, so these readings follow ratio and panel type together. EMAformer's phase embedding added to iTransformer, with no new run, raises the error on Traffic and TrafficC and lowers it by at most 9.6% on the other panels below P/L=1.71, but by 29% to 62% at that ratio (Appendix B).

4.4 Ablation

A 2^3 factorial at one seed, whose reruns differ by a few percent (Appendix H), switches off I (the state construction), II (the transport, mean field and read-out) and III (the objective, replaced by MSE). Table 2 omits Traffic (Appendix B). Over its fourteen cells the complete model is the best subset in eight (strictly in seven, and tied with the drop-transport subset at ETTh1, H=96), and a subset is strictly better in the remaining six: ETTh1 at H=720, Solar and Exchange at both horizons, and PEMS03 at H=720. II and III are complementary in thirteen of the fourteen cells, Solar at H=96 excepted, and in 14 of all 16.

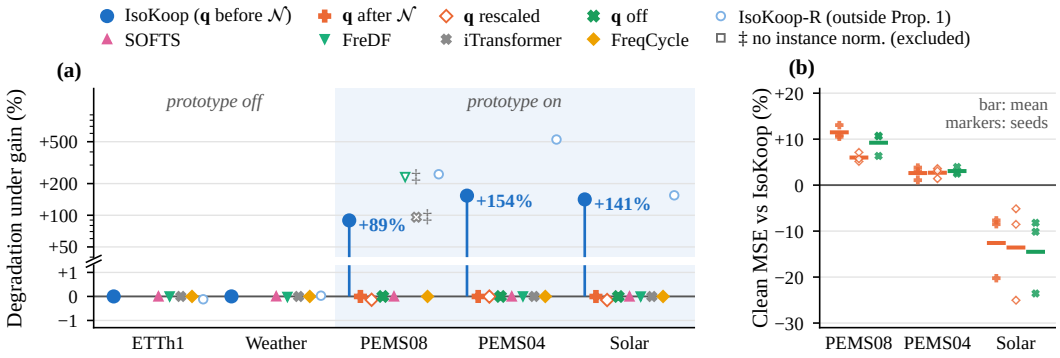


Figure 4: Proposition 1 on trained networks. **(a)** Change of test MSE when the whole window is multiplied by $g \in \{0.5, 2\}$ (H=96, three seeds, broken axis). Every normalising comparison model stays on the 0.0% line; IsoKoop with its prototype in raw units leaves it, and each in-coordinate placement of Corollary 1 returns it. ‡Run without instance normalisation, as in those models' papers. IsoKoop-R: Section 4.6. **(b)** Clean-MSE change of each placement against the raw-unit design (+: worse), per seed (markers) and seed mean (bar).

4.5 Proposition 1 on trained networks

The gain test multiplies window and target by $g \in \{0.5, 2\}$ and divides the forecast by g before scoring, so an exactly equivariant model scores 0.0% (Figure 4a, Table 8). The four comparison models tested score 0.0% wherever they normalise, as part (i) predicts. IsoKoop with its prototype in raw units degrades by +153.6%, +89.4% and +141.5% on PEMS04, PEMS08 and Solar, as part (ii) permits (Solar: Appendix C), and by 0.0% where the prototype is off. Moving FreqCycle's cycle table in front of its normaliser, nothing else changed, degrades that published model by +144.3%, +131.9% and +210.6% on the same three panels, and rescaling the moved table by the window's scale brings it back within 0.8%. IsoKoop's in-coordinate placements return to the line, the rescaled one within 0.1% (Figure 6).

At H=96 the in-coordinate placements (prototype off included) do not lower the error under real shift (Table 9), cost 2.6% to 3.1% on PEMS04 and 6.0% to 11.5% on PEMS08, and gain 12.6% to 14.5% on Solar against a reference with one outlier seed (Figure 4b).

4.6 The benchmarks through the coordinate lens

The test period moves the scale. Between 14.3% and 83.9% of a dataset's variates have a test-period standard deviation outside the range it takes over training windows of equal length (36.5% to 83.9% where the prototype is on, Figure 8a): the shift normalisation is built for. Noise-robust IsoKoop-R wins 100 of 200 corrupted cells at H=96 against four models (Appendix F).

Nine sensors make the Traffic ranking. Nine of Traffic's 862 sensors change level and scale after the training period (Figure 7) and carry 22% to 29% of the H=96 error of every model whose per-sensor error we recorded (Figure 8b). Removing them and retraining all nine models (TrafficC) lowers every model's error by 20.9% to 29.4% and reorders the table: IsoKoop from sixth to first, 0.300 against 0.310 for SOFTS and FreEformer, PatchTST from a tie for fourth to ninth (Table 10).

Exchange: persistence wins. The last-value forecast (Zeng et al., 2023) is below all nine models at every horizon (Figure 8c, Table 5), though IsoKoop's anchor tends to it.

5 Conclusion

Proposition 1 fixes where out-of-window knowledge may enter, and the ratio of the cycle to the look-back says when it is needed. The guarantee belongs to the placement, not to a model: FreqCycle loses it when its table moves in front of the normaliser, and IsoKoop, reported without it, has the best mean rank of nine models on thirteen public benchmarks. Report which side of the normaliser such knowledge enters.

AI USE STATEMENT

This work was carried out with the assistance of generative AI tools, which supported us in the following tasks with required disclosure: helping to develop the theoretical framework and to formulate its mathematical claims, supplying ingredients for the proofs, assisting in the writing of proofs, proposing hypotheses, designing experiments, and implementing the method and the analysis code. We did not use them to generate synthetic data sets, to clean or reformat datasets, for translation, for qualitative analysis, or to interpret results.

Among the tasks with recommended disclosure, we used these tools to assist with writing and with the preparation of figures and tables.

We have reviewed all AI-assisted work. The authors set the research direction; the statements and proofs in the appendix were checked against their assumptions; the model, the comparison-model ports and the analysis code were run by the authors, and every table and figure is generated from our own run outputs; and every numerical claim was checked against the table that prints it. We take responsibility for the final content of this work, including its text, claims and artifacts.

ETHICS STATEMENT

This work studies forecasting on public benchmark datasets of physical and infrastructure measurements and introduces no human-subject data; no ethical risk specific to it is apparent. One finding bears on deployment: a small number of sensors whose behaviour changes after the training period can dominate a network's aggregate error (Section 4.6), so an aggregate metric alone can hide the sensors that changed.

REPRODUCIBILITY STATEMENT

The model, the eight comparison models and the data pipeline live in one repository, so the pipeline is identical across models by construction. Section 3 specifies the architecture and Appendix H the datasets, splits, seeds and protocol; the proof of Proposition 1 is in Appendix C. The model, the comparison models, the data pipeline and the script that builds TrafficC are provided as anonymous supplementary material, together with the figures and tables that appear in this paper. Every configuration behind our results is stated in that material: `hyperparameters_md` gives the final per-dataset setting of each of the nine models, and the released training scripts run those settings as reported.

REFERENCES

Jonathan T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4326–4334, 2019. doi: 10.1109/CVPR.2019.00446.

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. doi: 10.1093/biomet/81.3.425.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-TS: A general paradigm for alleviating distribution shift in time series forecasting. In *AAAI Conference on Artificial Intelligence*, pp. 7522–7529, 2023. doi: 10.1609/aaai.v37i6.25914.

Wei Fan, Kun Yi, Hangting Ye, Zhiyuan Ning, Qi Zhang, and Ning An. Deep frequency derivative learning for non-stationary time series forecasting. In *International Joint Conference on Artificial Intelligence*, pp. 3944–3952, 2024. doi: 10.24963/ijcai.2024/436.

N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16(1):109–114, 1965. doi: 10.1090/s0002-9939-1965-0174934-9.

Ulf Grenander and Gábor Szegő. *Toeplitz Forms and Their Applications*. University of California Press, 1958.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. SOFTS: Efficient multivariate time series forecasting with series-core fusion. In *Advances in Neural Information Processing Systems*, pp. 64145–64175, 2024.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3rd edition, 2021. URL https://otexts.com/fpp3/.

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

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. doi: 10.1214/aoms/1177698505.

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*, pp. 4189–4201, 2019.

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.

Zhaoyang Li, Minghao Han, and Xunyuan Yin. MamKO: Mamba-based Koopman operator for modeling and predictive control. In *International Conference on Learning Representations*, 2025.

Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. Revisiting long-term time series forecasting: An investigation on affine mapping. *Academia AI and Applications*, 2(2), 2026. doi: 10.20935/AcadAI8236.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. CycleNet: Enhancing time series forecasting through modeling periodic patterns. In *Advances in Neural Information Processing Systems*, pp. 106315–106345, 2024.

Xiachong Lin, Du Yin, Hao Xue, Wen Hu, Imran Razzak, Arian Prabowo, Matthew Amos, and Flora D. Salim. AsyTO: Asymmetric temporal operator for parameter-efficient multivariate time series forecasting. *arXiv preprint arXiv:2608.16098*, 2026.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. TimeBridge: Non-stationarity matters for long-term time series forecasting. In *International Conference on Machine Learning*, pp. 39815–39840, 2025.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, pp. 9881–9893, 2022.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with Koopman predictors. In *Advances in Neural Information Processing Systems*, pp. 12271–12290, 2023a.

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

Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. In *Advances in Neural Information Processing Systems*, pp. 14273–14292, 2023b.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, 2018.

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.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *ACM International Conference on Information and Knowledge Management*, pp. 4454–4458, 2022. doi: 10.1145/3511808.3557702.

Lloyd N. Trefethen. *Spectral Methods in MATLAB*. Society for Industrial and Applied Mathematics, 2000.

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*, 2025.

Rui Wang, Yihe Dong, Sercan Ö. Arık, and Rose Yu. Koopman neural operator forecaster for time-series with temporal distributional shifts. In *International Conference on Learning Representations*, 2023.

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.

Peter R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960. doi: 10.1287/mnsc.6.3.324.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, 2024.

Weiwei Ye, Songgaojun Deng, Qiaosha Zou, and Ning Gui. Frequency adaptive normalization for non-stationary time series forecasting. In *Advances in Neural Information Processing Systems*, pp. 31350–31379, 2024.

Kun Yi, Jingru Fei, Qi Zhang, Hui He, Shufeng Hao, Defu Lian, and Wei Fan. FilterNet: Harnessing frequency filters for time series forecasting. In *Advances in Neural Information Processing Systems*, pp. 55115–55140, 2024.

Wenzhen Yue, Yong Liu, Xianghua Ying, Bowei Xing, Ruohao Guo, and Ji Shi. FreEformer: Frequency enhanced transformer for multivariate time series forecasting. In *International Joint Conference on Artificial Intelligence*, pp. 3606–3614, 2025.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, pp. 11121–11128, 2023.

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

Zhiwei Zhang, Xinyi Du, Xuanchi Guo, Weihao Wang, and Wenjuan Han. EMAformer: Enhancing transformer through embedding armor for time series forecasting. In *AAAI Conference on Artificial Intelligence*, pp. 28636–28644, 2026b.

A FULL RESULTS

Table 3 gives every (dataset, horizon) setting of Table 1 in MSE and MAE, Table 4 the seed standard deviation behind each of those cells, and Table 5 Exchange against the persistence forecast. One run diverged: TimeMixer on ETTm1 at $H = 336$, whose seed-2028 run ends at test MSE 444.97 (seeds 2027 and 2029: 0.3633 and 0.3657). A cell whose seed-mean MSE exceeds 20 times the median over models is reported as measured and, in both metrics, left out of the averages and ranked last in its own cell.

Table 3: Every cell of Table 1: test MSE and MAE, mean over seeds 2027/2028/2029, and each dataset's average over the four horizons (*Avg*). **Red**: best; **blue**: second best, per metric at the printed precision. *1st Count* counts the 56 (dataset, horizon) cells, not the averages. Seed standard deviations: Table 4. ‡: a diverged run (seed mean above 20 times the median over models in that cell), reported as measured, left out of the average and ranked last in its own cell.

Dataset	H	**IsoKoop**		EMAformer		FreEformer		FreqCycle		SOFTS		FreDF		TimeMixer		iTransformer		PatchTST	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Exchange	96	0.087	**0.204**	**0.084**	**0.206**	0.089	0.212	**0.086**	**0.206**	0.092	0.213	0.124	0.253	0.090	0.211	0.102	0.225	0.102	0.222
	192	0.193	0.308	**0.175**	**0.299**	0.187	0.312	0.191	0.312	**0.184**	0.309	0.236	0.354	**0.175**	**0.301**	0.204	0.325	0.195	0.316
	336	0.376	0.440	**0.336**	**0.422**	0.362	0.441	**0.356**	0.432	**0.336**	**0.424**	0.409	0.472	0.373	0.444	0.362	0.440	0.366	0.440
	720	0.901	0.709	**0.833**	**0.687**	**0.837**	**0.690**	0.890	0.710	0.863	0.702	0.987	0.760	0.934	0.713	0.911	0.724	0.923	0.720
	Avg	0.389	0.415	**0.357**	**0.404**	**0.369**	0.414	0.381	0.415	**0.369**	**0.412**	0.439	0.460	0.393	0.417	0.395	0.428	0.397	0.425
ETTh1	96	**0.364**	**0.391**	0.381	0.403	0.383	0.408	0.379	**0.397**	0.386	0.406	**0.377**	0.401	0.390	0.404	0.389	0.408	0.378	0.402
	192	**0.424**	**0.427**	0.434	0.431	0.439	0.437	**0.428**	**0.430**	0.439	0.437	0.431	0.433	0.438	**0.430**	0.444	0.441	0.433	0.437
	336	**0.453**	**0.441**	0.464	0.446	0.472	0.457	**0.462**	**0.445**	0.480	0.461	0.463	0.448	0.470	0.446	0.481	0.460	0.490	0.471
	720	**0.466**	**0.467**	**0.466**	0.469	0.484	0.490	0.473	**0.468**	0.468	0.475	**0.464**	0.471	0.551	0.511	0.505	0.498	0.598	0.543
	Avg	**0.427**	**0.432**	0.436	0.437	0.445	0.448	0.436	**0.435**	0.443	0.444	**0.434**	0.438	0.462	0.448	0.455	0.452	0.475	0.463
ETTh2	96	**0.290**	**0.339**	0.296	0.347	0.301	0.351	0.296	0.343	0.297	0.351	**0.290**	**0.341**	**0.292**	0.348	0.302	0.354	0.307	0.357
	192	**0.363**	**0.391**	0.365	**0.393**	0.374	0.398	**0.364**	**0.393**	0.368	0.396	0.371	0.395	0.365	0.398	0.378	0.401	0.378	0.407
	336	**0.400**	**0.422**	0.408	0.424	**0.404**	0.424	0.405	**0.423**	0.405	0.425	0.416	0.429	0.418	0.437	0.424	0.433	0.419	0.440
	720	**0.428**	**0.443**	0.432	0.444	0.431	0.448	0.430	**0.441**	**0.427**	0.446	**0.427**	0.445	0.446	0.461	0.441	0.454	0.447	0.466
	Avg	**0.370**	**0.399**	0.375	0.402	0.378	0.405	**0.374**	**0.400**	**0.374**	0.404	0.376	0.403	0.380	0.411	0.386	0.411	0.388	0.417
ETTm1	96	**0.290**	**0.336**	0.295	0.347	0.305	0.353	**0.292**	**0.342**	0.302	0.351	0.302	0.349	0.307	0.352	0.303	0.353	0.298	0.348
	192	**0.332**	**0.366**	0.339	0.373	0.343	0.376	0.336	**0.365**	0.356	0.380	0.345	0.375	0.342	0.373	0.350	0.380	**0.334**	0.370
	336	**0.362**	**0.386**	0.374	0.395	0.398	0.408	**0.365**	**0.392**	0.392	0.404	0.379	0.395	148.6‡	4.884‡	0.388	0.405	0.368	**0.392**
	720	**0.421**	**0.424**	0.448	0.434	0.450	0.440	**0.425**	**0.428**	0.451	0.439	0.441	0.431	0.647	0.527	0.454	0.443	0.429	**0.428**
	Avg	**0.351**	**0.378**	0.364	0.387	0.374	0.394	**0.355**	**0.382**	0.375	0.393	0.367	0.387	0.432	0.418	0.374	0.396	0.357	0.385
ETTm2	96	**0.167**	**0.248**	0.175	0.259	0.176	0.259	**0.168**	**0.250**	0.173	0.259	0.172	0.256	0.171	0.255	0.183	0.270	0.177	0.264
	192	**0.229**	**0.289**	0.239	0.302	0.238	0.300	**0.227**	**0.287**	0.245	0.308	0.236	0.296	0.234	0.297	0.254	0.316	0.243	0.309
	336	**0.283**	**0.326**	0.301	0.341	0.291	0.335	**0.285**	**0.328**	0.307	0.345	0.295	0.334	0.286	0.332	0.310	0.350	0.297	0.344
	720	**0.377**	**0.386**	0.390	0.395	0.384	0.392	**0.378**	**0.386**	0.397	0.400	0.380	**0.387**	0.382	0.393	0.390	0.398	0.406	0.413
	Avg	**0.264**	**0.312**	0.276	0.324	0.272	0.321	**0.264**	**0.313**	0.280	0.328	0.271	0.318	**0.268**	0.319	0.284	0.333	0.281	0.333
Solar	96	0.202	0.240	**0.184**	0.233	0.194	**0.225**	0.188	0.241	0.195	0.228	**0.183**	**0.219**	0.211	0.243	0.190	0.230	0.188	0.252
	192	**0.203**	**0.257**	0.213	**0.257**	0.207	0.258	0.220	0.263	0.216	**0.251**	0.211	**0.251**	0.241	0.264	0.223	0.261	**0.206**	0.262
	336	**0.216**	0.270	0.229	0.272	**0.218**	**0.264**	0.244	0.278	0.221	**0.262**	0.221	**0.264**	0.263	0.278	0.228	0.268	0.222	0.274
	720	**0.219**	0.280	0.239	0.279	0.226	0.276	0.251	0.284	**0.224**	**0.264**	0.230	**0.264**	0.259	0.275	0.228	**0.269**	0.226	0.274
	Avg	**0.210**	0.262	0.216	0.260	**0.211**	0.256	0.226	0.266	0.214	**0.251**	**0.211**	**0.250**	0.244	0.265	0.217	0.257	**0.211**	0.265
Weather	96	**0.148**	**0.192**	0.154	**0.200**	0.155	0.202	**0.153**	0.202	0.170	0.213	0.168	0.209	0.156	0.205	0.172	0.215	0.165	0.208
	192	**0.196**	**0.241**	0.201	0.245	0.199	0.244	**0.197**	**0.243**	0.219	0.259	0.214	0.252	0.198	**0.243**	0.218	0.256	0.209	0.248
	336	**0.255**	**0.285**	0.257	**0.286**	**0.256**	**0.286**	0.257	0.287	0.272	0.296	0.268	0.293	**0.255**	**0.286**	0.271	0.295	0.264	0.290
	720	0.337	0.342	0.335	**0.339**	**0.334**	**0.339**	0.335	0.341	0.344	0.345	0.343	0.343	**0.333**	**0.339**	0.345	0.346	0.340	**0.340**
	Avg	**0.234**	**0.265**	0.237	**0.267**	**0.236**	0.268	**0.236**	0.268	0.251	0.278	0.248	0.274	**0.236**	0.268	0.252	0.278	0.245	0.272
Electricity	96	**0.123**	**0.216**	**0.126**	**0.221**	0.131	0.227	0.130	0.225	0.129	0.222	0.130	0.223	0.135	0.227	0.134	0.230	0.144	0.246
	192	**0.143**	**0.235**	**0.148**	**0.241**	0.155	0.249	**0.148**	**0.241**	0.151	0.242	0.149	**0.241**	0.155	0.245	0.154	0.248	0.162	0.262
	336	**0.159**	**0.253**	**0.163**	**0.257**	0.169	0.266	0.165	0.259	0.166	0.259	0.165	0.259	0.173	0.263	0.170	0.266	0.180	0.278
	720	**0.184**	**0.280**	0.212	0.304	0.210	0.300	0.203	0.294	**0.199**	**0.289**	0.214	0.302	0.211	0.299	0.211	0.301	0.218	0.310
	Avg	**0.152**	**0.246**	0.162	0.256	0.166	0.260	0.162	0.255	**0.161**	**0.253**	0.165	0.256	0.169	0.259	0.167	0.261	0.176	0.274
Traffic	96	0.373	**0.247**	0.389	0.266	0.374	**0.255**	0.410	0.273	**0.353**	**0.247**	**0.363**	**0.255**	0.396	0.264	0.366	0.259	0.392	0.271
	192	0.407	**0.265**	0.416	0.281	0.407	0.273	0.428	0.281	**0.382**	**0.260**	**0.388**	0.266	0.422	0.274	0.389	0.270	0.414	0.280
	336	0.430	0.277	0.431	0.290	0.426	0.281	0.441	0.287	**0.400**	**0.268**	**0.399**	**0.272**	0.438	0.276	**0.399**	0.275	0.428	0.287
	720	0.491	0.307	0.472	0.314	0.489	0.301	0.467	0.303	**0.431**	**0.286**	**0.435**	**0.291**	0.484	0.294	**0.435**	0.293	0.460	0.305
	Avg	0.425	0.274	0.427	0.288	0.424	0.278	0.437	0.286	**0.392**	**0.265**	**0.396**	**0.271**	0.435	0.277	0.397	0.274	0.424	0.286
TrafficC	96	**0.271**	**0.237**	0.292	0.253	0.279	0.246	0.302	0.261	**0.278**	**0.239**	0.288	0.247	0.300	0.255	0.291	0.252	0.310	0.263
	192	**0.291**	**0.253**	0.318	0.270	**0.304**	0.260	0.318	0.269	**0.304**	**0.252**	0.308	0.258	0.318	0.263	0.310	0.262	0.327	0.272
	336	**0.305**	0.265	0.325	0.279	0.317	0.268	0.325	0.275	0.317	**0.259**	**0.315**	**0.262**	0.327	0.266	0.316	0.265	0.338	0.278
	720	**0.333**	0.291	0.353	0.299	0.340	0.287	0.341	0.289	**0.339**	**0.276**	0.341	**0.281**	0.348	0.284	0.340	0.282	0.360	0.294
	Avg	**0.300**	**0.261**	0.322	0.275	**0.310**	0.265	0.322	0.273	**0.310**	**0.256**	0.313	0.262	0.323	0.267	0.314	0.265	0.334	0.277
PEMS03	96	0.150	0.258	**0.124**	**0.234**	**0.133**	**0.245**	0.183	0.293	0.162	0.275	0.178	0.289	0.184	0.293	0.229	0.338	0.242	0.351
	192	0.182	0.269	**0.146**	**0.247**	**0.146**	**0.249**	0.195	0.294	**0.177**	0.277	0.178	0.279	0.195	0.292	0.222	0.323	0.230	0.329

Table 3 (continued)

Dataset	H	**IsoKoop**		EMAformer		FreEformer		FreqCycle		SOFTS		FreDF		TimeMixer		iTransformer		PatchTST	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
	336	0.206	0.286	**0.174**	**0.264**	**0.171**	**0.266**	0.206	0.298	0.196	0.285	0.196	0.287	0.214	0.304	0.233	0.322	0.240	0.334
	720	0.245	0.313	**0.217**	**0.295**	**0.220**	**0.297**	0.256	0.336	0.249	0.321	0.246	0.325	0.266	0.345	0.251	0.330	0.287	0.369
	Avg	0.196	0.281	**0.165**	**0.260**	**0.168**	**0.264**	0.210	0.305	0.196	0.290	0.200	0.295	0.215	0.308	0.234	0.328	0.250	0.346
PEMS04	96	**0.100**	**0.205**	**0.107**	**0.212**	0.112	0.219	0.199	0.306	0.178	0.284	0.208	0.310	0.151	0.268	0.202	0.309	0.313	0.403
	192	**0.108**	**0.208**	0.136	0.231	**0.127**	**0.227**	0.209	0.308	0.190	0.281	0.197	0.292	0.182	0.284	0.199	0.298	0.282	0.371
	336	**0.125**	**0.221**	0.160	0.250	**0.136**	**0.236**	0.213	0.310	0.190	0.280	0.206	0.295	0.196	0.299	0.214	0.302	0.280	0.367
	720	**0.131**	**0.230**	0.186	0.277	**0.142**	**0.246**	0.244	0.341	0.211	0.303	0.255	0.339	0.217	0.320	0.242	0.330	0.335	0.407
	Avg	**0.116**	**0.216**	0.147	0.243	**0.129**	**0.232**	0.216	0.316	0.192	0.287	0.217	0.309	0.186	0.293	0.214	0.310	0.303	0.387
PEMS07	96	**0.091**	**0.185**	**0.093**	**0.193**	0.109	0.214	0.190	0.293	0.136	0.241	0.793	0.642	0.146	0.257	0.240	0.346	0.292	0.380
	192	**0.094**	**0.185**	**0.097**	**0.193**	**0.097**	0.196	0.192	0.293	0.125	0.225	0.146	0.244	0.159	0.261	0.340	0.413	0.254	0.346
	336	**0.101**	**0.189**	0.110	0.204	**0.102**	**0.199**	0.192	0.293	0.129	0.226	0.148	0.244	0.145	0.249	0.278	0.372	0.245	0.339
	720	**0.124**	**0.207**	0.144	0.234	**0.125**	**0.221**	0.230	0.325	0.160	0.254	0.191	0.282	0.181	0.283	0.311	0.394	0.296	0.378
	Avg	**0.103**	**0.191**	0.111	**0.206**	**0.108**	0.208	0.201	0.301	0.138	0.237	0.320	0.353	0.158	0.263	0.292	0.381	0.272	0.360
PEMS08	96	**0.179**	**0.236**	**0.180**	**0.237**	0.183	0.238	0.295	0.332	0.195	0.260	0.250	0.274	0.277	0.332	0.404	0.400	0.359	0.390
	192	**0.256**	**0.239**	0.258	**0.242**	0.258	0.244	0.372	0.342	0.259	0.258	**0.257**	0.274	0.350	0.333	0.506	0.471	0.400	0.374
	336	0.320	0.256	**0.313**	**0.251**	0.318	**0.252**	0.420	0.349	0.325	0.267	**0.297**	0.291	0.409	0.340	0.485	0.443	0.453	0.386
	720	**0.345**	**0.275**	0.357	0.282	**0.347**	**0.278**	0.459	0.377	0.380	0.302	0.421	0.383	0.434	0.361	0.476	0.441	0.484	0.419
	Avg	**0.275**	**0.251**	0.277	**0.253**	**0.276**	**0.253**	0.387	0.350	0.290	0.272	0.306	0.305	0.368	0.341	0.468	0.439	0.424	0.392
1st Count		**38**	**34**	7	9	2	1	1	4	5	10	6	3	3	1	1	0	0	0

Table 4: Sample standard deviation ($n-1$ denominator) over seeds 2027/2028/2029 of the test MSE and MAE whose means are Table 3, same column order. **0.0000**: a standard deviation below 5×10^{-5}, not identical seeds. ‡: the diverged cells of that table, reported as measured; their spread is why the seed mean there is not a score.

Dataset	H	**IsoKoop**		EMAformer		FreEformer		FreqCycle		SOFTS		FreDF		TimeMixer		iTransformer		PatchTST	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Exchange	96	0.0024	0.0037	0.0013	0.0025	0.0007	0.0011	0.0003	0.0005	0.0036	0.0055	0.0023	0.0028	0.0003	0.0003	0.0014	0.0016	0.0101	0.0084
	192	0.0122	0.0091	0.0027	0.0021	0.0047	0.0041	0.0032	0.0030	0.0063	0.0048	0.0019	0.0017	0.0017	0.0019	0.0015	0.0013	0.0083	0.0064
	336	0.0286	0.0182	0.0048	0.0023	0.0222	0.0159	0.0044	0.0033	0.0095	0.0077	0.0024	0.0019	0.0123	0.0094	0.0019	0.0010	0.0296	0.0163
	720	0.0459	0.0154	0.0096	0.0042	0.0374	0.0196	0.0040	0.0017	0.0233	0.0099	0.0026	0.0011	0.0307	0.0131	0.0031	0.0016	0.0379	0.0143
ETTh1	96	0.0006	0.0009	0.0045	0.0030	0.0049	0.0023	0.0037	0.0004	0.0022	0.0009	0.0025	0.0011	0.0068	0.0032	0.0015	0.0020	0.0047	0.0018
	192	0.0037	0.0029	0.0039	0.0023	0.0012	0.0003	0.0075	0.0046	0.0057	0.0053	0.0002	0.0012	0.0077	0.0062	0.0060	0.0046	0.0037	0.0021
	336	0.0027	0.0017	0.0033	0.0018	0.0043	0.0035	0.0041	0.0007	0.0079	0.0062	0.0086	0.0065	0.0173	0.0109	0.0050	0.0038	0.0131	0.0081
	720	0.0085	0.0034	0.0082	0.0035	0.0048	0.0018	0.0034	0.0026	0.0082	0.0051	0.0092	0.0072	0.0488	0.0229	0.0089	0.0047	0.0210	0.0080
ETTh2	96	0.0031	0.0016	0.0031	0.0026	0.0029	0.0027	0.0014	0.0015	0.0037	0.0033	0.0012	0.0009	0.0049	0.0044	0.0005	0.0004	0.0059	0.0029
	192	0.0048	0.0034	0.0026	0.0017	0.0026	0.0020	0.0030	0.0027	0.0010	0.0004	0.0053	0.0033	0.0068	0.0093	0.0007	0.0008	0.0090	0.0036
	336	0.0036	0.0028	0.0017	0.0015	0.0026	0.0004	0.0031	0.0004	0.0046	0.0017	0.0049	0.0024	0.0089	0.0073	0.0128	0.0061	0.0028	0.0062
	720	0.0024	0.0022	0.0031	0.0024	0.0044	0.0008	0.0012	0.0002	0.0075	0.0039	0.0037	0.0023	0.0153	0.0132	0.0133	0.0068	0.0088	0.0080
ETTm1	96	0.0002	0.0004	0.0007	0.0001	0.0027	0.0015	0.0010	0.0015	0.0036	0.0023	0.0003	0.0006	0.0187	0.0094	0.0034	0.0020	0.0059	0.0045
	192	0.0003	0.0005	0.0029	0.0018	0.0040	0.0026	0.0008	0.0005	0.0074	0.0027	0.0011	0.0014	0.0050	0.0048	0.0012	0.0012	0.0019	0.0007
	336	0.0006	0.0002	0.0016	0.0004	0.0112	0.0038	0.0007	0.0004	0.0020	0.0010	0.0003	0.0006	256.69‡	7.7809‡	0.0020	0.0013	0.0008	0.0015
	720	0.0011	0.0005	0.0014	0.0002	0.0081	0.0050	0.0010	0.0010	0.0044	0.0020	0.0014	0.0008	0.3429	0.1508	0.0030	0.0019	0.0027	0.0019
ETTm2	96	0.0016	0.0009	0.0024	0.0015	0.0034	0.0019	0.0006	0.0004	0.0015	0.0011	0.0027	0.0012	0.0017	0.0015	0.0025	0.0008	0.0020	0.0023
	192	0.0035	0.0027	0.0027	0.0033	0.0008	0.0011	0.0005	0.0006	0.0018	0.0012	0.0016	0.0007	0.0059	0.0050	0.0040	0.0018	0.0037	0.0039
	336	0.0015	0.0005	0.0017	0.0007	0.0002	0.0006	0.0017	0.0008	0.0017	0.0017	0.0051	0.0028	0.0068	0.0039	0.0065	0.0033	0.0030	0.0016
	720	0.0027	0.0019	0.0025	0.0018	0.0020	0.0008	0.0024	0.0022	0.0079	0.0045	0.0031	0.0007	0.0034	0.0043	0.0030	0.0016	0.0058	0.0031
Solar	96	0.0087	0.0048	0.0013	0.0012	0.0108	0.0033	0.0009	0.0013	0.0017	0.0011	0.0017	0.0013	0.0045	0.0048	0.0027	0.0003	0.0055	0.0079
	192	0.0039	0.0029	0.0023	0.0005	0.0040	0.0028	0.0013	0.0019	0.0036	0.0011	0.0034	0.0021	0.0088	0.0115	0.0041	0.0032	0.0014	0.0030
	336	0.0121	0.0077	0.0020	0.0013	0.0038	0.0010	0.0009	0.0006	0.0032	0.0003	0.0021	0.0011	0.0035	0.0072	0.0072	0.0035	0.0028	0.0021
	720	0.0129	0.0053	0.0023	0.0004	0.0018	0.0025	0.0015	0.0001	0.0010	0.0007	0.0007	0.0003	0.0016	0.0090	0.0010	0.0001	0.0015	0.0024
Weather	96	0.0009	0.0011	0.0015	0.0012	0.0016	0.0021	0.0004	0.0003	0.0028	0.0029	0.0031	0.0034	0.0015	0.0006	0.0017	0.0011	0.0018	0.0014
	192	0.0013	0.0008	0.0008	0.0008	0.0021	0.0022	0.0010	0.0012	0.0010	0.0024	0.0036	0.0037	0.0021	0.0030	0.0031	0.0035	0.0005	0.0008
	336	0.0018	0.0009	0.0038	0.0016	0.0014	0.0011	0.0004	0.0005	0.0058	0.0025	0.0012	0.0005	0.0021	0.0019	0.0016	0.0007	0.0010	0.0018
	720	0.0022	0.0024	0.0028	0.0010	0.0016	0.0018	0.0004	0.0005	0.0023	0.0013	0.0028	0.0017	0.0002	0.0007	0.0010	0.0007	0.0011	0.0004

Table 4 (continued)

Dataset	H	**IsoKoop**		EMAformer		FreEformer		FreqCycle		SOFTS		FreDF		TimeMixer		iTransformer		PatchTST	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	0.0002	0.0003	0.0012	0.0013	0.0012	0.0011	0.0002	0.0001	0.0001	0.0004	0.0012	0.0014	0.0005	0.0007	0.0001	0.0002	0.0001	0.0000
	192	0.0005	0.0002	0.0006	0.0006	0.0018	0.0017	0.0005	0.0002	0.0006	0.0006	0.0011	0.0016	0.0010	0.0002	0.0007	0.0011	0.0003	0.0007
	336	0.0012	0.0008	0.0011	0.0005	0.0017	0.0016	0.0006	0.0004	0.0019	0.0026	0.0012	0.0010	0.0004	0.0006	0.0019	0.0020	0.0004	0.0005
	720	0.0014	0.0017	0.0019	0.0023	0.0057	0.0034	0.0003	0.0003	0.0046	0.0027	0.0041	0.0033	0.0028	0.0019	0.0071	0.0060	0.0000	0.0002
Traffic	96	0.0009	0.0008	0.0036	0.0034	0.0004	0.0002	0.0006	0.0004	0.0005	0.0002	0.0001	0.0001	0.0037	0.0030	0.0001	0.0001	0.0002	0.0015
	192	0.0034	0.0024	0.0021	0.0011	0.0017	0.0018	0.0006	0.0004	0.0002	0.0002	0.0005	0.0002	0.0026	0.0041	0.0009	0.0003	0.0014	0.0018
	336	0.0017	0.0012	0.0034	0.0019	0.0003	0.0017	0.0009	0.0005	0.0002	0.0003	0.0007	0.0005	0.0065	0.0038	0.0003	0.0003	0.0006	0.0003
	720	0.0070	0.0018	0.0034	0.0022	0.0010	0.0004	0.0023	0.0004	0.0002	0.0001	0.0003	0.0006	0.0212	0.0014	0.0012	0.0003	0.0007	0.0002
TrafficC	96	0.0013	0.0011	0.0020	0.0007	0.0003	0.0002	0.0004	0.0004	0.0008	0.0002	0.0008	0.0007	0.0034	0.0030	0.0008	0.0004	0.0003	0.0016
	192	0.0025	0.0018	0.0017	0.0020	0.0005	0.0003	0.0002	0.0002	0.0003	0.0003	0.0002	0.0002	0.0055	0.0054	0.0005	0.0002	0.0012	0.0019
	336	0.0009	0.0020	0.0027	0.0010	0.0007	0.0008	0.0000	0.0005	0.0001	0.0001	0.0009	0.0007	0.0035	0.0042	0.0001	0.0003	0.0002	0.0006
	720	0.0005	0.0011	0.0014	0.0004	0.0012	0.0010	0.0002	0.0004	0.0007	0.0008	0.0019	0.0019	0.0043	0.0051	0.0008	0.0001	0.0005	0.0005
PEMS03	96	0.0040	0.0033	0.0016	0.0011	0.0019	0.0022	0.0009	0.0010	0.0017	0.0010	0.0007	0.0006	0.0111	0.0089	0.0750	0.0690	0.0063	0.0046
	192	0.0040	0.0022	0.0045	0.0028	0.0016	0.0008	0.0004	0.0006	0.0019	0.0014	0.0017	0.0014	0.0146	0.0115	0.0248	0.0264	0.0067	0.0059
	336	0.0057	0.0035	0.0044	0.0032	0.0013	0.0009	0.0003	0.0003	0.0026	0.0015	0.0003	0.0003	0.0153	0.0132	0.0483	0.0427	0.0050	0.0045
	720	0.0027	0.0017	0.0086	0.0058	0.0030	0.0028	0.0006	0.0006	0.0014	0.0009	0.0025	0.0021	0.0034	0.0036	0.0020	0.0011	0.0057	0.0053
PEMS04	96	0.0002	0.0003	0.0017	0.0020	0.0019	0.0009	0.0013	0.0011	0.0003	0.0004	0.0019	0.0017	0.0129	0.0087	0.0007	0.0012	0.0006	0.0007
	192	0.0032	0.0024	0.0055	0.0046	0.0027	0.0008	0.0012	0.0010	0.0026	0.0014	0.0010	0.0014	0.0112	0.0086	0.0039	0.0030	0.0007	0.0013
	336	0.0029	0.0027	0.0120	0.0083	0.0004	0.0007	0.0009	0.0011	0.0012	0.0004	0.0010	0.0007	0.0007	0.0009	0.0018	0.0011	0.0012	0.0009
	720	0.0015	0.0022	0.0093	0.0061	0.0020	0.0009	0.0004	0.0005	0.0029	0.0022	0.0025	0.0016	0.0288	0.0242	0.0049	0.0028	0.0016	0.0010
PEMS07	96	0.0001	0.0007	0.0033	0.0035	0.0276	0.0384	0.0014	0.0014	0.0001	0.0005	0.6168	0.2501	0.0079	0.0076	0.0667	0.0653	0.0012	0.0017
	192	0.0007	0.0010	0.0004	0.0021	0.0016	0.0012	0.0011	0.0010	0.0005	0.0005	0.0018	0.0014	0.0040	0.0059	0.0932	0.0510	0.0006	0.0007
	336	0.0002	0.0004	0.0033	0.0034	0.0025	0.0033	0.0005	0.0006	0.0009	0.0009	0.0006	0.0008	0.0144	0.0126	0.0134	0.0068	0.0005	0.0007
	720	0.0031	0.0016	0.0049	0.0047	0.0013	0.0016	0.0005	0.0004	0.0005	0.0005	0.0032	0.0023	0.0102	0.0086	0.0119	0.0073	0.0011	0.0015
PEMS08	96	0.0014	0.0112	0.0001	0.0033	0.0022	0.0027	0.0021	0.0014	0.0039	0.0013	0.0100	0.0058	0.0247	0.0253	0.0189	0.0077	0.0013	0.0027
	192	0.0023	0.0014	0.0037	0.0022	0.0068	0.0066	0.0029	0.0012	0.0070	0.0035	0.0044	0.0049	0.0133	0.0190	0.0350	0.0252	0.0034	0.0023
	336	0.0041	0.0049	0.0041	0.0093	0.0046	0.0069	0.0017	0.0021	0.0031	0.0011	0.0110	0.0091	0.0101	0.0077	0.0264	0.0124	0.0016	0.0010
	720	0.0026	0.0015	0.0036	0.0046	0.0035	0.0014	0.0024	0.0011	0.0016	0.0017	0.0191	0.0232	0.0087	0.0110	0.0128	0.0084	0.0019	0.0019

Table 5: Exchange against the naive last-value (persistence) forecast under this protocol ($L = 168$, the same test split): test MSE, and in parentheses each model's excess over persistence. Persistence is lower than every model at every horizon.

Model	H=96	H=192	H=336	H=720
Persistence	0.081	0.167	0.306	0.810
IsoKoop	0.087 (+7.4%)	0.193 (+15.6%)	0.376 (+22.9%)	0.901 (+11.2%)
EMAformer	0.084 (+3.7%)	0.175 (+4.8%)	0.336 (+9.8%)	0.833 (+2.8%)
FreEformer	0.089 (+9.9%)	0.187 (+12.0%)	0.362 (+18.3%)	0.837 (+3.3%)
FreqCycle	0.086 (+6.2%)	0.191 (+14.4%)	0.356 (+16.3%)	0.890 (+9.9%)
SOFTS	0.092 (+13.6%)	0.184 (+10.2%)	0.336 (+9.8%)	0.863 (+6.5%)
FreDF	0.124 (+53.1%)	0.236 (+41.3%)	0.409 (+33.7%)	0.987 (+21.9%)
TimeMixer	0.090 (+11.1%)	0.175 (+4.8%)	0.373 (+21.9%)	0.934 (+15.3%)
iTransformer	0.102 (+25.9%)	0.204 (+22.2%)	0.362 (+18.3%)	0.911 (+12.5%)
PatchTST	0.102 (+25.9%)	0.195 (+16.8%)	0.366 (+19.6%)	0.923 (+14.0%)

Where IsoKoop ranks low. In Table 3, IsoKoop is eighth or ninth of the nine models in five cells: in MSE on Exchange at H=336 (0.376; best EMAformer, 0.336), on Solar at H=96 (0.202; best FreDF, 0.183) and on Traffic at H=720 (ninth, 0.491; best SOFTS, 0.431), and in MAE on Solar at H=720 (0.280; best SOFTS, 0.264) and on Traffic at H=720 (0.307; best SOFTS, 0.286). Its lowest remaining rank is seventh, in MAE on TrafficC at H=720 (0.291; best SOFTS, 0.276).

Seed sensitivity of the ranking. Table 6 repeats the ranking of Section 4.2 for each seed on its own. Three seeds give a range, not a confidence interval, and no test is implied.

Table 6: Seed sensitivity of the ranking. For each seed (2027, 2028, 2029) separately, the nine models are ranked within every (dataset, horizon) setting by that seed's full-precision test MSE and MAE (ties averaged; no ties occurred), using the per-seed values whose three-seed mean is the cell printed in Table 1. Columns: IsoKoop's mean rank, the mean rank of the best-ranked competitor (named; its range mixes models), and the number of settings in which IsoKoop has the lowest error. *range* is the minimum–maximum over the three seeds. Three seeds only; no statistical test.

	MSE			MAE		
Seed	IsoKoop	best competitor	wins	IsoKoop	best competitor	wins
All 14 datasets (56 settings)						
2027	2.43	3.91 (EMAformer)	31	2.46	3.93 (EMAformer)	30
2028	2.23	3.98 (FreEformer)	32	2.27	4.11 (FreEformer)	31
2029	2.59	3.88 (EMAformer)	29	2.52	4.00 (EMAformer)	27
range	2.23–2.59	3.88–3.98	29–32	2.27–2.52	3.93–4.11	27–31
Without TrafficC (13 datasets, 52 settings)						
2027	2.54	3.67 (EMAformer)	27	2.40	3.60 (EMAformer)	28
2028	2.33	3.81 (EMAformer)	28	2.17	3.90 (EMAformer)	30
2029	2.71	3.69 (EMAformer)	25	2.44	3.71 (EMAformer)	26
range	2.33–2.71	3.67–3.81	25–28	2.17–2.44	3.60–3.90	26–30

B READINGS OF THE CRITERION

The ratio P/L of a dataset's cycle to the look-back is fixed by the protocol. The readings below are a window-only operator we trained, a read-out of the trained model, a measurement of the raw series, and a comparison between a published model and its own backbone. Traffic, the eighth factorial dataset, is omitted from the rows of Table 2. Its cells, in that table's column order, are $0.372, 0.390, 0.412, 0.365, 0.413, 0.373, 0.461, 0.451$ at $H{=}96$ and $0.483, 0.547, 0.470, 0.465, 0.474, 0.448, 0.520, 0.506$ at $H{=}720$ (seed 2027).

The window-only operator. The $\emptyset$ column of Table 2 is per-window instance normalisation followed by a linear map of the window, an operator that reads nothing but the window. In percent of the complete model's test MSE it costs, at $H{=}96$: PEMS04 609, PEMS08 386, PEMS03 382, Electricity 126, Solar 123, Traffic 121, ETTh1 104, Exchange 99; and at $H{=}720$: PEMS04 375, PEMS08 201, PEMS03 180, Solar 137, Electricity 121, Traffic 105, ETTh1 98, Exchange 87 (Traffic's factorial, which Table 2 omits, from the same runs). Though not within groups, the ordering follows P/L between them, and on Exchange, where persistence already beats every model, reading the window is as good as reading it and supplying the phase.

What the prototype learns. The prototype is zero-initialised, so whatever profile it holds after training was learned from the data. Figure 3(b,c) reads it on the runs of Table 1 (seed 2027): where validation turned it on, it is almost entirely phase-varying on the four PEMS networks (0.87 to 1.00 of its RMS amplitude) and mostly a per-sensor level on Traffic and TrafficC (0.35 to 0.44), with an effective dimension of 2.0 to 3.3 of $R{=}8$ on the PEMS networks and 1.14 to 1.16 of $R{=}16$ on Solar, whose variates share one daily cycle.

The jump the window carries. The circular derivative's error is the jump of the window's periodic extension, and removing the per-phase mean is what takes it out (Statement 2). The probe that measures both on the raw series, with no model involved, ships with the code (`tools/probes/t006_jump_statistic.py`). On Solar and PEMS08, removing the per-phase training mean shrinks the median $|\tilde{\Delta}|$ to $0.805\times$ and $0.863\times$ (Appendix J).

A published architecture. EMAformer is iTransformer's backbone with an explicit phase embedding added, and both are trained in this repository on all fourteen datasets at $L = 168$, four

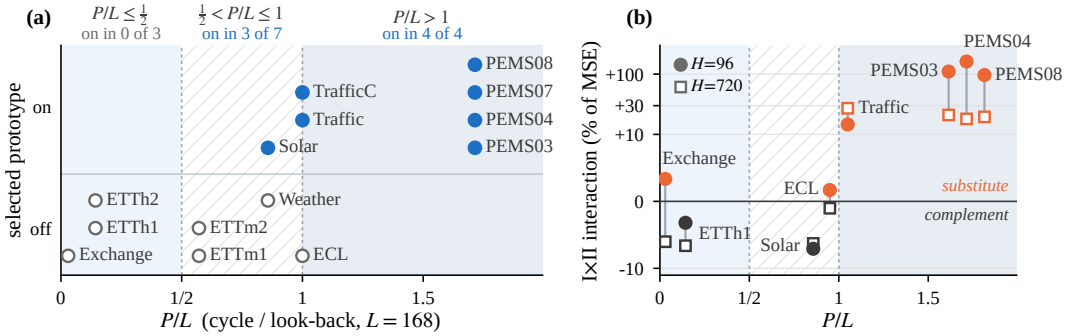


Figure 5: The prototype against P/L. (a) Each dataset at P/L, the dataset cycle over the look-back L=168, and whether validation turned the prototype on (•) or off (○): off at $P/L \le \frac{1}{2}$, on for every sensor network at $P/L > 1$, and on for three of the seven in between (Solar, Traffic and TrafficC): seven of the fourteen in all. **(b)** I×II interaction of the 2^3 factorial (seed 2027, $H \in \{96, 720\}$), % of full-model MSE; > 0: I and II together recover less than the sum of what each recovers alone, an observed pattern.

horizons and three seeds (Table 1). The embedding helps by 29%, 31%, 41% and 62% on PEMS03, PEMS04, PEMS08 and PEMS07, the four panels whose cycle is 1.7 times the look-back, and by 0.5% to 9.6% on the seven whose cycle is shorter than the look-back. Its only losses are at $P = L$, on Traffic (7.6%) and TrafficC (2.5%), where the window already holds one full cycle; on the third such panel, Electricity, it gains 3.0%. Its parameter count grows with the product of the variate count and the cycle length (the joint channel–phase table is one row per phase over Nd columns), which is what carries it to 278.97M parameters on the 883-variate PEMS07 (Section 4.2).

Attention in place of the transport. Table 7 lists every cell of the swap quoted in Section 3.3: the transport, mean field and read-out of stage II replaced by attention of the same depth and width, everything else at the configuration of the factorial, seed 2027. On Electricity both swap runs stop after 8 and 6 epochs at the same validation error (0.5928 at H=96 and 720, against 0.103 and 0.165 for the complete model), a collapse to a near-constant forecast. We report them as failed runs and leave them out of the counts in the text.

Table 7: Attention in place of stage II (seed 2027). Test MSE of the complete model and of the swap, and the swap's MSE and MAE as multiples of the complete model's (> 1: the swap is worse). *Collapsed run (see text).

Dataset	H	complete	attention	MSE ratio	MAE ratio
ETTh1	96	0.365	0.382	1.05	1.03
ETTh1	720	0.466	0.578	1.24	1.14
Solar	96	0.186	0.226	1.22	1.08
Solar	720	0.235	0.222	0.94	0.98
Electricity*	96	0.123	0.676	5.51	2.94
Electricity*	720	0.185	0.659	3.56	2.15
Traffic	96	0.372	0.387	1.04	1.07
Traffic	720	0.483	0.466	0.96	1.03
Exchange	96	0.084	0.084	0.99	1.00
Exchange	720	0.962	0.831	0.86	0.94
PEMS03	96	0.145	0.372	2.57	1.73
PEMS03	720	0.245	0.249	1.01	1.01
PEMS04	96	0.100	0.315	3.16	1.79
PEMS04	720	0.132	0.312	2.37	1.76

C Proof of Proposition 1 and Corollary 1

Proposition 1 (restated). *(i) If $\boldsymbol{\mu}, \mathbf{s}$ are positively homogeneous and $\mathbf{s}(\mathbf{x}) > 0$, every coordinate-native forecaster is scale-equivariant, $f(g\mathbf{x}) = g\, f(\mathbf{x})$ for all $g > 0$, whatever the inner network F. (ii) Let $\boldsymbol{\mu}$ be the per-variate mean and $\mathbf{s}$ the per-variate standard deviation, and let a profile*

$\mathbf{q}_h \in \mathbb{R}^{L\times N}$ *that does not depend on the window's values be subtracted in raw units before* $\mathcal{N}$ *and* $\mathbf{q}_f \in \mathbb{R}^{H\times N}$ *added after* $\mathcal{N}^{-1}$*. Then* f *is scale-equivariant for every* F *if and only if* $\mathbf{q}_h = \mathbf{q}_f = \mathbf{1}c^\top$ *is a constant level.*

In (ii) the forecaster is $f(\mathbf{x}) = \mathcal{N}^{-1}_{\mathbf{x}-\mathbf{q}_h}(F(\mathcal{N}_{\mathbf{x}-\mathbf{q}_h}(\mathbf{x}-\mathbf{q}_h))) + \mathbf{q}_f$, and $\mathbf{q}_h$ may depend on the sample index. In (i) it is $f = \mathcal{N}^{-1}_{\mathbf{x}} \circ F \circ \mathcal{N}_{\mathbf{x}}$.

Corollary 1 (restated). *Under the hypotheses of part (i), each of the other three placements IsoKoop implements, switching the profile off, subtracting it inside the coordinates* ($\mathbf{v} \mapsto F(\mathbf{v} - \tilde{\mathbf{q}}_h) + \tilde{\mathbf{q}}_f$) *and rescaling it by* $\mathbf{s}(\mathbf{x})$ *before* $\mathcal{N}$ *(exactly while the implementation's clamp on the scale ratio is inactive at* $\mathbf{x}$ *and* $g\mathbf{x}$ *and its horizon blend is off), is scale-equivariant for every* F*. So is a published cycle table subtracted after its normaliser.*

Proof. First half. Fix $g > 0$. Positive homogeneity gives $\boldsymbol{\mu}(g\mathbf{x}) = g\boldsymbol{\mu}(\mathbf{x})$ and $\mathbf{s}(g\mathbf{x}) = g\mathbf{s}(\mathbf{x}) > 0$, so $\mathcal{N}_{g\mathbf{x}}(g\mathbf{x}) = (g\mathbf{x} - g\boldsymbol{\mu}(\mathbf{x}))/(g\mathbf{s}(\mathbf{x})) = \mathcal{N}_{\mathbf{x}}(\mathbf{x})$ and F receives an identical input $\mathbf{v}$. Hence $f(g\mathbf{x}) = g\mathbf{s}(\mathbf{x})F(\mathbf{v}) + g\boldsymbol{\mu}(\mathbf{x}) = gf(\mathbf{x})$. No property of F was used.

Second half. Here $\boldsymbol{\mu}$ is the per-variate mean, which is linear, and $\mathbf{s}$ the per-variate standard deviation, which is positively homogeneous and unchanged by adding a per-variate constant; a positively homogeneous statistic in general need be neither (the median is not linear). Write $G(\mathbf{u}) = \mathcal{N}^{-1}_{\mathbf{u}}(F(\mathcal{N}_{\mathbf{u}}(\mathbf{u}))) = \mathbf{s}(\mathbf{u})F(\mathcal{N}_{\mathbf{u}}(\mathbf{u})) + \boldsymbol{\mu}(\mathbf{u})$, so $f(\mathbf{x}) = G(\mathbf{x}-\mathbf{q}_h) + \mathbf{q}_f$. The mean shifts with a per-variate constant and the standard deviation does not, so $\mathcal{N}_{\mathbf{u}+\mathbf{1}c^\top}(\mathbf{u}+\mathbf{1}c^\top) = \mathcal{N}_{\mathbf{u}}(\mathbf{u})$ and $G(\mathbf{u}+\mathbf{1}c^\top) = G(\mathbf{u}) + \mathbf{1}c^\top$ for every $c \in \mathbb{R}^N$.

If. With $\mathbf{q}_h = \mathbf{1}c^\top$ and $\mathbf{q}_f = \mathbf{1}c^\top$, $f(\mathbf{x}) = G(\mathbf{x}) - \mathbf{1}c^\top + \mathbf{1}c^\top = G(\mathbf{x})$, which is equivariant by the first half.

Only if. Suppose f is equivariant for every F. With $F \equiv 0$, $f(\mathbf{x}) = \mathbf{1}(\boldsymbol{\mu}(\mathbf{x}) - \boldsymbol{\mu}(\mathbf{q}_h))^\top + \mathbf{q}_f$ by linearity of the mean, and $f(g\mathbf{x}) - gf(\mathbf{x}) = (1-g)\big(\mathbf{q}_f - \mathbf{1}\boldsymbol{\mu}(\mathbf{q}_h)^\top\big)$; equivariance at any $g \neq 1$ forces $\mathbf{q}_f = \mathbf{1}c^\top$ with $c = \boldsymbol{\mu}(\mathbf{q}_h)$. Put $\mathbf{p} = \mathbf{q}_h - \mathbf{1}c^\top$, which has zero mean in every variate; the shift property gives $f(\mathbf{x}) = G(\mathbf{x}-\mathbf{p})$. Suppose some column $\mathbf{p}_{:,i} \neq 0$; being non-zero with zero mean it is non-constant. Choose $\mathbf{x}$ with $\mathbf{x}_{:,j} = 0$ for every j with $\mathbf{p}_{:,j} \neq 0$ and $\mathbf{x}_{:,j}$ an arbitrary non-constant series for every j with $\mathbf{p}_{:,j} = 0$. Then every variate of $g\mathbf{x} - \mathbf{p}$ is non-constant, and its normalised form $\mathbf{v}$ does not depend on g: the first group is $-\mathbf{p}_{:,j}$ for every g, the second is $g\mathbf{x}_{:,j}$, which the normaliser maps to the same series. Take $F \equiv \mathbf{1}\mathbf{1}^\top$. Then $f_{:,i}(g\mathbf{x}) = \mathbf{s}_i(\mathbf{p})\mathbf{1} + \boldsymbol{\mu}_i(-\mathbf{p})\mathbf{1} = \mathbf{s}_i(\mathbf{p})\mathbf{1}$ for every g, while $gf_{:,i}(\mathbf{x}) = g\,\mathbf{s}_i(\mathbf{p})\mathbf{1}$; since $\mathbf{s}_i(\mathbf{p}) > 0$ they differ at every $g \neq 1$, a contradiction. Hence $\mathbf{p} = 0$ and $\mathbf{q}_h = \mathbf{1}c^\top = \mathbf{q}_f$. □

Scope of the only-if half. Its witnesses are constant networks, so it removes a guarantee, not the equivariance of every F. In general the before-normaliser f is equivariant iff $g\,[G(\mathbf{x} - \mathbf{q}_h/g) - G(\mathbf{x} - \mathbf{q}_h)] = (g-1)\,\mathbf{q}_f$ for all $\mathbf{x}$ and $g > 0$, a property of F as well as of the placement. For example, let $F(\mathbf{v}) = M\mathbf{v}$ act per variate with $M \in \mathbb{R}^{H\times L}$ and $M\mathbf{1}_L = \mathbf{1}_H$. Then $G(\mathbf{u}) = \mathbf{s}(\mathbf{u})M\mathcal{N}_{\mathbf{u}}(\mathbf{u}) + \boldsymbol{\mu}(\mathbf{u}) = M\mathbf{u}$, so $f(\mathbf{x}) = M(\mathbf{x} - \mathbf{q}_h) + \mathbf{q}_f$, which is exactly equivariant whenever $\mathbf{q}_f = M\mathbf{q}_h$, for a prototype with any time profile. The degradation measured in Section 4.5 is therefore realised by the two trained non-linear networks, IsoKoop's and FreqCycle's, which the placement admits.

Per-variate gains. Because $\boldsymbol{\mu}$ and $\mathbf{s}$ are per variate, the first half holds verbatim for a per-variate gain $\mathbf{x} \mapsto \mathbf{x}\,\mathrm{diag}(\mathbf{g})$, $\mathbf{g} \in \mathbb{R}^N_{>0}$: the normalised input is unchanged and $f(\mathbf{x}\,\mathrm{diag}(\mathbf{g})) = f(\mathbf{x})\,\mathrm{diag}(\mathbf{g})$, for any F, including one that mixes variates.

The variance floor. The implementations use $\mathbf{s}_\epsilon(\mathbf{x}) = \sqrt{\mathrm{var}(\mathbf{x}) + \epsilon}$ with $\epsilon = 10^{-5}$, which is not positively homogeneous: $\mathbf{s}_\epsilon(g\mathbf{x}) = g\,\mathbf{s}_{\epsilon/g^2}(\mathbf{x})$. The first half therefore holds with the floor ϵ replaced by ϵ/g^2 on the scaled window, a relative change of the normalised input of $O(|1 - g^{-2}|\,\epsilon/\sigma^2)$ for window variance σ^2: of order 10^{-5} on windows of unit-order variance, and of order one on nearly constant ones. IsoKoop-R's scale subtracts a fixed training-time noise reference and is not positively homogeneous at all, so neither half applies to it.

Proof of Corollary 1. Switching the profile off is part (i). Subtracting a table independent of the window's values $(\tilde{\mathbf{q}}_h, \tilde{\mathbf{q}}_f)$ in normalised units, after the normaliser, gives $f(\mathbf{x}) = \mathcal{N}_{\mathbf{x}}^{-1}(F(\mathcal{N}_{\mathbf{x}}(\mathbf{x}) - \tilde{\mathbf{q}}_h) + \tilde{\mathbf{q}}_f)$, which is the first half with the admissible $F'(\mathbf{v}) = F(\mathbf{v} - \tilde{\mathbf{q}}_h) + \tilde{\mathbf{q}}_f$. Equivariance holds for every table. Rescaling the prototype by the window's own scale, $\mathbf{q} \mapsto (\mathbf{s}(\mathbf{x})/\mathbf{s}_{\mathrm{ref}})\mathbf{q}$, makes both $\mathbf{q}_h$ and $\mathbf{q}_f$ positively homogeneous of degree one in $\mathbf{x}$, so $f(g\mathbf{x}) = G(g(\mathbf{x} - \mathbf{q}_h(\mathbf{x}))) + g\mathbf{q}_f(\mathbf{x}) = gf(\mathbf{x})$. In the implementation the ratio is clamped to $[0.1, 10]$, so this holds exactly for window-scale ratios inside the clamp (and up to the variance floor above).

The blended horizon profile (Solar). With `shrink` on (Solar), $\mathbf{q}_h$ is unchanged and the horizon profile is $\mathbf{q}_f(\mathbf{x}) = (\mathbf{1}\mathbf{1}^\top - \Lambda) \odot \mathbf{q}_f + \Lambda \odot A\mathbf{x}$, where A is a fixed linear map, per variate, that averages the in-window samples at each horizon step's phase, and $\Lambda \in [0, 1)^{H \times N}$ a learned per-variate weight (zero at a phase the window does not hold). The added term is linear in $\mathbf{x}$, so it scales with g. In the proof, $F \equiv 0$ now forces $(\mathbf{1}\mathbf{1}^\top - \Lambda) \odot \mathbf{q}_f = \mathbf{1}\boldsymbol{\mu}(\mathbf{q}_h)^\top$, and the second step is unchanged, because the constructed window is zero on every variate with $\mathbf{p}_{:,j} \neq 0$, where $A\mathbf{x}$ vanishes too. Hence f is equivariant for every F if and only if $\mathbf{q}_h = \mathbf{1}c^\top$ and $(\mathbf{1}\mathbf{1}^\top - \Lambda) \odot \mathbf{q}_f = \mathbf{1}c^\top$: Solar's row is covered, and a prototype with any time profile removes the guarantee there as well.

Offsets, and the units of the gain test. The second half concerns gains only. The shift property in the proof, $G(\mathbf{u} + \mathbf{1}c^\top) = G(\mathbf{u}) + \mathbf{1}c^\top$, gives $f(\mathbf{x} + \mathbf{1}c^\top) = G(\mathbf{x} - \mathbf{q}_h + \mathbf{1}c^\top) + \mathbf{q}_f = f(\mathbf{x}) + \mathbf{1}c^\top$ for every $\mathbf{q}_h, \mathbf{q}_f$ and every F: a prototype subtracted before the normaliser keeps exact equivariance to a whole-window offset. With Solar's blended profile the added term $\Lambda \odot A\mathbf{x}$ moves by $\Lambda \odot \mathbf{1}c^\top$ instead of $\mathbf{1}c^\top$, a departure of the size of Λ. The gain test is applied in the model's train-standardised input units. A gain g in raw units is there the map $\mathbf{x} \mapsto g\mathbf{x} + \mathbf{1}b^\top$ with a per-variate b, and since the offset part passes through either placement exactly, the raw-unit test breaks or keeps equivariance exactly as the train-standardised one does.

What the gain test measures. Positive homogeneity of G (Section 3.1) gives the test a direct reading. For a prototype in raw units $\frac{1}{g}f(g\mathbf{x}) = G(\mathbf{x} - \frac{1}{g}\mathbf{q}_h) + \frac{1}{g}\mathbf{q}_f$, so scoring $f(g\mathbf{x})/g$ against $\mathbf{y}$ scores the original window with its prototype at $1/g$ of its amplitude. The test is a controlled mis-scaling of the profile, and a network departs from the line as far as its forecast depends on that amplitude. Under the in-coordinate placements $f(g\mathbf{x})/g = f(\mathbf{x})$.

Measurement. Figure 4, Table 8 and, seed by seed, Figure 6. Every IsoKoop run there, the raw-unit reference included, is a separate training of the design of Table 1 at the same seeds. The rescaled placement clamps the scale ratio to $[0.1, 10]$, and with the variance floor above it returns to the line within the tenth of a percent the table prints. The other two return exactly up to the variance floor $\epsilon = 10^{-5}$.

C.1 The in-coordinate placements across horizons and under real shift

Table 9 follows the in-coordinate placements beyond the gain test, each against the raw-unit design. Panel (a) trains the after-normaliser placement at every horizon on the three datasets of the gain test: it is dearer on PEMS04 and PEMS08 at every horizon, and on Solar it gains only at H=96. Panel (b) bins every test window and variate by how far the window's scale has moved from the training scale, $|\log r_{w,i}|$ with $r_{w,i} = s(\mathbf{x}_{w,i})/s_{\mathrm{train},i}$ the ratio of the normaliser's scale to the variate's training scale. Every scored arm has $S > 0$: each placement fares relatively worse where the scale moved most, and on Solar its gain shrinks there. Because $s_{\mathrm{train},i}$ is taken over the whole training split, the bins also track where a window sits in the daily cycle. Panel (c) takes Traffic's nine drifting sensors, which carry a real train-to-test level and scale shift (Appendix D): both placements raise the error on these sensors, and the total MSE, on every seed. On the real shifts tested, the placements do not lower the error.

D Traffic and TrafficC

TrafficC removes the 1% of Traffic's sensors ($\lceil 0.01 \times 862 \rceil = 9$) with the largest variance of the train-standardised test series, the quantity the metric is computed in (Table 10b). None is stuck or dead. Each changes level and scale after the training period, so under the benchmark's train-statistics

Table 8: Proposition 1 on trained networks, $H{=}96$, three seeds. **(a)** Change of test MSE under a whole-window gain (mean over $g \in \{0.5, 2\}$; window and target multiplied by g, the forecast divided by g), against each seed's clean MSE; the three columns after IsoKoop are its in-coordinate placements (Corollary 1). IsoKoop-R's scale is not positively homogeneous, so the proposition does not cover it. ‡FreDF and iTransformer run PEMS08 without instance normalisation, their papers' setting. **(b)** Clean test MSE of each placement per seed, its change against the raw-unit IsoKoop (on the seed means, then per seed) and its gain degradation; FreqCycle with its table moved in front of its normaliser, and moved and rescaled, is compared with FreqCycle as published. On Solar the raw-unit reference is a separate training of the same design; against the runs of Table 1 the three placements change clean MSE by -10.82%, -11.83% and -12.78%, and the raw-unit reference's own seeds are 0.1931, 0.2292 and 0.1965, the second an outlier, so that row is read as a three-seed mean with its spread in view. **(c)** IsoKoop-R's clean change against IsoKoop on the eight corruption-benchmark datasets; the Solar cells are separate retrainings of the same design in the robustness campaign, and Solar has the widest seed spread of the fourteen datasets (0.192 to 0.209 at $H{=}96$), which is what separates the 0.1925 here from the 0.202 of Table 3.

(a) Gain degradation

Dataset	prototype	IsoKoop	$\mathbf{q}$ after $\mathcal{N}$	$\mathbf{q}$ rescaled	$\mathbf{q}$ off	IsoKoop-R	SOFTS	FreDF	iTransformer	FreqCycle
ETTh1	off	0.0%	–	–	–	−0.1%	0.0%	−0.0%	0.0%	−0.0%
Weather	off	0.0%	–	–	–	0.0%	0.0%	−0.0%	−0.0%	−0.0%
PEMS08	on	+89.4%	−0.0%	−0.1%	−0.0%	+245.4%	0.0%	+230.8%‡	+95.8%‡	−0.0%
PEMS04	on	+153.6%	−0.0%	−0.0%	−0.0%	+524.7%	0.0%	0.0%	−0.0%	−0.0%
Solar	on	+141.5%	−0.0%	−0.1%	−0.0%	+154.7%	−0.0%	−0.0%	−0.0%	−0.0%

(b) Clean cost of each placement

Dataset	arm	2027	2028	2029	cost	cost per seed	gain degr.
PEMS04	IsoKoop (raw units)	0.0995	0.0993	0.1000	–	–	+153.6%
	$\mathbf{q}$ after $\mathcal{N}$	0.1025	0.1031	0.1011	+2.62%	+2.98 / +3.82 / +1.07	−0.0%
	$\mathbf{q}$ rescaled	0.1031	0.1023	0.1014	+2.69%	+3.59 / +3.09 / +1.41	−0.0%
	$\mathbf{q}$ off	0.1034	0.1020	0.1025	+3.06%	+3.94 / +2.74 / +2.49	−0.0%
	FreqCycle	0.2001	0.2003	0.1979	–	–	−0.0%
	FreqCycle, table moved	0.2128	0.2130	0.2120	+6.59%	+6.35 / +6.33 / +7.11	+144.3%
	FreqCycle, moved and rescaled	0.2303	0.2306	0.2302	+15.50%	+15.10 / +15.15 / +16.27	−0.0%
PEMS08	IsoKoop (raw units)	0.1797	0.1778	0.1805	–	–	+89.4%
	$\mathbf{q}$ after $\mathcal{N}$	0.1986	0.1971	0.2040	+11.48%	+10.51 / +10.87 / +13.05	−0.0%
	$\mathbf{q}$ rescaled	0.1889	0.1880	0.1933	+6.00%	+5.14 / +5.73 / +7.12	−0.1%
	$\mathbf{q}$ off	0.1990	0.1966	0.1919	+9.22%	+10.74 / +10.61 / +6.34	−0.0%
	FreqCycle	0.2972	0.2932	0.2942	–	–	−0.0%
	FreqCycle, table moved	0.3216	0.3203	0.3215	+8.91%	+8.22 / +9.25 / +9.25	+131.9%
	FreqCycle, moved and rescaled	0.3248	0.3268	0.3346	+11.48%	+9.30 / +11.45 / +13.71	+0.8%
Solar	IsoKoop (raw units)	0.1931	0.2292	0.1965	–	–	+141.5%
	$\mathbf{q}$ after $\mathcal{N}$	0.1767	0.1828	0.1814	−12.59%	−8.50 / −20.25 / −7.67	−0.0%
	$\mathbf{q}$ rescaled	0.1766	0.1718	0.1864	−13.57%	−8.54 / −25.04 / −5.14	−0.1%
	$\mathbf{q}$ off	0.1773	0.1752	0.1766	−14.51%	−8.20 / −23.57 / −10.14	−0.0%
	FreqCycle	0.1880	0.1883	0.1893	–	–	−0.0%
	FreqCycle, table moved	0.1780	0.1772	0.1786	−5.62%	−5.31 / −5.87 / −5.67	+210.6%
	FreqCycle, moved and rescaled	0.1867	0.1857	0.1867	−1.14%	−0.69 / −1.36 / −1.38	−0.4%

(c) IsoKoop-R

Dataset	IsoKoop	IsoKoop-R	cost
Electricity	0.1228	0.1228	+0.04%
ETTh1	0.3640	0.3652	+0.33%
ETTm1	0.2899	0.2902	+0.12%
PEMS04	0.0996	0.1026	+2.97%
PEMS08	0.1793	0.1796	+0.16%
Solar	0.1925	0.1899	−1.31%
Traffic	0.3725	0.3664	−1.66%
Weather	0.1475	0.1478	+0.17%

standardisation its test values reach 15 to 39 standard deviations, against a median variance of 1.12 over the 853 kept sensors. Every model was retrained on the 853-sensor file (Table 10a): all nine improve by 20.9% to 29.4%, and the ranking changes, IsoKoop from sixth to first and PatchTST from a fourth-place tie at the printed precision to ninth. The script that builds TrafficC ships with the code. Figure 7 places all 862 sensors on the same criterion.

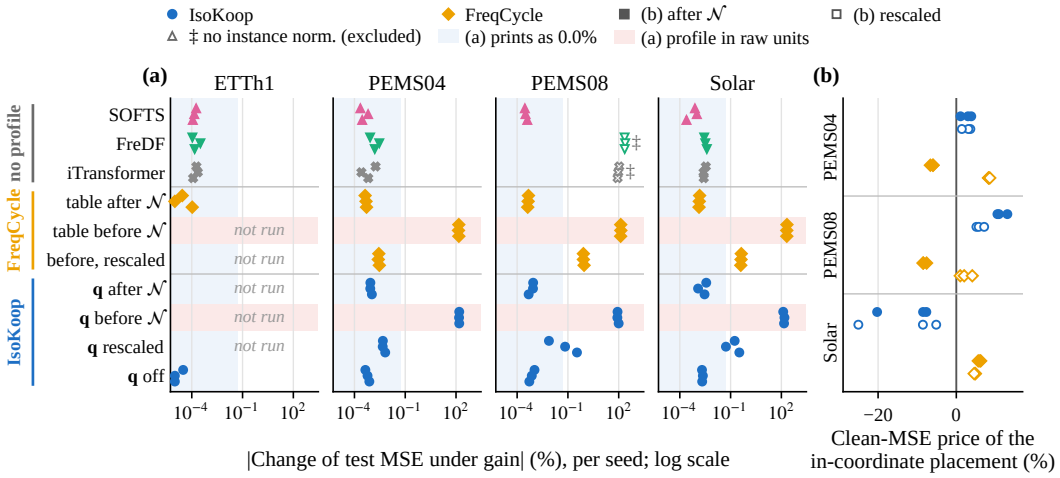


Figure 6: Seed by seed, the gain test separates the placements of Corollary 1 on both networks that carry a prototype, and the clean price of the in-coordinate placement has no common sign. **(a)** Absolute change of test MSE under a whole-window gain (log scale; $H=96$; each marker is one of three seeds) for the three comparison models without a profile, for FreqCycle with its cycle table after its normaliser (as published), moved in front of it, and moved and rescaled, and for IsoKoop's four placements of **q**. Shaded: values that print as 0.0% in Table 8; tinted rows: the profile in raw units. Apart from FreDF and iTransformer on PEMS08‡, every placement after the normaliser or without a profile stays at floating-point level, the rescaled placements stay near the line, and moving the profile into raw units takes either network orders of magnitude off the line, on every seed. On ETTh1 the validated IsoKoop has no prototype and is drawn in the **q** off row. ‡No instance normalisation (excluded, as in Figure 4). **(b)** Clean-MSE change of the in-coordinate placement against the raw-unit one, per seed. For the after-normaliser placement (filled) the sign is opposite for the two networks on every seed of every dataset. Electricity has no gain-test runs.

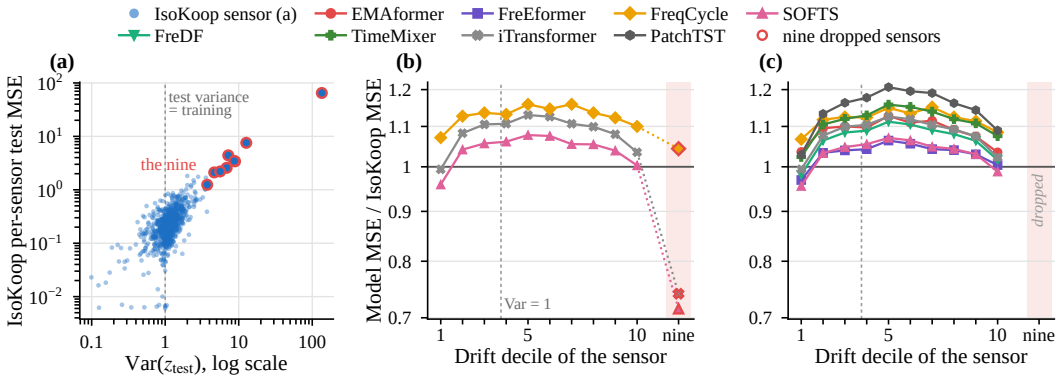


Figure 7: Sensor by sensor, how each comparison model's error relative to IsoKoop's varies with drift strength on Traffic and TrafficC. Drift strength is $\mathrm{Var}(z_{\mathrm{test}})$, the variance of a sensor's test series standardised by training statistics, which is the criterion that defines TrafficC (Table 10, part (b)). Deciles are taken over the 853 sensors TrafficC keeps (dashed: $\mathrm{Var}(z_{\mathrm{test}})=1$). **(a)** IsoKoop's per-sensor test MSE on Traffic ($H=96$, seed 2027); ringed: the nine sensors TrafficC drops. **(b)** Bin-mean MSE of each comparison model divided by IsoKoop's, on the runs of Figure 8b (above 1: IsoKoop lower). **(c)** The same ratio on TrafficC, with every model re-trained on the 853-sensor file ($H=96$, per-sensor MSE averaged over three seeds).

E THE BENCHMARKS THROUGH THE COORDINATE LENS

Figure 8 collects the three readings of Section 4.6: the share of a dataset's variates whose test-period scale leaves the range it takes over training windows, the share of the $H=96$ error carried by Traffic's nine drifting sensors for every model whose per-sensor error we recorded, and Exchange against the persistence forecast. The lens also reads the error as well as the input: on some panels a handful of fixed channels carry most of the error, and on others it is spread evenly. Chosen by IsoKoop's error,

Table 9: The in-coordinate placements beyond the gain test. **(a)** Clean test MSE of the after-normaliser placement against the raw-unit design, in percent, three-seed means, at every horizon. **(b)** Per test window (H=96, seed 2027): window–variate pairs binned by quartiles of $|\log r_{w,i}|$, the movement of the window scale against the training scale (b_1 least moved); $D(b)$ is the placement's relative test MSE against the raw-unit design in bin b (%), and $S = D(b_4) - D(b_1)$ in points; arms whose re-run did not reproduce its first run within 1% are not scored. **(c)** Traffic's nine drifting sensors (H=96): total MSE, absolute error of the worst 1% of variates (the nine) and MSE of the other 99%, with Δ in percent against the raw-unit design of the same seed.

(a) Across horizons

Dataset	H=96	H=192	H=336	H=720
PEMS04	+2.62	+9.16	+6.23	+6.63
PEMS08	+11.48	+6.83	+2.91	+3.40
Solar	−12.59	+4.95	+8.86	+12.07

(b) Per test window

Dataset	arm	$D(b_1)$	$D(b_2)$	$D(b_3)$	$D(b_4)$	S
PEMS04	**q** after $\mathcal{N}$	not scored				
	q rescaled	+1.77	+2.55	+1.18	+6.98	+5.21
	q off	+0.72	+1.71	+0.34	+9.26	+8.54
PEMS08	**q** after $\mathcal{N}$	+2.12	+1.56	+1.26	+26.95	+24.83
	q rescaled	−0.85	−0.77	−1.66	+16.70	+17.55
Solar	**q** after $\mathcal{N}$	not scored				
	q rescaled	−13.78	−12.40	−9.43	−4.94	+8.84

(c) Traffic's nine drifting sensors

Seed	design	MSE	Δ	worst 1%	Δ	other 99%	Δ
2027	raw units	0.3715		0.1074		0.2669	
	q after $\mathcal{N}$	0.3880	+4.45	0.1242	+15.69	0.2665	−0.12
	q rescaled	0.3755	+1.08	0.1121	+4.46	0.2661	−0.29
2028	raw units	0.3732		0.1075		0.2684	
	q after $\mathcal{N}$	0.3857	+3.35	0.1233	+14.68	0.2651	−1.23
	q rescaled	0.3735	+0.08	0.1100	+2.33	0.2661	−0.84
2029	raw units	0.3729		0.1089		0.2666	
	q after $\mathcal{N}$	0.3820	+2.46	0.1190	+9.27	0.2657	−0.35
	q rescaled	0.3731	+0.08	0.1103	+1.27	0.2655	−0.41

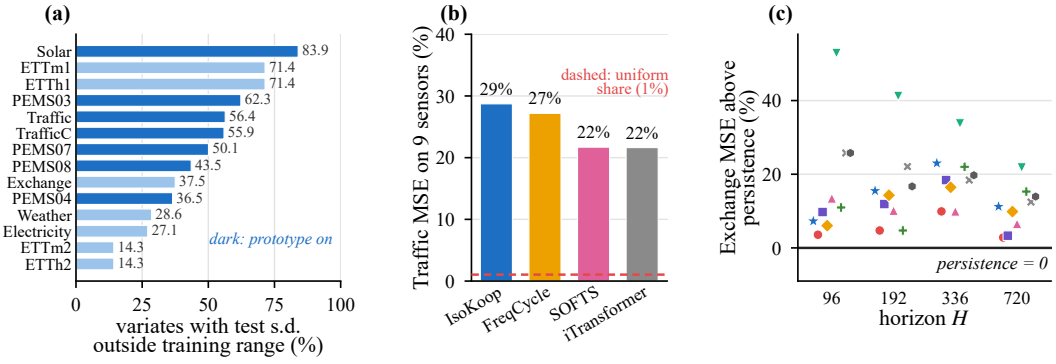


Figure 8: The benchmarks through the coordinate lens. **(a)** Share of variates whose test-period s.d. leaves its range over equal-length training windows (dark: prototype on). **(b)** Share of Traffic's H=96 test MSE carried by nine drifting sensors, 1% of the network, in the four seed-2027 runs that recorded per-sensor error; for SOFTS and iTransformer these are not table runs, hence two models in Table 10a. **(c)** Exchange: each model's MSE above the persistence forecast (markers as in Figure 19); no model reaches it at any horizon.

the worst 1% of channels carry 36% to 61% of its MSE on PEMS08 (2 of 170) and 29% to 32% on Traffic (9 of 862) at seed 2027, and 0.9% to 1.0% on Solar (1 of 137, against 0.73% if even).

Table 10: Traffic and TrafficC. **(a)** Every model retrained on TrafficC (853 sensors), averages over the four horizons and three seeds; *top*-1% *share* is the fraction of a run's Traffic MSE held by its worst 1% of sensors, for the two models whose runs recorded it. **(b)** The nine sensors removed to form TrafficC: the 1% of sensors with the largest variance of the train-standardised test series. None is stuck or dead; each changes level and scale after the training period.

(a) Every model on both panels

Model	Traffic	TrafficC	change	Traffic rank	TrafficC rank	top-1% share	
						H=96	all H
IsoKoop	0.4252	0.3000	−29.4%	6	1	29.0%	30.0%
SOFTS	0.3915	0.3095	−20.9%	1	2	–	–
FreEformer	0.4240	0.3100	−26.9%	5	3	–	–
FreDF	0.3963	0.3130	−21.0%	2	4	–	–
iTransformer	0.3973	0.3143	−20.9%	3	5	–	–
FreqCycle	0.4365	0.3215	−26.3%	9	6	27.2%	27.2%
EMAformer	0.4270	0.3220	−24.6%	7	7	–	–
TimeMixer	0.4350	0.3233	−25.7%	8	8	–	–
PatchTST	0.4235	0.3337	−21.2%	4	9	–	–

(b) The nine sensors

Sensor	train mean	train sd	test mean	test sd	test/train sd	$\mathrm{Var}(z_{\text{test}})$	$\max \lvert z_{\text{test}} \rvert$
840	0.0318	0.0154	0.1005	0.1790	11.61	134.9	39.3
567	0.0254	0.0156	0.0476	0.0556	3.56	12.7	26.8
724	0.0278	0.0161	0.0394	0.0481	2.98	8.9	15.0
728	0.0243	0.0236	0.0550	0.0696	2.95	8.7	17.0
722	0.0253	0.0244	0.0296	0.0655	2.68	7.2	23.0
642	0.0240	0.0173	0.0482	0.0453	2.62	6.9	16.6
668	0.0324	0.0232	0.0580	0.0549	2.36	5.6	15.8
512	0.0290	0.0169	0.0375	0.0365	2.15	4.6	20.9
817	0.0352	0.0166	0.0459	0.0321	1.93	3.7	15.8

F ROBUSTNESS PROTOCOL

Corruptions. Each trained model is evaluated on the test split under 25 corruptions from seven families, applied per window in train-standardised units ($T = L$ input steps): *scale* $s \in \{0.5, 1.5, 2\}$, the last $T/4$ input steps and the whole target multiplied by s; *offset* $o \in \{-1, 0.5, 1, 2\}$, the same span and the target shifted by o; *missing* $p \in \{0.1, 0.2, 0.3, 0.5\}$, Bernoulli-lost entries filled by the last observation; *noise* $\sigma \in \{0.05, 0.1, 0.2, 0.5, 1\}$, i.i.d. Gaussian noise on every input entry; *stuck* $f \in \{0.1, 0.3, 0.5\}$, a fraction f of channels frozen at their value $T/2$ steps back; *dead* $f \in \{0.1, 0.3, 0.5\}$, a fraction f of channels reading the training mean; and *lag* $d \in \{2, 6, 24\}$, half the channels delayed by d steps. Scale and offset are undone on the forecast before scoring, so every metric is in the unshifted units. Because they alter only part of the window, Proposition 1 does not make an equivariant model score zero on them. The random draws are seeded by the corruption and the batch index, so every model sees the identical corruption. The gain test of Section 4.5 is the scale family applied to the whole window.

Cells and wins. The eight datasets are Electricity, ETTh1, ETTm1, PEMS04, PEMS08, Solar, Traffic and Weather, at $H = 96$ and three seeds, so there are $8 \times 25 = 200$ cells. The line-up is one IsoKoop variant and four comparison models (iTransformer, FreDF, SOFTS, FreqCycle). A cell is won by the lowest three-seed mean corrupted test MSE, so the exchangeability reference is $1/5$. IsoKoop-R wins 100 of the 200 cells and IsoKoop 85. Resampling datasets, IsoKoop-R's bootstrap interval for its share of won cells is $[35.5, 63.5]\%$ (88 of its 100 wins hold on every seed) and IsoKoop's $[27.5, 57.5]\%$. On clean inputs IsoKoop-R costs at most 0.33% on seven of the eight datasets and 2.97% on PEMS04 (Table 8c).

IsoKoop-R's scale. For variate i of window $\mathbf{x}$, let $\hat{\sigma}_i^2(\mathbf{x})$ be the median of $|\mathrm{rfft}(\mathbf{x}_i)|^2$ over the upper half of the frequency axis divided by $L \ln 2$ (for white noise $|X(\omega)|^2$ is exponential with mean $L\sigma^2$), and $\bar{\sigma}_i^2$ its running mean over training windows. The median-over-frequency estimate

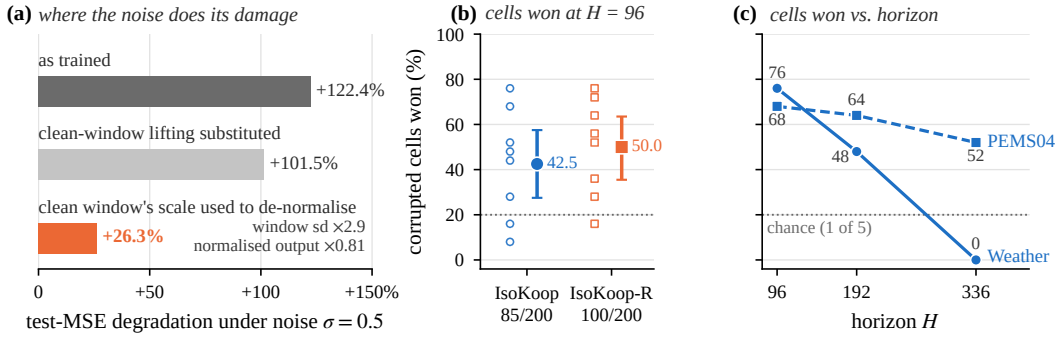


Figure 9: Input noise acts through de-normalisation. **(a)** One Weather model (seed 2027, H=96, noise σ=0.5): substituting the clean window's lifting barely helps, while de-normalising with the clean window's mean and scale removes most of the damage (each substitution alone). **(b)** Share of the 200 corrupted cells won against the four comparison models; bar: 95% dataset-bootstrap interval; open markers: single datasets; dotted: chance. **(c)** IsoKoop's share on Weather and PEMS04 with the same line-up also trained at H=192 and 336 (25 cells per dataset and horizon); the benchmark of (b) is H=96 only.

is a periodogram analogue of the median-absolute-deviation noise estimate of Donoho & Johnstone (1994). IsoKoop-R replaces the standard deviation in equation 1 by $s_i^{\mathrm{R}}(\mathbf{x}) = \big(\max\{\mathrm{var}(\mathbf{x}_i) - \max(0, \hat{\sigma}_i^2(\mathbf{x}) - \bar{\sigma}_i^2),\ 0.05\,\mathrm{var}(\mathbf{x}_i)\} + 10^{-5}\big)^{1/2}$: on a window whose noise floor is at or below the training reference it is the ordinary standard deviation, and it is not positively homogeneous.

The noise path. On one trained Weather model (seed 2027, H=96, noise σ=0.5), substituting the lifting computed on the clean window into the corrupted forward pass leaves the degradation at $+101.5\%$ against $+122\%$, and de-normalising the unmodified corrupted pass with the clean window's mean and scale cuts it to $+26\%$ (Figure 9a): noise inflates the window's standard deviation while the normalised forecast barely moves, and de-normalisation multiplies it back in.

G Forecasts, error spectra and mechanism read-outs

Four figures show what the models forecast (Figures 10, 11, 15 and 17): a typical and a hard window per dataset, and whether IsoKoop's hardest window is hard for every model. Three locate the error in frequency (Figures 12, 16 and 18), separating a level offset from a spectral error. Two put the order read-out of Section 3.3 against each variate's spectral entropy (Figures 13 and 14). Variates and windows follow the fixed rule stated in each caption. Every figure comes from single-seed training runs at the configuration of Table 1 and quotes no error values.

H Protocol and datasets

Every model runs the same pipeline: $L = 168$, $H \in \{96, 192, 336, 720\}$, all variates in and out, seeds 2027/2028/2029, at most 20 epochs with each paper's own patience, and per-window instance normalisation. ETT runs on its own split, the PEMS panels on a 0.6/0.2/0.2 split and every other dataset on 0.7/0.1/0.2. Comparison models use the hyperparameters published in their official repositories (PatchTST through the Time-Series-Library port). A model with no script for a dataset uses the nearest dataset's recipe from the same repository, and PEMS07 and TrafficC use the PEMS04 and Traffic recipes. FreDF is the FreDF loss on its paper's headline backbone, iTransformer. iTransformer and FreDF run PEMS08 without instance normalisation, their papers' setting. As in their public code, iTransformer, FreDF, SOFTS and TimeMixer embed the timestamp features as extra tokens except on PEMS and Solar, which carry none; IsoKoop, EMAformer and FreqCycle read the phase from the sample index, and IsoKoop, like iTransformer, also takes the timestamp features as extra tokens where a dataset has them.

IsoKoop's per-dataset capacity, trainer (learning rate, batch size, dropout, schedule, patience) and operator variants (coupling, lifting variant, anchor mode) are chosen on the validation split, never on test. They are listed per dataset with the released code. The level anchor is on for Exchange and

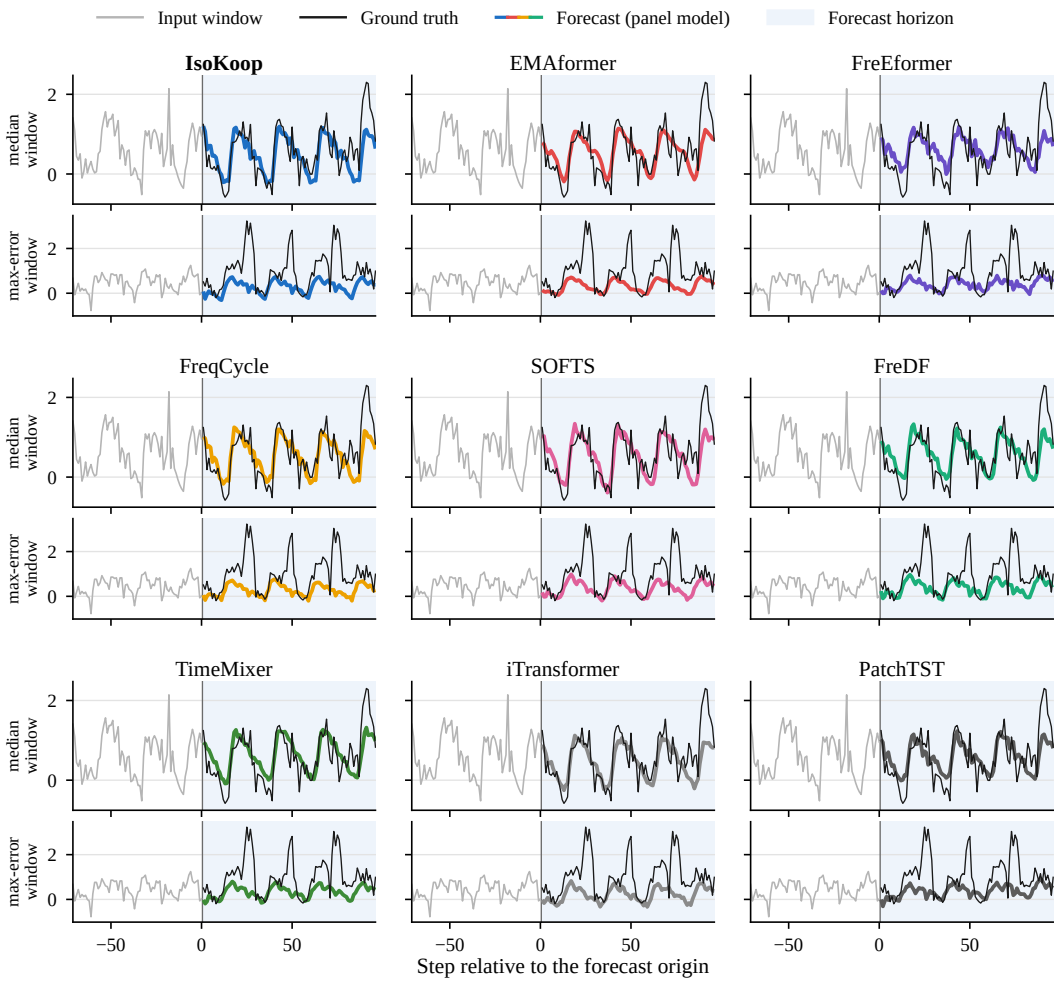


Figure 10: ETTh1, $H{=}96$: on IsoKoop's median-error window all nine models follow the daily shape; on its largest-error window all nine stay below the same sudden rise in level and swing. As at $H{=}336$ (Figure 11), the worst case is shared by every model. Variate HULL; test windows 847 (median error) and 1835 (largest error), selected by rule, not by hand: the variate whose IsoKoop test MSE is the median over the variates, and on it the test windows of median and of largest IsoKoop MSE. Grey: the last 72 of the 168 input steps; black: ground truth; coloured: the panel model's forecast (standardised units; one y range per row of windows). One training run per model with its configuration of Table 1, same order and names; no error value is shown.

Solar, the prototype for Solar, Traffic, TrafficC and the four PEMS panels, and the banded lifting on PEMS03, PEMS04, PEMS07, Traffic and TrafficC. Validation chooses whether the prototype is on and its rank. The placement is fixed in raw units on every dataset. The loss balance a is fixed at 0.85 on Weather. PEMS07 and TrafficC take the designs validated on PEMS04 and Traffic, and PEMS07's capacity is chosen on its own validation split. The comparison models keep their published recipes in Table 1. We hold $40{,}826$ distinct run records (training runs and evaluations), among them the 168 IsoKoop runs of Table 1 (14 datasets, 4 horizons, 3 seeds) and 208 factorial and ablation arms. The same design at the same seed moves within a few percent between campaigns (Solar at $H{=}96$, seed 2027: 0.186 in the factorial, 0.192 in Table 1's runs). Table 11 characterises every panel on its loader's own splits.

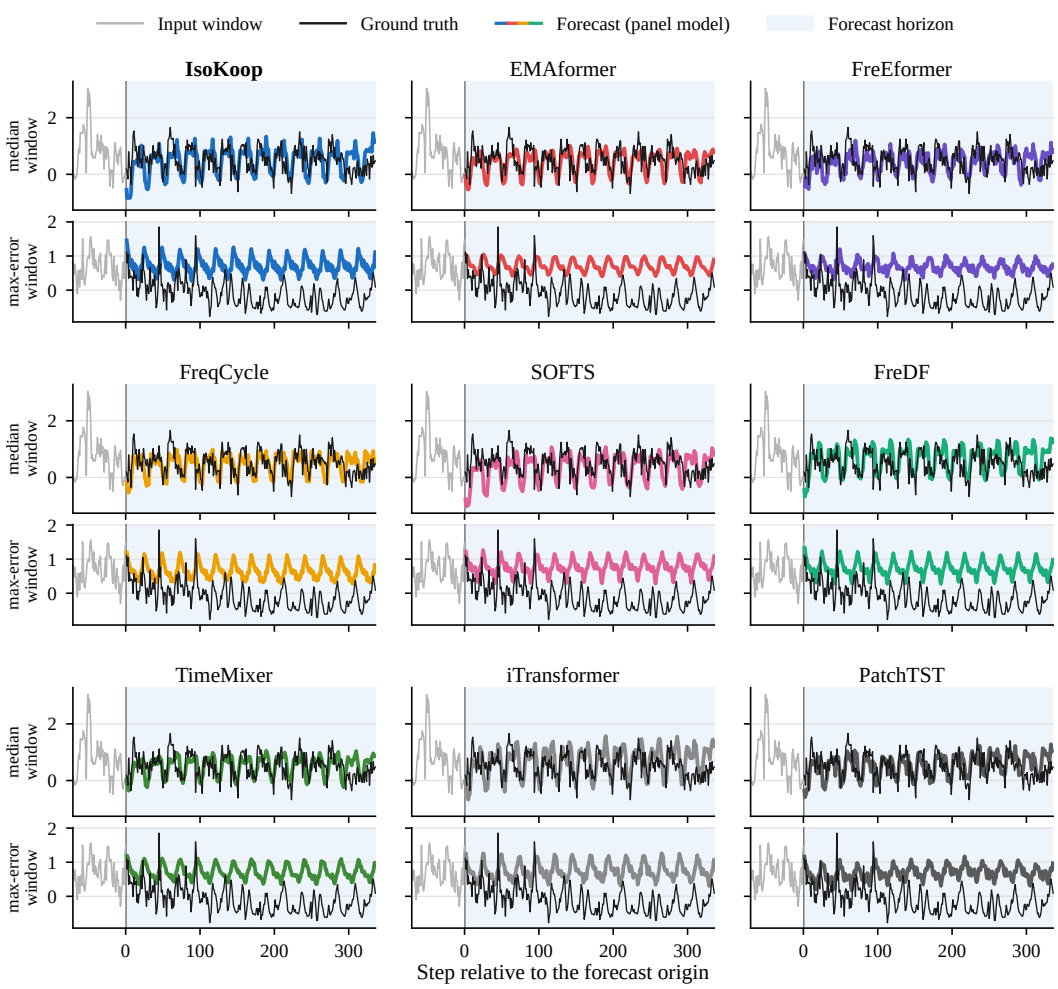


Figure 11: ETTh1, $H{=}336$: on IsoKoop's median-error window all nine models follow the daily cycle; on its largest-error window all nine overshoot the same fall in level. The worst case is shared by every model, not specific to IsoKoop. Variate HULL; test windows 1959 (median error) and 2544 (largest error), selected by rule, not by hand: the variate whose IsoKoop test MSE is the median over the variates, and on it the test windows of median and of largest IsoKoop MSE. Grey: the last 72 of the 168 input steps; black: ground truth; coloured: the panel model's forecast (standardised units; one y range per row of windows). One training run per model with its configuration of Table 1, same order and names; no error value is shown.

H.1 Cost at the reported configuration

Figure 19 places accuracy against training cost on the three datasets IsoKoop leads by the widest margins (PEMS04, Electricity and PEMS07; Section 4.2), and Table 12 gives parameters, peak memory and per-sample time for every model at the configuration of Table 1.

Table 13 gives IsoKoop's rank among the nine models in each of the three costs on every dataset: it is not the cheapest model, with median ranks of seven in parameters, four in peak memory and seven in time per sample.

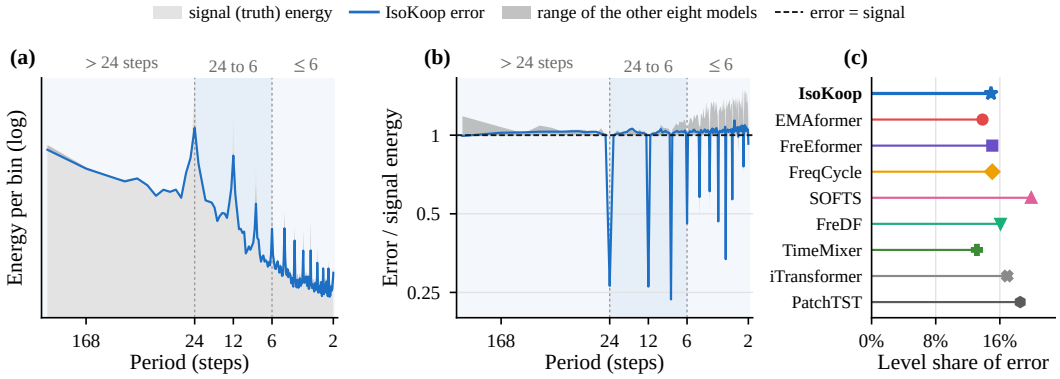


Figure 12: ETTh1, $H{=}336$: the nine models remove error energy only at the daily harmonics; at every other period their error carries as much energy as the signal itself. What separates them is mainly how much of the error is a level offset (c), and the error some add at periods under six steps (b). Error spectrum of the runs of Figure 11, over every test window and variate, per DFT bin of the horizon. **(a)** Energy of the truth (window mean removed) and of the forecast error, log scale. **(b)** Error energy over signal energy; on the dashed line the error carries as much energy as the signal, below it the model has removed some. Blue: IsoKoop; grey band: the range of the other eight models. **(c)** Share of each model's error energy in bin 0, the error's window mean, i.e. a level offset. Shaded: the period bands of Figure 13(d). One training run per model; no error value is quoted.

Table 11: The datasets, measured on each loader's own splits. P is the protocol cycle, fixed per dataset from the sampling rate and not tuned: one day where the panel is sub-daily, one week for the hourly Electricity and Traffic, and Exchange's five-day business week; Forecast. is forecastability, one minus the normalised spectral entropy of the detrended training split (1 a pure tone, 0 white noise); ADF the augmented Dickey–Fuller t-statistic with constant and trend, averaged over variates (more negative, more stationary). *Shift*: the share of variates whose test-split mean or standard deviation, in train-standardised units, falls outside the range that the same statistic takes over sliding training windows of the test split's length (any, the mean alone, the standard deviation alone), and the largest such excursion of the mean and of the standard deviation. The standard-deviation columns are the train-to-test scale shift of Section 4.6.

								shift, % of variates			largest excursion	
Dataset	N	T	interval	P	P/L	Forecast.	ADF	any	mean	s.d.	mean	s.d.
Exchange	8	7588	1 d	5	0.03	0.763	−2.00	75.0	62.5	37.5	1.86	0.13
ETTh1	7	17420	1 h	24	0.14	0.520	−6.17	85.7	28.6	71.4	0.45	0.09
ETTh2	7	17420	1 h	24	0.14	0.563	−5.04	71.4	71.4	14.3	2.57	0.06
ETTm1	7	69680	15 min	96	0.57	0.581	−13.97	85.7	28.6	71.4	0.45	0.10
ETTm2	7	69680	15 min	96	0.57	0.620	−6.85	71.4	71.4	14.3	2.56	0.05
Weather	21	52696	10 min	144	0.86	0.467	−25.96	42.9	38.1	28.6	0.56	0.19
Solar	137	52560	10 min	144	0.86	0.697	−30.61	100.0	100.0	83.9	0.09	0.10
Electricity	321	26304	1 h	168	1.00	0.663	−7.64	40.8	25.5	27.1	1.88	0.42
Traffic	862	17544	1 h	168	1.00	0.512	−13.84	68.9	52.2	56.4	3.94	10.56
TrafficC	853	17544	1 h	168	1.00	0.512	−13.87	68.6	51.8	55.9	1.12	0.86
PEMS03	358	26208	5 min	288	1.71	0.681	−14.76	97.2	91.3	62.3	0.76	0.89
PEMS04	307	16992	5 min	288	1.71	0.660	−11.86	75.2	69.1	36.5	0.82	0.60
PEMS08	170	17856	5 min	288	1.71	0.714	−12.58	62.9	40.0	43.5	1.80	3.52
PEMS07	883	28224	5 min	288	1.71	0.704	−15.65	63.5	35.4	50.1	4.28	1.94

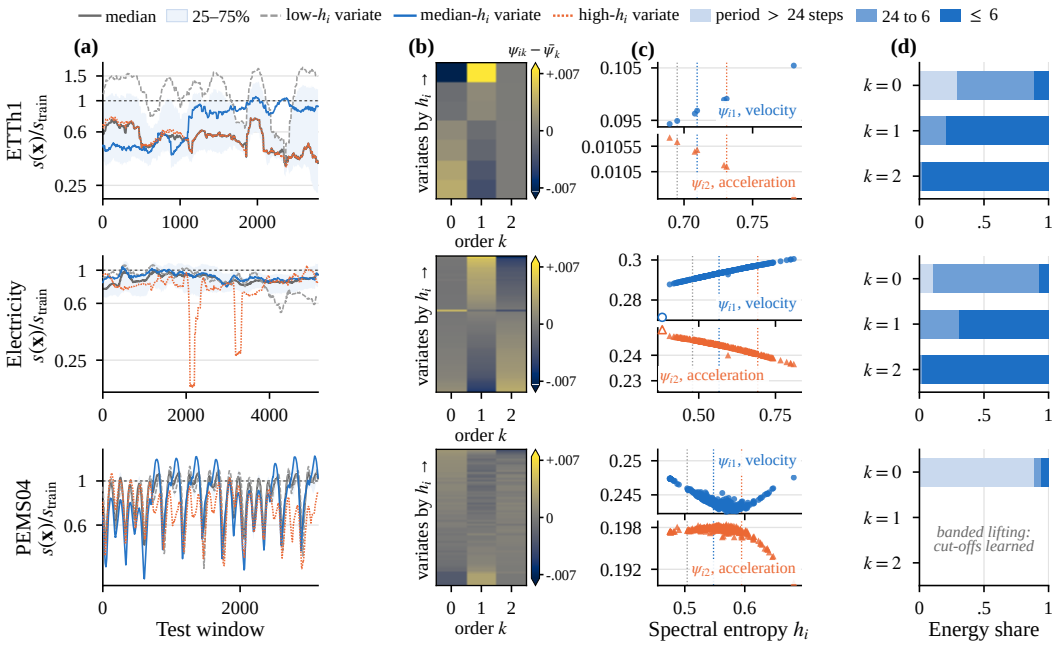


Figure 13: Measured on trained IsoKoop runs, the order read-out ψ_i varies systematically with the spectral entropy h_i, in a direction learned per dataset: on ETTh1 and Electricity a higher h_i gets more velocity and less acceleration weight; on PEMS04 the velocity weight falls with h_i for most variates and rises again for the highest-entropy ones, and the acceleration weight has no monotone trend. On ETTh1 the acceleration weight stays close to its floor ε. Rows: datasets at H=96, one run each with the configuration of Table 1. **(a)** Scale $s(\mathbf{x})$ of each raw test window over the variate's training-split standard deviation: median and 25–75% band over variates, and three variates selected by rule (those at the 10th, 50th and 90th percentile of h_i, nearest rank), marked in (c) by dotted lines of the same colour. **(b)** ψ_{ik} averaged over the test set minus its mean over variates for each order k, variates sorted by h_i; the colour scale saturates at the 98th percentile of the deviation. **(c)** ψ_{i1} and ψ_{i2} against h_i, one point per variate; the correlation is shown, not quoted. One Electricity variate is nearly constant ($h_i \approx 0$) and is drawn open at the left edge. **(d)** Share of order-k energy $|\omega|^{2k}P(\omega)$ in three period bands, P the mean power spectrum of the normalised input windows, before the learned equaliser of equation 2. PEMS04 runs the banded lifting, whose learned cut-offs the run does not save, so (d) shows only k=0; its model subtracts the prototype before normalising, so its own scale differs from the raw-window s of (a).

Table 12: Cost of each model at the configuration of Table 1, on one idle GPU ($L = 168$, $H = 96$): parameters (M), peak memory of a forward and backward pass (GB) and forward-plus-backward time per training sample (ms). Accuracy against cost: Figure 19.

	Electricity			PEMS04			Traffic			PEMS07		
Model	par.	GB	ms	par.	GB	ms	par.	GB	ms	par.	GB	ms
IsoKoop	19.53	2.50	8.45	19.52	2.44	8.35	7.26	3.97	11.21	20.12	6.52	27.21
iTransformer	4.87	0.83	1.60	25.48	2.96	5.10	6.45	5.00	8.86	25.48	12.54	17.84
PatchTST	7.35	8.92	65.80	7.35	8.54	62.56	4.20	3.77	60.63	7.35	24.25	199.86
FreDF	4.87	0.83	1.60	25.48	2.96	5.08	6.45	5.00	8.85	25.48	12.54	17.89
TimeMixer	0.29	7.22	17.95	1.22	0.47	0.68	0.30	9.60	68.28	1.52	0.67	0.81
EMAformer	32.87	0.85	1.29	108.51	2.49	2.89	79.68	2.21	3.71	278.97	7.21	8.64
SOFTS	4.48	0.52	1.14	9.60	1.50	1.76	4.48	1.24	2.48	9.60	3.92	4.72
FreEformer	14.30	4.01	4.69	14.25	3.78	4.51	18.14	10.17	20.30	18.36	20.94	20.46
FreqCycle	0.19	2.98	0.39	0.23	2.85	0.37	0.28	7.88	1.05	0.39	8.07	1.08

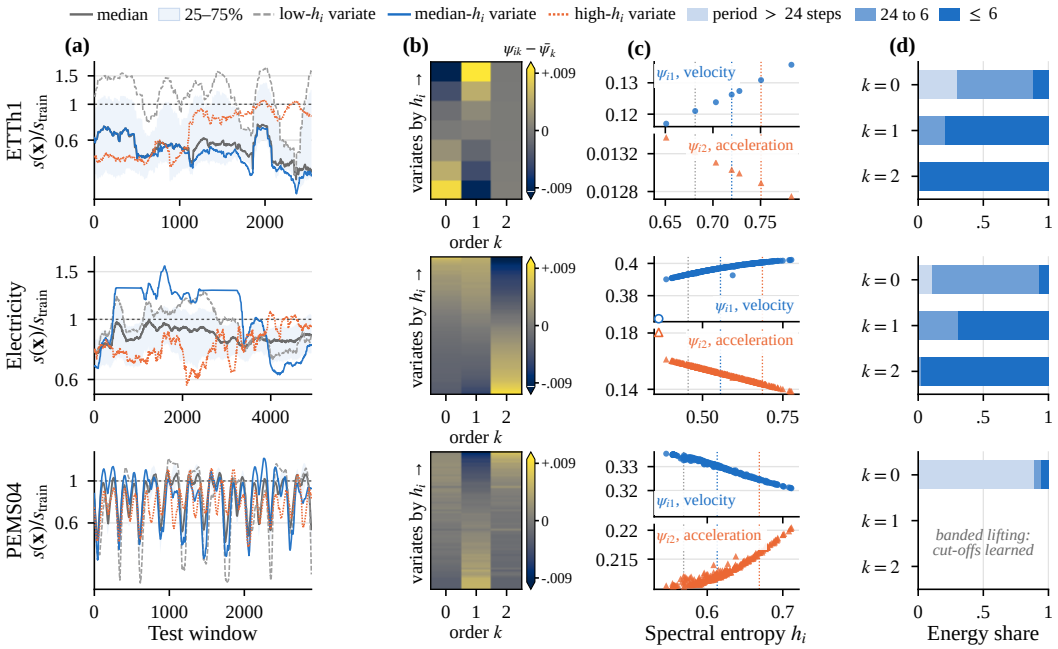


Figure 14: As at H=96 (Figure 13), the order read-out ψ_i varies systematically with the spectral entropy h_i, in a direction learned per dataset: on ETTh1 and Electricity a higher h_i gets more velocity and less acceleration weight; on PEMS04 the reverse, less velocity and more acceleration weight. On ETTh1 the acceleration weight stays close to its floor ε. Rows: datasets at H=336, one run each with the configuration of Table 1. **(a)** Scale $s(\mathbf{x})$ of each raw test window over the variate's training-split standard deviation: median and 25–75% band over variates, and three variates selected by rule (those at the 10th, 50th and 90th percentile of h_i, nearest rank), marked in (c) by dotted lines of the same colour. **(b)** ψ_{ik} averaged over the test set minus its mean over variates for each order k, variates sorted by h_i; the colour scale saturates at the 98th percentile of the deviation. **(c)** ψ_{i1} and ψ_{i2} against h_i, one point per variate; the correlation is shown, not quoted. One Electricity variate is nearly constant ($h_i \approx 0$) and is drawn open at the left edge. **(d)** Share of order-k energy $|\omega|^{2k} P(\omega)$ in three period bands, P the mean power spectrum of the normalised input windows, before the learned equaliser of equation 2. PEMS04 runs the banded lifting, whose learned cut-offs the run does not save, so (d) shows only k=0; its model subtracts the prototype before normalising, so its own scale differs from the raw-window s of (a).

Table 13: IsoKoop's cost rank among the nine models on each dataset (rank 1 = cheapest), measured on one idle GPU with $L = 168$, $H = 96$, each model at the configuration of Table 1: parameters, peak memory of a forward and backward pass, and forward-plus-backward time per training sample (batch time divided by the recipe's batch size, as in Table 12). Ranks use the unrounded measurements. Last row: median over the 14 datasets (min–max).

Dataset	Params	Peak mem.	ms / sample
ETTh1	4	2	7
ETTh2	7	6	7
ETTm1	1	2	7
ETTm2	1	1	7
Weather	9	7	7
Solar	9	6	8
Electricity	8	5	7
Traffic	7	4	6
Exchange	8	6	7
PEMS03	8	6	8
PEMS04	6	3	8
PEMS08	3	1	4
TrafficC	7	4	6
PEMS07	6	3	8
Median (range)	7 (1–9)	4 (1–7)	7 (4–8)

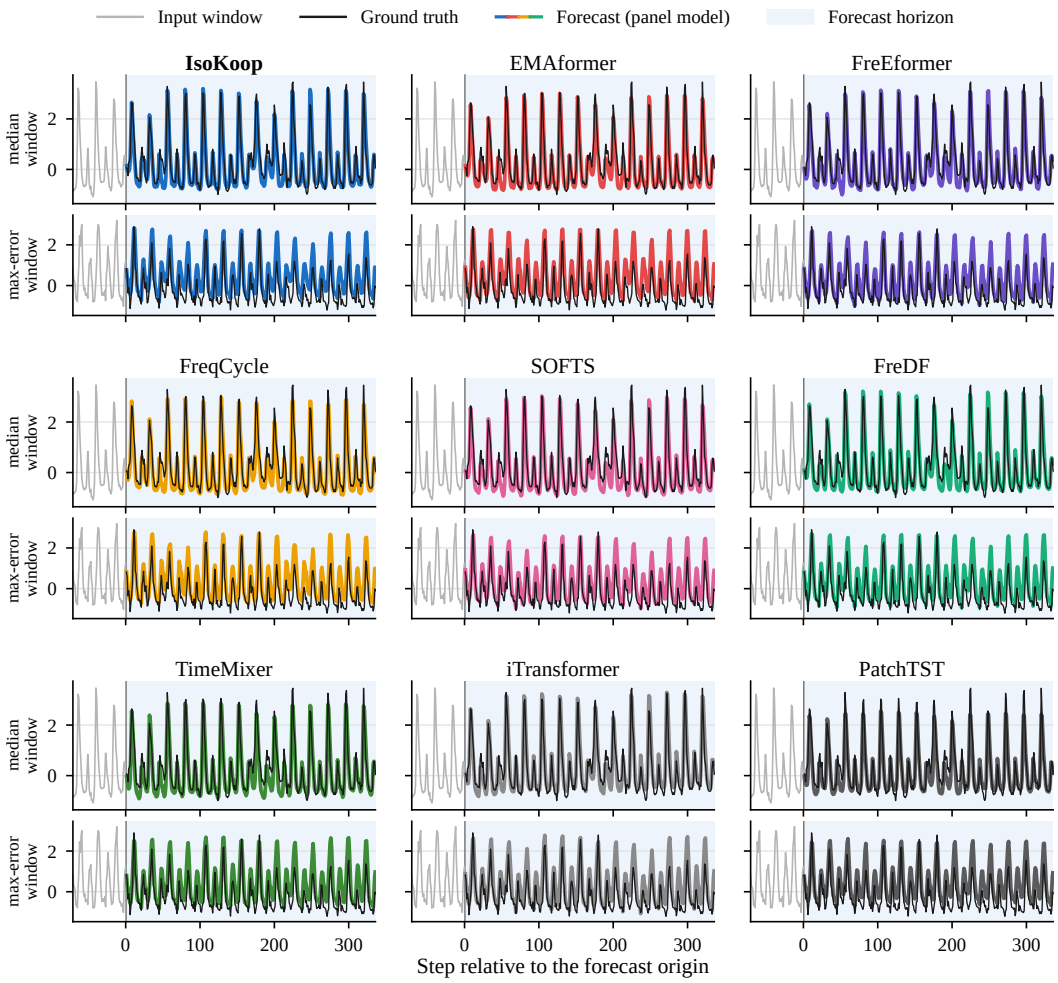


Figure 15: Electricity, $H{=}336$: on IsoKoop's median-error window all nine models follow the daily cycle; on its largest-error window all nine keep their daily peaks near the input's height while most of the true peaks fall well below it, and every forecast lies above the truth almost throughout. As on ETTh1 (Figure 11), the worst case is shared by every model. Variate 44; test windows 4648 (median error) and 1574 (largest error), selected by rule, not by hand: the variate whose IsoKoop test MSE is the median over the variates, and on it the test windows of median and of largest IsoKoop MSE. Grey: the last 72 of the 168 input steps; black: ground truth; coloured: the panel model's forecast (standardised units; one y range per row of windows). One training run per model with its configuration of Table 1, same order and names; no error value is shown.

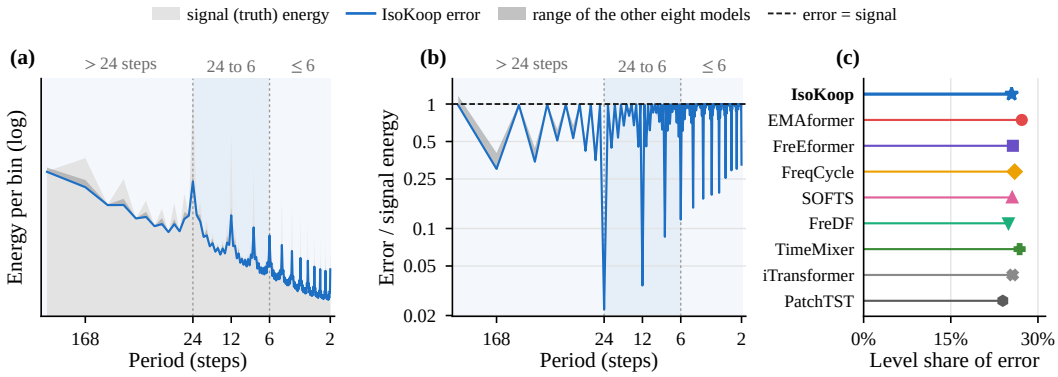


Figure 16: Electricity, H=336: the nine models remove nearly all error energy at the daily harmonics and part of it at many other harmonics of the week; at the remaining periods their error carries about as much energy as the signal. Unlike on ETTh1 (Figure 12), the nine level shares (c) lie close together. Error spectrum of the runs of Figure 15, over every test window and variate, per DFT bin of the horizon. **(a)** Energy of the truth (window mean removed) and of the forecast error, log scale. **(b)** Error energy over signal energy; on the dashed line the error carries as much energy as the signal, below it the model has removed some. Blue: IsoKoop; grey band: the range of the other eight models. **(c)** Share of each model's error energy in bin 0, the error's window mean, i.e. a level offset. Shaded: the period bands of Figure 13(d). One training run per model; no error value is quoted.

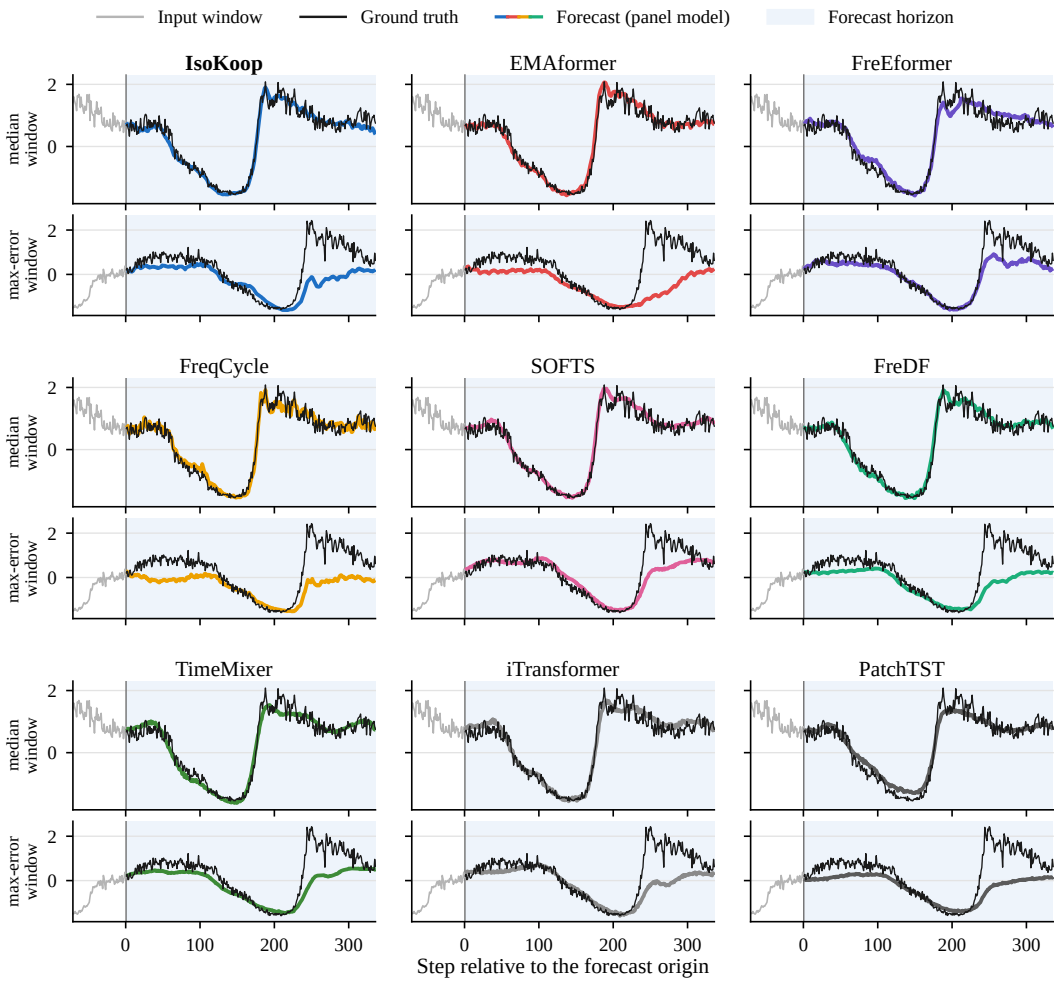


Figure 17: PEMS04, $H{=}336$: on IsoKoop's median-error window all nine models follow the fall into the trough and the sharp rise that ends it; on its largest-error window all nine reach the trough but none reaches the level the truth jumps to after it, and every forecast stays below the truth for most of the remaining horizon. As on ETTh1 and Electricity (Figures 11 and 15), the worst case is shared by every model. Variate sensor 284; test windows 1379 (median error) and 452 (largest error), selected by rule, not by hand: the variate whose IsoKoop test MSE is the median over the variates, and on it the test windows of median and of largest IsoKoop MSE. Grey: the last 72 of the 168 input steps; black: ground truth; coloured: the panel model's forecast (standardised units; one y range per row of windows). One training run per model with its configuration of Table 1, same order and names; no error value is shown.

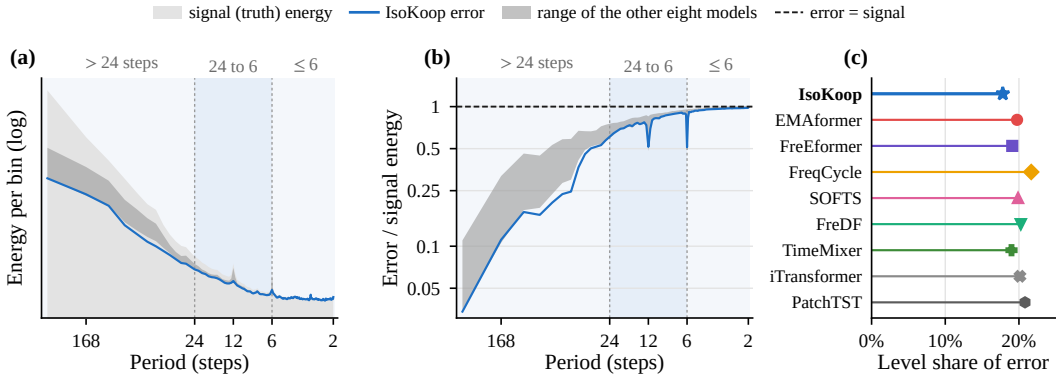


Figure 18: PEMS04, H=336: the nine models remove most of the error energy at the longest periods and less and less toward shorter ones; at periods of six steps and under their error carries about as much energy as the signal. Unlike on ETTh1 and Electricity (Figures 12 and 16), (b) has no deep dips at isolated harmonics; the shallow dips at periods of one hour and half an hour appear for only some of the models. Error spectrum of the runs of Figure 17, over every test window and variate, per DFT bin of the horizon. **(a)** Energy of the truth (window mean removed) and of the forecast error, log scale. **(b)** Error energy over signal energy; on the dashed line the error carries as much energy as the signal, below it the model has removed some. Blue: IsoKoop; grey band: the range of the other eight models. **(c)** Share of each model's error energy in bin 0, the error's window mean, i.e. a level offset. Shaded: the period bands of Figure 13(d). One training run per model; no error value is quoted.

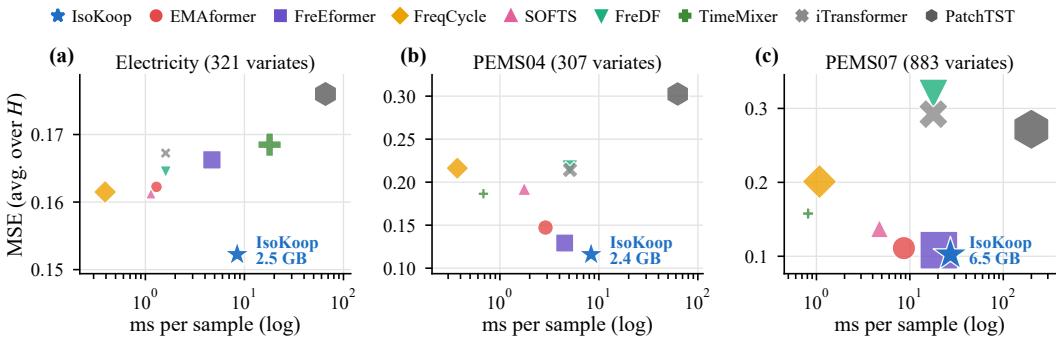


Figure 19: Accuracy against cost on the three datasets IsoKoop leads by the widest margins: test MSE (average over the four horizons) against training time per sample; marker area is proportional to peak memory. IsoKoop (★) is the most accurate model on each; per training sample only PatchTST (and, on Electricity, TimeMixer) is slower.

I NOTATION, ASSUMPTIONS AND THE BANDED VARIANT

Notation. D^k is the spectral derivative: on a window $x \in \mathbb{R}^L$ it multiplies DFT bin m by $(j\omega_m)^k$, $\omega_m = 2\pi m/L$. Δ is the *circular* forward difference on a horizon, $(\Delta e)_t = e_{(t+1) \bmod H} - e_t$; the implemented time term of equation 4 uses the non-circular forward difference truncated to its valid length, which is where it departs from the objective remark (Design remark 9). $\mathbf{V} \in \mathbb{R}^{N\times R}$ is the prototype's variate coordinate and $\mathbf{V}_i^{\mathrm{eq}} \in \mathbb{R}^r$ the equaliser's per-variate fingerprint. They are separate parameters. $\mathbf{q}_h \in \mathbb{R}^{L\times N}$ and $\mathbf{q}_f \in \mathbb{R}^{H\times N}$ are the history and horizon parts of a subtracted-then-restored profile. $\mathbf{1}$ is an all-ones vector of the length the context requires.

One name per object. The *prototype* $\mathbf{Q} = \mathbf{U}\mathbf{V}^\top$ is the one calendar-indexed component; in Proposition 1 $\mathbf{q}_h, \mathbf{q}_f$ stand for any subtracted profile. The *level anchor* has gain a_i. The loss weight $a = \sigma(\alpha_{\mathrm{freq}})$ is unrelated. The *gated diagonal transport* equation 3 has gate $A(\mathbf{z})$, mixing matrix C and residual weight α. g is only the gain of the equivariance test. ψ_i are the read-out weights, $\nu^{\mathrm{time}}, \nu^{\mathrm{freq}}$ the objective's order weights, β_k the equaliser's per-order frequency profile, $\tilde{\Delta}$ the jump statistic and Δ the forward difference. *Scale equivariance* is the property of the forecaster, $f(g\mathbf{x}) = gf(\mathbf{x})$; the normalised input is *invariant*. Symbols defined inside an assumption, proof or remark (the matrices A, B of A3 and A of A7, M, $\mathcal{C}$, c_i, κ, Λ) are local to it. Two symbols carry a second, unrelated scope: m is the DFT bin index of the circular-derivative statement and the whitening multiplier of the equaliser's, and c_i is the power-law scale of A5 and a centroid offset in the mean-field remark; neither pair appears in the same statement. Also, s_i is the window scale $s_i(\mathbf{x})$ outside A5 and the spectral exponent inside it and in Appendix J.3; ρ is the AR(1) coefficient except in the objective remark, where $\rho(\omega) = e^{i\omega} - 1$; and c without the index i is the constant level of Proposition 1, the mean-field scale, a band cut-off c_j or a circulant entry c_r, as the context fixes.

Assumptions. The classical statements of Appendix J and the design remarks use eight hypotheses. Proposition 1 uses none. Seven are idealisations that no dataset satisfies exactly. A2, as stated, holds for every window of even length, and its content is how $\tilde{\Delta}$ is read. They are hypotheses of the statements, never assertions about the data.

Assumption A1 (AR(1) level)**.** *$x_t = \rho x_{t-1} + \varepsilon_t$ for $t \geq 1$, with $\rho \in [0, 1]$, $(\varepsilon_t)_{t\geq 1}$ i.i.d. with mean zero and variance σ^2, and x_0 independent of (ε_t) with $\mathbb{E}x_0^2 < \infty$; $\mathcal{F}_L = \sigma(x_0, \ldots, x_L)$.* Used by *the level statement (Statement 1). No dataset is AR(1); the assumption fixes the anchor's functional form and nothing about any dataset.*

Assumption A2 (Band-limited window)**.** *$x[n] = u(n) + \tilde{\Delta}\, n/L$ for $n = 0, \ldots, L-1$, where u is a real trigonometric polynomial of period L and degree $M < L/2$.* Used by *the circular-derivative statement (Statement 2) and part (iii) of the jet remark (Design remark 6). As stated, every sampled window of even length satisfies it, with $\tilde{\Delta} = -2\sum_n (-1)^n x[n]$: the ramp is the only term with a Nyquist coefficient, and a sequence without one samples a unique trigonometric polynomial of degree $< L/2$. The circular-derivative statement is therefore an identity on every window. The assumption's content is the reading of $\tilde{\Delta}$, which equals the jump $x(L) - x(0)$ of the periodic extension when the underlying signal is band-limited well below Nyquist, and is otherwise a Nyquist-band statistic (Remark 1).*

Assumption A3 (Second-order linear transverse flow)**.** *The transverse state obeys $\ddot{\mathbf{x}} = A\dot{\mathbf{x}} + B\mathbf{x}$ with constant $A, B \in \mathbb{R}^{N\times N}$.* Used by *the jet remark (Design remark 6). No dataset obeys it, and the transport equation 3 is not linear in any case. Near an attracting periodic orbit the transverse dynamics are periodic in time (Floquet), so constant A, B is at best a description in phase-removed coordinates; the phase code is how the model is allowed a phase-dependent operator.*

Assumption A4 (Stationarity on a band)**.** *Each variate X_i is weakly stationary with a spectral density $S_i > 0$ on a band $\Omega = [\omega_{\min}, \omega_{\max}]$, $0 < \omega_{\min} < \omega_{\max}$, and a multiplier acts on it as a linear time-invariant filter.* Used by *the whitening statement (Statement 3). The band restriction is supplied by the DFT grid ($\omega_{\min} = 2\pi/L$, the mean removed by the normaliser); stationarity is not tested. On a finite window a DFT mask is a circular filter, not a time-invariant one, so the statement describes the mask only approximately.*

Assumption A5 (Power-law spectra)**.** *$S_i(\omega) = c_i|\omega|^{-2s_i}$ on Ω with $c_i > 0$.* Used by *parts (i), (iii), (iv) and the slope form of part (ii) of the whitening statement. The power law itself is not tested; the per-variate slopes fitted on each training split (Appendix J.3) spread materially only on Weather.*

Assumption A6 (Label model). $(\mathbf{x}, y)$ *have finite second moments and* $y = \mu(\mathbf{x}) + \eta \in \mathbb{R}^H$ *with* $\mathbb{E}[\eta \mid \mathbf{x}] = 0$ *and* $\mathrm{Cov}(\eta \mid \mathbf{x}) = \Sigma \succ 0$ *constant in* $\mathbf{x}$ *(one variate).* Used by *the frequency-term remark (Design remark 10). The horizon-label covariance is not measured.*

Assumption A7 (Linear forecaster class). *The forecaster class is* $\{A\beta : \beta \in \mathbb{R}^p\}$ *with* $A \in \mathbb{R}^{H\times p}$ *of full column rank,* $p < H$. Used by *parts (iii)–(iv) of the frequency-term remark, which are illustrative: our class is not linear.*

Assumption A8 (Circular stationarity). Σ *is circulant,* $\Sigma_{ts} = c_{(t-s) \bmod H}$ *with* $c_r = c_{H-r}$. Used by *part (v) of the frequency-term remark. An ordinary stationary process has a Toeplitz* Σ*, for which the statement holds only asymptotically in* H.

Three further facts come from the configuration: every reported cell runs the equaliser, so part (iii) of the jet remark describes the un-equalised jet only (so does the fixed profile e_k of the circular-derivative statement, which holds for D^k itself, not for the equalised channel, whose error profile depends on the learned mask and the variate). Five datasets run the banded operator variant, which forms no $(j\omega)^k$ factor and is outside the circular-derivative statement, the jet remark and part (iv) of the whitening statement. The read-out gate reads only the spectral entropy, not the transported tokens, which the read-out remark (Design remark 8) needs and the code satisfies by construction.

Which rows each statement covers. Part (i) of Proposition 1 covers the seven datasets that run the prototype off (the four ETT panels, Weather, Electricity and Exchange), and part (ii) the seven that run it in raw units, Solar through the extension of Appendix C. The level statement concerns the two datasets with the anchor on, Exchange and Solar, neither of them AR(1). The circular-derivative statement and part (iii) of the jet remark describe the un-equalised operator, which no reported cell runs, and the five banded datasets are outside them. The whitening statement's premise, differing spectral slopes, holds materially only on Weather. No dataset obeys the jet remark's hypothesis A3. The mean-field remark and parts (iii)–(v) of the frequency-term remark describe linear classes, which IsoKoop is not; and the objective remark is an identity for a squared seminorm, which equation 4 applies through $\log(1 + \cdot)$. The read-out floor holds on every row, by construction. The statements motivate the modules; Sections 4.4 and 4.5 measure what the modules do.

The banded variant. Selected on PEMS03, PEMS04, PEMS07, Traffic and TrafficC (the spectral derivative of equation 2 is replaced). Channel k of variate i is $\mathrm{IFFT}[\, b_{k,i}(\omega)\, \sigma(\beta_0(\omega) + \mathbf{V}_i^{\mathrm{eq}\top} W_0)\, \mathrm{FFT}(\mathbf{x}_i)]$, with $b_{0,i} = 1 - u_{1,i}$, $b_{1,i} = u_{1,i}(1 - u_{2,i})$, $b_{2,i} = u_{1,i}u_{2,i}$ and $u_{j,i}(\omega) = \mathrm{sigmoid}\big((\omega/\pi - c_j - \frac{1}{4}\tanh\delta_{ij})/\tau_j\big)$: ordered learned cut-offs $c_1 < c_2$ with a per-variate shift δ_{ij} and learned widths τ_j. The $b_{k,i}$ sum to one at every frequency, so the three channels are a low, a middle and a high band of the equalised window and sum back to it. The mask is shared across orders (one β_0, one W_0), and no $(j\omega)^k$ factor, hence no derivative, is formed. As for the derivative jet equation 2, the three channels are divided by the norm of channel 0.

What reading in the coordinates fixes. Between $\mathcal{N}$ and $\mathcal{N}^{-1}$ the jet, the codes, the transport, the read-out and the anchor see only $\tilde{\mathbf{x}}$ and calendar inputs, none of which depends on the window's values, so together they form the inner network F of Proposition 1, which holds for every value of their parameters. Reading $\tilde{\mathbf{x}}$ instead of $\mathbf{x}$ also fixes what each of them can express. For a fixed equaliser mask, T_k in equation 2 is linear and annihilates a constant for $k \geq 1$, so

$$T_k\big(\mathcal{N}_{\mathbf{x}}(\mathbf{x})_i\big) = \frac{T_k \mathbf{x}_i}{s_i(\mathbf{x})}, \qquad k \geq 1 : \tag{5}$$

the derivative channels in the coordinates are the raw-unit ones divided by the window's scale, and dividing by channel 0's norm removes a gain in either units. They differ in the level, since channel 0 has an empty DC bin on $\tilde{\mathbf{x}}$ and passes $\mu_i(\mathbf{x})$ through the mask value on a raw window. The gate and the mean field both read the tokens' scale, so on a raw window a level far from zero shrinks a variate's derivative channels toward zero in what they see, whereas in the coordinates the scale of a variate's tokens is set by the shape of its variation. The read-out's entropy is likewise destroyed by a level on a raw window and describes how the variation spreads over frequency on $\tilde{\mathbf{x}}$. De-normalised, the anchor is

$$s_i(\mathbf{x})\, a_i \hat{\rho}_i^{\,h} \tilde{x}_{L,i} + \mu_i(\mathbf{x}) = \mu_i(\mathbf{x}) + a_i \hat{\rho}_i^{\,h} \big(x_{L,i} - \mu_i(\mathbf{x})\big), \tag{6}$$

an AR(1) mean (Statement 1) that reverts to the window mean at rate $\hat{\rho}_i$ with gain a_i, and is the last value $x_{L,i}$ at every h when $\hat{\rho}_i = a_i = 1$. The same form on the raw window reverts to the origin

Table 14: Where this paper's claims are stated and cited. **(a)** Every statement and design remark of the appendix, the entry that states or proves it, and the sections of the main text that cite it. **(b)** The mechanism readings quoted in the main text that are printed in no table, with the runs behind them. Section numbers refer to the main text.

(a) Statements and design remarks

Claim	What it says	Appendix	Cited from
Proposition 1	Where out-of-window knowledge may enter	C (proof)	1, 2, 3.1–3.3, conclusion
Corollary 1	The other three placements	C (proof)	appendix only
Statement 1	The level: the window contributes nothing beyond its last value	J	3.2
Statement 2	Circular-derivative error, an identity	J	3.1, 3.2
Statement 3	Per-variate whitening	J	3.2
Lemma 1	Kernel of D^1	J	appendix only
Remark 1	What the residual measures on a real window	J	appendix only
Design remark 4	The period is not a window feature	K	3.2
Design remark 5	A window that does not close the cycle does not fix the cycle length	K	1, 3.1, 3.2
Design remark 6	Koopman-invariant lifting	K	3.1, 3.2
Design remark 7	Centred coupling	K	3.1, 3.3
Design remark 8	Simplex floor	K	3.3
Design remark 9	One Sobolev seminorm, two domains	K	3.1, 3.4
Design remark 10	What the frequency domain buys	K	appendix only
Assumptions A1–A8	The conditions the claims above use	I	3.2

(b) Readings that appear in no table

Reading	In the text	Quoted in	Where it comes from
Structural codes	carry 21.3% on PEMS04 and 11.9% on PEMS03	3.2	single-arm ablation, seed 2027, H=96
Solar's anchor gain	0.054 at H=96, falling to 0.027 at H=720	3.2	the fitted anchor gain, seed 2027
Anchor removal	costs 4.9% on Solar and 8.5% on Exchange	3.2	single-arm ablation, seed 2027, H=96
Exchange's anchor	its last-value mode, a_i=0.99, $\hat{\rho}_i$=1	3.2	the fitted anchor, seed 2027
Mean field	removing it costs 49.6% on PEMS04, 16.9% on PEMS03 and 6.3% on Electricity, a tie on ETTh1 and Solar	3.3	single-arm ablation, seed 2027, H=96
Loss balance a	initial 0.54, settling between 0.69 and 0.999	3.4	the learned balance over the seed-2027 runs
Balance frozen at 0.85	the error moves by −3.1% to +0.8%	3.4	the frozen-balance probe, seed 2027, H=96

of the data's units instead, and equation 6 is equivariant to gains and offsets for every a_i, as $\hat{\rho}_i$ is invariant to both. The banded variant forms no $(j\omega)^k$ factor and is outside equation 5.

Table 14 registers the claims below: the entry that states or proves each, the main-text sections that cite it, and the mechanism readings quoted in Section 3 that are printed in no table.

J THREE CLASSICAL STATEMENTS: THE LEVEL, THE CIRCULAR DERIVATIVE AND WHITENING

Three classical statements shaped IsoKoop's modules: the level anchor, the spectral jet and the equaliser.

The chain in one paragraph. For a band-limited window (A2) the circular derivative's error is exactly $\tilde{\Delta}\, e_k$, linear in the statistic $\tilde{\Delta} = -2\sum_n (-1)^n x[n]$, which for a signal band-limited below Nyquist is the periodic-extension jump $x(L) - x(0)$ (not $x[L-1] - x[0]$), with a profile e_k that depends on L and k only and, for $k = 1$, has edge value $-\ln 2 + O(1/L)$ per unit jump (Statement 2). If variates have power-law spectra with different exponents (A4, A5), no shared multiplier whitens them on a band bounded away from zero (Statement 3). Under a zero-mean AR(1) level (A1) the window's risk-minimising forecast is $\rho^h x_L$ (Statement 1).

J.1 THE LEVEL

Statement 1 (The level: the window contributes nothing beyond its last value)**.** *Under A1, for every $h \geq 1$ and every square-integrable $\mathcal{F}_L$-measurable $\hat{Z}$,*

$$\mathbb{E}\big(x_{L+h} - \hat{Z}\big)^2 = \sigma^2\, \frac{1-\rho^{2h}}{1-\rho^2} + \mathbb{E}\big(\rho^h x_L - \hat{Z}\big)^2,$$

where the fraction is read as h at $\rho = 1$. The minimum over all window functionals is attained uniquely (almost surely) at $\hat{Z} = \rho^h x_L$, which is $\mathbb{E}[x_{L+h} \mid \mathcal{F}_L]$.

This is the textbook conditional mean of an AR(1) process (Hyndman & Athanasopoulos, 2021), stated in the form the anchor uses.

Proof. Iterating the recursion, $x_{L+h} = \rho^h x_L + \xi$ with $\xi = \sum_{j=1}^{h} \rho^{h-j} \varepsilon_{L+j}$. Since $\mathcal{F}_L \subseteq \sigma(x_0, \varepsilon_1, \ldots, \varepsilon_L)$ and x_0 is independent of the innovations, ξ is independent of $\mathcal{F}_L$ with mean zero and variance $\sigma^2 \sum_{j=1}^{h} \rho^{2(h-j)}$, which is the stated fraction. $\mathbb{E}x_L^2 < \infty$ because $\mathbb{E}x_0^2 < \infty$ and the innovations have finite variance, so $\rho^h x_L - \hat{Z} \in L^2$ and $\mathbb{E}[(\rho^h x_L - \hat{Z})\xi] = \mathbb{E}[\rho^h x_L - \hat{Z}]\,\mathbb{E}\xi = 0$. Expanding $(x_{L+h} - \hat{Z})^2 = ((\rho^h x_L - \hat{Z}) + \xi)^2$ gives the identity; the second term vanishes iff $\hat{Z} = \rho^h x_L$ almost surely. The argument holds for any real ρ; $[0, 1]$ matches the clip below. □

Four departures of the implementation. The anchor computes $a_i \hat{\rho}_i^h \tilde{x}_{L,i}$ on the normalised window $\tilde{\mathbf{x}} = \mathcal{N}_{\mathbf{x}}(\mathbf{x})$ and is de-normalised with the rest of the forecast. (1) The gain $a_i = \tanh(\cdot) \in (-1, 1)$ is learned (on Exchange the selected design fixes it at 0.99), where the statement's coefficient is one. A gain below one is shrinkage, an estimation choice rather than a consequence. (2) $\hat{\rho}_i$ replaces ρ: the window's lag-1 least-squares coefficient of the demeaned series is biased low by $O(1/L)$, and at $\rho = 1$ this is the unit-root (Dickey–Fuller) bias of the demeaned estimator, present before any clipping, so $\hat{\rho}^{\,h}$ under-persists at long horizons. (3) The clip to $[0, 1]$ excludes $\rho < 0$, whose conditional mean alternates in sign. The upper clip is not inert either, since on a unit-root window the unclipped ratio exceeds one on about 4% of windows at $L = 168$, contributing a further 0.0002 of downward bias in the same direction as (2). (4) Because the anchor acts in normalised units, its level component is $\hat{\mu}_i + a_i \hat{\rho}_i^h (x_{L,i} - \hat{\mu}_i)$: an AR(1) about the *window* mean, which coincides with the statement's $\rho^h x_L$ only at $\rho = 1$, $a_i = 1$, or on a window whose mean vanishes (a null event for a continuously distributed window). For $\rho < 1$ it substitutes the window mean for the process mean. On Exchange the validated design runs the anchor with ρ fixed at one (mode `last`), which removes departures (2)–(3) exactly and reduces (4) to its residual: with $a_i < 1$ the window-mean term survives with weight $1 - a_i$. At $a_i = 1$ that mode is the last-value normalisation of NLinear (Zeng et al., 2023). Two further implemented behaviours are selected by no reported cell: a `mix` mode that interpolates $\hat{\rho}_i$ toward one with a learned share, which is precisely a repair for (2), and the default initialisation of the gain at zero, at which the anchor contributes nothing until training moves it. The same design also subtracts $a_i \tilde{x}_{L,i}$ from the window the jet reads, so the operator sees the window re-centred on its last value and the anchor restores that level on the forecast.

Measurement. The naive last-value forecast, the statement's minimiser at $\rho = 1$, $a = 1$, is below every model at every Exchange horizon (Table 5). The anchor tends to it as $a \to 1$, and the trained model, with $a = 0.99$, stays above it.

J.2 THE CIRCULAR DERIVATIVE'S ERROR

Statement 2 (Circular-derivative error, an identity)**.** *Let $L \geq 4$ be even and let D^k act on $x \in \mathbb{R}^L$ by multiplying the one-sided DFT bin $m \in \{0, \dots, L/2\}$ by $(j\omega_m)^k$ and inverting under Hermitian symmetry, as* `irfft` *does. Write x as in A2, which every window of even length admits, and take as reference derivative $x^{(k)}[n] = u^{(k)}(n) + r^{(k)}(n)$ with $r(t) = \tilde{\Delta}t/L$. Then, on every window: (i) $D^k u = u^{(k)}$ on the grid, for every $k \geq 0$; (ii) $D^k x - x^{(k)} = \tilde{\Delta}\, e_k$ with $e_k \in \mathbb{R}^L$ depending on L and k only; (iii) with $d_m = (-1)^m(\pi/L)\cot(\pi m/L)$ for $1 \leq m \leq L-1$ and $S(n) = \sum_{m=n+1}^{L-1} d_m$, $e_1[n] = S(n) - \frac{1}{L}\big(1 + \sum_{m=1}^{L-1} m\, d_m\big)$; (iv) $e_1[0] = e_1[L-1] = -\ln 2 + O(1/L)$; (v) $\sum_n e_1[n] = -1$, the discrete counterpart of the Dirac mass $-\tilde{\Delta}$ that the derivative of the discontinuous periodic extension carries.*

Lemma 1 (Kernel of D^1)**.** *D^1 is circular convolution with d ($d_0 = 0$), the periodic spectral-differentiation matrix (Trefethen, 2000, ch. 3) in sample units: $(D^1 x)[n] = \sum_{m=0}^{L-1} d_m\, x[(n-m) \bmod L]$. Moreover $\sum_m d_m = 0$ and $d_{L-m} = -d_m$. The null space of D^1 is spanned by $\mathbf{1}$ and $((-1)^n)_n$, and that of D^2 by $\mathbf{1}$.*

Proof. D^1 is diagonal in the DFT basis, hence circulant. On the full DFT its multiplier is $2\pi i\kappa/L$ on bin κ for $|\kappa| < L/2$ and 0 on the Nyquist bin, where the real coefficient times $j\pi$ is imaginary and `irfft` keeps the real part. The kernel is the inverse DFT of this multiplier: for $m \neq 0$, $d_m = -\frac{4\pi}{L^2}\sum_{\kappa=1}^{L/2-1} \kappa \sin(2\pi\kappa m/L)$. Differentiating $\sum_{\kappa=1}^{n} \cos\kappa\theta = \frac{\sin((n+\frac{1}{2})\theta)}{2\sin(\theta/2)} - \frac{1}{2}$ gives $\sum_{\kappa=1}^{n} \kappa \sin\kappa\theta = \frac{(n+1)\sin n\theta - n\sin(n+1)\theta}{4\sin^2(\theta/2)}$; at $n = L/2-1$, $\theta = 2\pi m/L$ one has $\sin((n+1)\theta) = \sin\pi m = 0$ and $\sin n\theta = -(-1)^m \sin\theta$, so the sum is $-\frac{L}{4}(-1)^m\cot(\pi m/L)$ and $d_m = (-1)^m\frac{\pi}{L}\cot\frac{\pi m}{L}$. $\sum_m d_m$ is the multiplier at DC, which is 0; $d_{L-m} = -d_m$ because $\cot(\pi - x) = -\cot x$ and L is even. The multiplier of D^1 vanishes exactly on bin 0 and the Nyquist bin, whose basis vectors are $\mathbf{1}$ and $(-1)^n$; that of D^2 is $-\omega_m^2$, real, vanishing only on bin 0. □

Proof of the circular-derivative statement. (i) By linearity take $u(t) = e^{2\pi i\kappa t/L}$ with $|\kappa| \leq M < L/2$, combined in conjugate pairs for real u. Its samples occupy the single bin $\kappa \bmod L$, which is not the Nyquist bin; D^k multiplies it by $(2\pi i\kappa/L)^k$, the factor relating $u^{(k)}$ to u, and the inverse transform returns the samples of $u^{(k)}$. The Hermitian pairing is consistent because $\overline{(j\omega_\kappa)^k} = (j\omega_{-\kappa})^k$.

(ii) Write $x = u + (\tilde{\Delta}/L)q$ with $q[n] = n$. Both D^k and the reference derivative are linear, so $D^k x - x^{(k)} = (D^k u - u^{(k)}) + (\tilde{\Delta}/L)(D^k q - q^{(k)})$, where $q^{(1)} \equiv 1$ and $q^{(k)} \equiv 0$ for $k \geq 2$. The first bracket vanishes by (i), so $e_k = (D^k q - q^{(k)})/L$.

(iii) By Lemma 1, $(D^1 q)[n] = \sum_m d_m\,((n-m) \bmod L)$. For $m \leq n$ the index is $n-m$ and for $m > n$ it is $n-m+L$, so $(D^1 q)[n] = n\sum_m d_m - \sum_m m\, d_m + L\, S(n) = -\sum_m m\, d_m + L\, S(n)$. Subtract 1 and divide by L.

(iv) $S(L-1) = 0$ as an empty sum and $S(0) = \sum_{m=1}^{L-1} d_m = 0$, so $e_1[0] = e_1[L-1] = -\frac{1}{L}\big(1 + \sum_m m\, d_m\big)$. Pairing m with $L-m$ and using $d_{L-m} = -d_m$, $d_{L/2} = 0$ gives $-\frac{1}{L}\sum_m m\, d_m = \sum_{m=1}^{L/2-1}\big(1 - \frac{2m}{L}\big)d_m$. With $x_m = \pi m/L \in (0, \pi/2)$, $d_m = (-1)^m x_m \cot x_m/m$, so this sum equals $\sum_{m<L/2}(1-\frac{2m}{L})\frac{(-1)^m}{m} + \frac{\pi}{L}\sum_{m<L/2}(-1)^m\phi(x_m)$ with $\phi(x) = (1-\frac{2x}{\pi})\frac{x\cot x - 1}{x}$, which is C^1 on $[0, \pi/2]$ since $x\cot x - 1 = -x^2/3 + O(x^4)$. Grouping consecutive terms, the alternating sum of $\phi(x_m)$ is bounded by $\|\phi\|_\infty + \frac{\pi}{4}\|\phi'\|_\infty$, so the second part is $O(1/L)$. In the first part, $\frac{2}{L}\big|\sum_{m<L/2}(-1)^m\big| \leq \frac{2}{L}$ and $\sum_{m=1}^{L/2-1}(-1)^m/m = -\ln 2 + O(1/L)$ by the alternating-series remainder. Adding the $-1/L$ term gives (iv).

(v) Summing $D^1 x$ over the grid returns L times its DC coefficient, which is 0. Summing the reference derivative gives $\sum_n u'(n) + \tilde{\Delta} = \tilde{\Delta}$, because u' is a trigonometric polynomial of degree $< L/2$ without constant term. Hence $\sum_n \tilde{\Delta} e_1[n] = -\tilde{\Delta}$. □

Remark 1. On a real window the decomposition of A2 always exists, so the statement is exact; what changes is what $\tilde{\Delta}$ measures. For a smooth cycle whose period does not divide L, $\tilde{\Delta}$ is close to its

periodic-extension jump, and removing the cycle removes that contribution. White noise of variance σ^2 adds a term of standard deviation $2\sigma\sqrt{L}$, which no prototype removes; the measurement below shows how much of it removing the per-phase mean takes out on Solar and PEMS08, where without the profile the normalised statistic is already near its pure-noise value. That the two edge samples carry the largest error, and that the profile decays into the window like the tail of the alternating harmonic series, is checked numerically, not proved here.

What it gives. The error is proportional to $\tilde{\Delta}$, not to $x[L-1]-x[0]$: a smooth component whose period divides L, and whose harmonics stay strictly below the Nyquist frequency, contributes nothing to $\tilde{\Delta}$, however large its in-window endpoint difference, and one whose period does not contributes about its periodic-extension jump. (At Nyquist the claim would fail: $x[n]=(-1)^n$ has period $2 \mid L$ and $\tilde{\Delta}=-2L$.) That reading (Remark 1, not the identity) is the design reason for removing the prototype before the lifting. The identity holds on every panel whose lifting forms derivatives. The reading concerns the six of them with $P \nmid L$ (ETTm1, ETTm2, Weather, Exchange, Solar, PEMS08), and the prototype is on in two, Solar ($P=144$) and PEMS08 ($P=288$). The other five prototype-carrying panels (PEMS03, PEMS04, PEMS07, Traffic, TrafficC) run the banded variant, which forms no D^k, and on Traffic $P=L$. Because $\tilde{\Delta}$ also carries the noise's Nyquist coefficient, the identity does not size the benefit. Items (iii)–(v) characterise $k=1$ only; the second-order profile e_2 is not characterised here. The identity does not say on which side of the normaliser the prototype must be removed: removing it after the normaliser and before the lifting shrinks $\tilde{\Delta}$ equally. A boundary variant that subtracts a ramp fitted to $x[L-1]-x[0]$ removes the wrong jump. It was measured and is used on none of the fourteen panels. Subtracting the ramp $\tilde{\Delta}\, n/L$, which by (ii) makes D^k exact on the window, has not been run.

Measurement. On Solar and PEMS08, subtracting the per-phase training mean (a model-free stand-in for the prototype) shrinks the median $|\tilde{\Delta}|$ over test windows and variates to $0.805\times$ and $0.863\times$.

J.3 The equaliser

Statement 3 (Per-variate whitening)**.** *Assume A4. (ii) One multiplier whitens both X_1 and X_2 on Ω, i.e. makes each of $|m|^2S_1$ and $|m|^2S_2$ a positive constant there, iff S_1/S_2 is constant on Ω; under A5 this is $s_1=s_2$. Assume now also A5. (i) A multiplier m whitens X_i on Ω iff $|m(\omega)|=\kappa_i|\omega|^{s_i}$ for a constant $\kappa_i>0$. (iii) The whitening log-multipliers $\log|m_i|$ of all variates lie in $\operatorname{span}\{1,\log\omega\}$, a two-dimensional space of functions on Ω, independent of N. (iv) A channel carrying the extra factor $|\omega|^k$ is whitened by $|m|\propto|\omega|^{s_i-k}$, so the multiplier must also depend on the order k.*

Proof. By the spectral representation, $X_i=\int_\Omega e^{i\omega t}dZ_i(\omega)$ with orthogonal increments and $\mathbb{E}|dZ_i|^2=S_i\,d\omega$; an LTI filter maps dZ_i to $m\,dZ_i$, so the output density is $|m|^2S_i$. (ii) If m whitens both, $S_1/S_2=(|m|^2S_1)/(|m|^2S_2)$ is constant on Ω; conversely, if $S_1=cS_2$ then $|m|^2=1/S_1$ whitens both. Under A5, $S_1/S_2=(c_1/c_2)|\omega|^{2(s_2-s_1)}$, constant on a set with more than one point iff $s_1=s_2$. (i) $|m|^2c_i|\omega|^{-2s_i}=$ const iff $|m|=\kappa_i|\omega|^{s_i}$. (iii) Take logarithms in (i); 1 and $\log\omega$ are linearly independent on Ω. (iv) Replace S_i by $|\omega|^{2k}S_i$ in (i). □

What it gives. Necessity of a per-variate and per-order multiplier; the minimality of ours is left open. The implemented factor $\sigma(\beta_k+\mathbf{V}_i^{\mathrm{eq}\,\top}W_k)$ is additive in the logit, which matches an additive log-multiplier only where the mask attenuates strongly and saturates towards a shared multiplier elsewhere. Its rank is a hyper-parameter, not the two of (iii). On the banded variant the code shares β_0 and W_0 across orders, so (iv) is not implemented there.

The premise, measured. We fitted s_i per variate on each training split (Welch periodogram, least squares of $\log S$ on $\log\omega$ over $1/L\le f<1/2$ with the cycle harmonics excluded), with the reading fixed before computing: the premise of part (ii) is visible if the interquartile range of the s_i exceeds 0.5. It does so on Weather only (0.646). On every other dataset it is 0.025 to 0.174, and on Solar 0.064, among the most uniform. This measures the slope form of the premise. Spectra that are not proportional but have equal fitted slopes (for instance through the excluded harmonics) meet

the weaker form of part (ii), which is not measured. The share of variate pairs whose slopes differ beyond their standard errors is high everywhere, but it is 0.49 on synthetic variates with identical slopes, so it is read against that rate, not against zero.

Is per-variate equalisation needed? The same design with a single mask shared by all variates, on four of the ablation datasets whose design forms the derivative (Electricity, ETTh1, Exchange, Solar), H=96, three seeds, each against a control re-run in the same stage, raises test MSE by $+1.50\%$ pooled over the twelve paired cells ($+0.008\%$ once Solar is removed), with the sign agreeing on three of the four datasets. The effect is carried by Solar ($+5.96\%$, seed spread $+1.6$ to $+9.0\%$). On the other three the dataset mean is within $\pm 0.05\%$ of the control, but that agreement is a mean effect and not a per-seed one: Exchange swings from $+5.38\%$ to -4.74% across the three seeds and changes sign (standard deviation 5.10%), while Electricity and ETTh1 stay within $\pm 0.26\%$. The arm therefore tests the design choice rather than the premise of part (ii): none of the four datasets reaches the interquartile range at which that premise is visible (0.064 to 0.155, against the 0.5 bar), so the number says what a shared mask costs where the spectra are close to proportional, not what per-variate whitening buys where they are not.

K Design remarks

Seven design remarks shaped module choices, each with its proof and its classical source: the period remark's part (ii) is a continuous-time form of the periodicity lemma of Fine & Wilf (1965); the jet remark is the Koopman invariance of the linear observables of a linear system; the mean-field remark's part (i) is the characterisation of permutation-equivariant linear maps of Zaheer et al. (2017) restricted to zero row sums, read in the linear case; the objective remark is Parseval's identity; and parts (iii)–(v) of the frequency-term remark are the condition of Kruskal (1968), Aitken's generalised Gauss–Markov theorem and the diagonalisation of circulant matrices by the DFT.

K.1 The period

Design remark 4 (The period is not a window feature)**.** *Let $\gamma \in C^1(\mathbb{R})$ be periodic with minimal period P, observed on $[0, L]$. (i) For every $P' > L$ with $P' \notin P\mathbb{Z}$ there is a non-constant P'-periodic $\gamma' \in C^1(\mathbb{R})$ with $\gamma'|_{[0,L]} = \gamma|_{[0,L]}$, whose minimal period divides P' and so differs from P; if $L < P$, every $P' \in (L, P)$ qualifies. Hence on any class containing γ and such a γ', no functional of the window returns the minimal period. (ii) Conversely, on a class of continuous non-constant periodic signals with minimal periods at most $P_{\max}$, a window with $L \geq P + P_{\max}$ determines the minimal period, as it does for integer periods sampled at $0, \dots, L-1$; in particular $L \geq 2P$ suffices when P is the longest period present.*

Proof. We use one fact twice: the periods of a continuous non-constant periodic function form a closed proper subgroup of $\mathbb{R}$, hence equal $p\mathbb{Z}$ for its minimal period p; the periods of a sequence on $\mathbb{Z}$ form a subgroup $p\mathbb{Z}$ of $\mathbb{Z}$.

(i) Let $a = \gamma|_{[0,L]}$. Since $P' > L$ the interval $[L, P']$ is non-degenerate, and the cubic Hermite interpolant c on it with $c(L) = a(L)$, $c'(L) = a'(L)$, $c(P') = a(0)$, $c'(P') = a'(0)$ exists and is unique. If a is constant, c is the same constant; in that case add to c any non-zero C^1 function supported in a compact subset of (L, P'), which changes no endpoint datum. Define γ' as a on $[0, L]$ and c on $[L, P']$ and extend P'-periodically. Values and derivatives match at $t = L$ by the interpolation conditions and at the wrap $t = P'$ because $c(P') = a(0)$ and $c'(P') = a'(0)$, so $\gamma' \in C^1(\mathbb{R})$; it is P'-periodic, agrees with γ on $[0, L]$, and is non-constant (either a is, or the bump makes it so). Its minimal period p' satisfies $P' \in p'\mathbb{Z}$. If $p' = P$ then $P' \in P\mathbb{Z}$, which was excluded; so $p' \neq P$. When $L < P$, any $P' \in (L, P)$ has $0 < P' < P$ and hence $P' \notin P\mathbb{Z}$. A functional of the window takes the same value on γ and γ', whose minimal periods differ, so it cannot return the minimal period on both.

(ii) Let γ_1, γ_2 be in the class with minimal periods $P_1 \geq P_2$, agreeing on $[0, L]$ with $L \geq P_1 + P_2$ (this is implied by $L \geq P + P_{\max}$ when one of them is the observed signal, and by $L \geq 2P_{\max}$ always). For $t \in [0, P_1]$ both t and $t + P_2$ lie in $[0, L]$, so $\gamma_1(t + P_2) = \gamma_2(t + P_2) = \gamma_2(t) = \gamma_1(t)$. The function $h(t) = \gamma_1(t + P_2) - \gamma_1(t)$ is P_1-periodic and vanishes on a full period, so $h \equiv 0$ and

P_2 is a period of γ_1. Then $P_2 \in P_1\mathbb{Z}$ with $0 < P_2 \le P_1$, so $P_2 = P_1$. For sequences observed at $0, \dots, L-1$ the same argument runs over $t \in \{0, \dots, P_1 - 1\}$, which needs samples up to index $P_1 + P_2 - 1 \le L - 1$. Only continuity was used, not C^1. The special case is $P_{\max} = P$. □

Scope. Inside the class $\{t \mapsto \gamma(t + \varphi)\}$ with γ fixed, the phase is a function of the window whenever $\varphi \mapsto \gamma|_{[\varphi,\varphi+L]}$ is injective, which holds for generic γ and every L. The remark concerns the period, not the phase, and says nothing about whether supplying the phase improves accuracy or about the band $P \le L < P + P_{\max}$.

The period is learnable from the training split, which holds many windows at different phases. The remark says that a forecaster which must commit to a cycle length cannot take that length from a window that does not hold a full cycle; the length comes from outside, and the phase within the cycle is defined only once it is (Design remark 5).

K.2 The cycle a window cannot determine

Design remark 5 (A window that does not close the cycle does not fix the cycle length)**.** *On any class containing a C^1 periodic γ of minimal period P and a continuous non-constant P'-periodic γ' with $P' \in (L, P)$ that agrees with γ on $[0, L]$, where $L < P$: no functional of the window returns the minimal period, and the two signals differ on some interval of positive length beyond L. A forecaster that must commit to a cycle length before it can use the phase therefore cannot take that length from the window; the length comes from the protocol (Design remark 4).*

Proof. Design remark 4(i) supplies the γ': it agrees with γ on $[0, L]$ and has a different minimal period. Every period of γ' is at most $P' < P$, so its minimal period is not P, whichever γ' is taken. The two signals cannot agree on all of $[0, \infty)$: if they did, P' would be a period of γ, so $P' \in P\mathbb{Z}$ by the subgroup argument of that remark, which is impossible for $0 < P' < P$. A point where they differ lies beyond L, and since both are continuous they differ on an interval of positive length around it, and a forecaster whose input is a functional of the window takes the same value on both. Part (ii) of the remark is the complementary statement for $L \ge P + P_{\max}$. □

K.3 The jet

Design remark 6 (Koopman-invariant lifting)**.** *Assume A3 and let $z = (\mathbf{x}, \dot{\mathbf{x}}) \in \mathbb{R}^{2N}$, so $\dot{z} = Fz$ with $F = \left(\begin{smallmatrix} 0 & I \\ B & A \end{smallmatrix}\right)$, flow $\Phi_t = e^{Ft}$, and let $J(z) = Tz$ with $T = \left(\begin{smallmatrix} I & 0 \\ 0 & I \\ B & A \end{smallmatrix}\right) \in \mathbb{R}^{3N \times 2N}$, whose blocks are $\mathbf{x}, \dot{\mathbf{x}}, \ddot{\mathbf{x}}$. (i) The span $\mathcal{J}$ of the $3N$ components of J is invariant under the Koopman generator and semigroup; with $M = TFT^+$, $\frac{d}{dt}J(z(t)) = MJ(z(t))$ along every trajectory, and $J(\Phi_t z) = K_t J(z)$ with $K_t = Te^{Ft}T^+$. (ii) $\dim \mathcal{J} = 2N$, and $\mathcal{J}$ is the space of all linear observables of z, already spanned by the components of $(\mathbf{x}, \dot{\mathbf{x}})$. Consequently, whenever the window samples $\mathbf{x}(0), \dots, \mathbf{x}(L-1)$ determine z, their span is Koopman-invariant as well. (iii) Under A2 with $\tilde{\Delta} = 0$ and the equaliser off, channel $k \in \{0, 1, 2\}$ of equation 2, before the division by channel 0's norm, equals the sampled k-th derivative, and D^1, D^2 are injective off their kernels, subspaces of dimension 2 and 1.*

Proof. T has the identity as its top $2N \times 2N$ block, so it has full column rank $2N$ and $T^+T = I$. Then $MT = TFT^+T = TF$. Along a trajectory, $\frac{d}{dt}Tz = TFz = MTz$; componentwise $\mathcal{L}J_i(z) = \nabla J_i \cdot Fz = (MJ(z))_i \in \mathcal{J}$, which is generator invariance. For the semigroup, $K_t J(z) = Te^{Ft}T^+Tz = Te^{Ft}z = J(\Phi_t z)$, so every $J_i \circ \Phi_t$ lies in $\mathcal{J}$. (ii) The rows of T span a $2N$-dimensional subspace of the $2N$-dimensional dual of $\mathbb{R}^{2N}$, hence all of it, and the first two blocks already do; the third is the linear combination $B\mathbf{x} + A\dot{\mathbf{x}}$. Each sample $\mathbf{x}(n)$ is the linear functional $z \mapsto [I\ 0]e^{Fn}z$; if the samples determine z, their span is the whole dual space, which the linear flow maps into itself. (iii) is parts (i)–(ii) of the circular-derivative statement (Statement 2) with $\tilde{\Delta} = 0$, and Lemma 1. □

What it gives. Three channels is a choice the remark makes exact, not minimal: the acceleration channel adds no dimension, but it makes the equation of motion a fixed linear read-out of the lifted state. The remark does not distinguish the jet from the raw window, both of which span $\mathcal{J}$ whenever

they determine z, and the transport equation 3 is not the linear operator it licenses. With the equaliser on, channel k is D^k of a differently filtered signal, so (iii) describes the un-equalised jet only. At initialisation the mask is within $O(10^{-3})$ of the constant $1/2$ (the fingerprint and interaction weights are drawn with standard deviations 0.02 and 0.01), so the model starts near the exact jet and learns away from it.

K.4 THE MEAN FIELD

Design remark 7 (Centred coupling). *Let $n \geq 2$, $M = I - \mathbf{1}\mathbf{1}^\top/n$ and $\mathcal{W}_0 = \{W \in \mathbb{R}^{n\times n} : W\mathbf{1} = 0\}$. (i) The matrices $W \in \mathcal{W}_0$ with $P_\pi W P_\pi^\top = W$ for every permutation matrix P_π form $\operatorname{span}\{M\}$, of dimension one ($M \neq 0$, since $M_{11} = (n-1)/n$ for $n \geq 2$), whereas $\dim \mathcal{W}_0 = n(n-1)$. (ii) For tokens $v_1, \dots, v_n \in \mathbb{R}^d$ with mean $\bar{v}$, the centred simplex set $\mathcal{C} = \{\sum_i \alpha_i v_i - \bar{v} : \alpha \in \Delta^{n-1} = \{\alpha \in \mathbb{R}^n : \alpha_i \geq 0,\ \sum_i \alpha_i = 1\}\}$ is contained in the zero-sum set $\mathcal{Z} = \{\sum_i w_i v_i : \sum_i w_i = 0\}$, with equality iff all v_i are equal.*

Proof. (i) $P_\pi W P_\pi^\top = W$ for all π means $W_{\pi(i)\pi(j)} = W_{ij}$, so W is constant on the orbits of the permutation action on ordered pairs, which for $n \geq 2$ are the diagonal and the off-diagonal: $W = aI + b(\mathbf{1}\mathbf{1}^\top - I)$. Zero row sums force $b = -a/(n-1)$, i.e. $W = \frac{an}{n-1}M$; conversely M is permutation-invariant with zero row sums. The map $W \mapsto W\mathbf{1}$ is onto $\mathbb{R}^n$, so its kernel $\mathcal{W}_0$ has dimension $n^2 - n$. (ii) With $c_i = v_i - \bar{v}$, $\sum_i c_i = 0$. Since $\sum_i \alpha_i = 1$, $\mathcal{C} = \operatorname{conv}\{c_i\}$; since $\sum_i w_i = 0$, $\sum_i w_i v_i = \sum_i w_i c_i$, and any $\sum_i \lambda_i c_i$ is reached with $w_i = \lambda_i - \bar{\lambda}$, writing $\bar{\lambda}$ for the mean of the λ_i, so $\mathcal{Z} = \operatorname{span}\{c_i\}$. A convex hull lies in the span. $\mathcal{C}$ is compact, and a bounded linear subspace is $\{0\}$, so $\mathcal{C} = \mathcal{Z}$ forces every $c_i = 0$; conversely equal tokens give $\{0\}$ for both. □

What it gives, and how the code maps onto it. The implemented update subtracts the mean of the $n = KN$ kinematic tokens, pooled over the three orders, and scales by $1 + c$: on those tokens it is $(1 + c)M$. Covariate tokens are shifted by the same mean, so their rows still sum to zero and the cross-token part stays rank one, but the full map is not invariant under exchanging a kinematic and a covariate token. (i) applies to the kinematic block. The remark concerns *linear* maps: a non-linear permutation-equivariant aggregator is outside it, and the `response` completion is an instance of that family. The restriction to maps that commute with relabelling is a design choice, not a property of the tokens, which the structural codes mark as distinct. The remark therefore covers only couplings under which the tokens are exchangeable. It cannot select centring over couplings that depend on the token index, and it does not compare centring with Han et al. (2024)'s aggregation.

K.5 THE READ-OUT FLOOR

Design remark 8 (Simplex floor). *Let $K \geq 2$, $\varepsilon \in (0, 1/K)$, $\psi = (1 - K\varepsilon)\operatorname{softmax}(u) + \varepsilon\mathbf{1}$ for $u \in \mathbb{R}^K$, and $z = \sum_k \psi_k z_k$ with ψ not depending on $(z_k)_k$. Then ψ lies in the probability simplex with $\varepsilon \leq \psi_k \leq 1 - (K-1)\varepsilon$; for any loss $\mathcal{L}$ differentiable at z, $\partial\mathcal{L}/\partial z_k = \psi_k\, \partial\mathcal{L}/\partial z$, so $\|\partial\mathcal{L}/\partial z_k\| \geq \varepsilon\|\partial\mathcal{L}/\partial z\|$; and $\inf_u \min_k \psi_k = \varepsilon$, whereas at $\varepsilon = 0$ the infimum is 0 and the hard selector $\psi = (1, 0, \dots, 0)$ gives $\partial\mathcal{L}/\partial z_k = 0$ for $k \geq 1$.*

Proof. With $p = \operatorname{softmax}(u)$, $p_k > 0$ and $\sum_k p_k = 1$, so $\sum_k \psi_k = (1 - K\varepsilon) + K\varepsilon = 1$, $\psi_k \geq \varepsilon$ since $1 - K\varepsilon > 0$, and $\psi_k = 1 - \sum_{j\neq k} \psi_j \leq 1 - (K-1)\varepsilon$. As ψ does not depend on z_k, $z_k \mapsto z$ is linear with Jacobian $\psi_k I$, and the chain rule gives the gradient identity. Taking $u_k = 0$ and $u_j = t \to \infty$ for $j \neq k$ sends $p_k \to 0$ and $\psi_k \to \varepsilon$; at $\varepsilon = 0$ the same sequence sends $\psi_k \to 0$. □

What it gives. The bound is relative: it guarantees connectivity, not a non-zero gradient and not that an order is used. The mean field is a second path between orders, so velocity and acceleration receive exactly zero gradient only when the read-out *and* the mean field are both removed. The hypothesis that ψ ignores the transported tokens holds by construction: the gate reads only the lifting's spectral entropy.

K.6 THE OBJECTIVE

Design remark 9 (One Sobolev seminorm, two domains). *For $e \in \mathbb{R}^H$, weights $w_0, \dots, w_{K-1}$, the circular difference Δ and the unnormalised DFT $E_m = \sum_t e_t e^{-2\pi i m t/H}$,*

$$\sum_{k=0}^{K-1} w_k \|\Delta^k e\|^2 = \frac{1}{H} \sum_{m=0}^{H-1} \Big(\sum_{k=0}^{K-1} w_k |\rho(\omega_m)|^{2k} \Big) |E_m|^2,$$

with $\omega_m = 2\pi m/H$, $\rho(\omega) = e^{i\omega} - 1$ and $|\rho(\omega_m)|^2 = 4\sin^2(\pi m/H)$.

Proof. With $(Se)_t = e_{(t+1) \bmod H}$, re-indexing over $\mathbb{Z}/H\mathbb{Z}$ gives $\widehat{Se}_m = e^{i\omega_m} E_m$, with no boundary term because the difference is circular; so $\widehat{\Delta e}_m = \rho(\omega_m) E_m$ and, by induction, $\widehat{\Delta^k e}_m = \rho(\omega_m)^k E_m$. Parseval for the unnormalised DFT, $\sum_t |x_t|^2 = \frac{1}{H} \sum_m |X_m|^2$, follows from $\sum_m e^{-2\pi i m(t-s)/H} = H \mathbb{1}[t = s]$. Apply it to $\Delta^k e$, multiply by w_k and sum. □

What it gives. A weighted H^2 seminorm of the error has two exact forms, per step and per frequency. The implemented objective equation 4 is not that seminorm, for several reasons: its weights $(m/H)^k$ behave at low frequency like $|\rho|^k$ rather than $|\rho|^{2k}$, so no non-negative combination of the exact symbols reproduces them (at $H = 96$ the order-one ratio to the exact symbol $|\rho(\omega_m)|^2$ varies by a factor of 22 across the one-sided (`rfft`) axis). Its time differences are non-circular and truncated, which reintroduces a boundary term of the kind the circular-derivative statement (Statement 2) isolates; both halves pass through $\log(1 + \cdot)$; and its single-sided spectral average gives the DC and Nyquist bins twice the relative weight of the others (at $H = 96$, $96/49$ against $48/49$). Since $\log(1 + \cdot)$ is concave, Jensen's inequality bounds each robustified half by the logarithm of its mean energy, and by the remark with w the k-th unit vector those two energies are the same quantity for the printed circular, two-sided sums (not for the implemented single-sided average); the halves share that upper bound and nothing more. As the error's scale grows, the gap between the two log measures converges to $\mathrm{mean}_m \log(|E_m|^2/H) - \mathrm{mean}_t \log e_t^2$, finite for windows with no vanishing sample and no vanishing DFT bin, a limit that does not scale with the error, and the reason the balance can be a dimensionless number.

K.7 THE FREQUENCY TERM

Design remark 10 (What the frequency domain buys). *Let F be the unitary DFT on $\mathbb{R}^H$. Under A6: (i) over all measurable forecasters, $\arg\min_f \mathbb{E}\|y - f(\mathbf{x})\|^2 = \mu$, whatever Σ; (ii) $\sum_m |E_m|^2 = H \sum_t e_t^2$ for the unnormalised DFT, so an unweighted squared-magnitude two-sided frequency loss is H times the per-step squared error, with gradients in the same ratio. Under A7, with target μ at a fixed $\mathbf{x}$: (iii) $\beta_{\mathrm{OLS}} = (A^\top A)^{-1} A^\top \mu$ and $\beta_{\mathrm{GLS}} = (A^\top \Sigma^{-1} A)^{-1} A^\top \Sigma^{-1} \mu$ agree for every μ iff $\Sigma\,\mathrm{range}(A) = \mathrm{range}(A)$, and for a given μ when $\mu \in \mathrm{range}(A)$; (iv) for $y = A\beta^\star + \eta$ both estimators are unbiased and $\mathrm{Cov}(\hat\beta_{\mathrm{OLS}}) - \mathrm{Cov}(\hat\beta_{\mathrm{GLS}}) \succeq 0$. (v) Under A8, $F\Sigma F^*$ is diagonal with entries $\sum_r c_r e^{-2\pi i m r/H}$; a covariance is diagonalised by F only if it is circulant.*

Proof. (i) $\mathbb{E}[(y_h - f_h)^2 \mid \mathbf{x}] = \mathbb{E}[(y_h - \mu_h)^2 \mid \mathbf{x}] + (\mu_h - f_h)^2$ for each step h, the cross term vanishing because $\mathbb{E}[y_h - \mu_h \mid \mathbf{x}] = 0$; summing over h no off-diagonal entry of Σ appears, and the minimum is at $f = \mu$. (ii) is Parseval, as in the objective remark (Design remark 9). (iii) Both objectives are strictly convex quadratics and their normal equations give the formulas. Both maps are left inverses of A; a left inverse L is determined by its kernel, since $\mathbb{R}^H = \mathrm{range}(A) \oplus \ker L$. The kernels are $\mathrm{range}(A)^\perp$ and $\Sigma\,\mathrm{range}(A)^\perp$, which coincide iff Σ preserves $\mathrm{range}(A)^\perp$, equivalently (by symmetry and invertibility) $\mathrm{range}(A)$. If $\mu = A\beta^\star$, both objectives vanish at $\beta^\star$. (iv) With $G = (A^\top \Sigma^{-1} A)^{-1} A^\top \Sigma^{-1}$ and $D = (A^\top A)^{-1} A^\top - G$, $DA = 0$, and $\mathrm{Cov}(\hat\beta_{\mathrm{OLS}}) = (G + D)\Sigma(G + D)^\top = G\Sigma G^\top + D\Sigma D^\top$ because $G\Sigma D^\top = (A^\top \Sigma^{-1} A)^{-1}(DA)^\top = 0$; and $G\Sigma G^\top = (A^\top \Sigma^{-1} A)^{-1} = \mathrm{Cov}(\hat\beta_{\mathrm{GLS}})$. (v) A circulant matrix is $\sum_r c_r S^r$ for the cyclic shift S, which F diagonalises (proof of the objective remark); conversely $F^* \Lambda F$ is circulant for any diagonal Λ. With $c_r = c_{H-r}$ the entries are real. □

What it gives. A reason to write an error term over frequency: the DFT is the basis in which a circularly stationary label covariance is exactly diagonal. The Toeplitz covariance of an ordinary stationary process is not circulant, so by (v) it is not diagonalised at finite H; the classical Szegő–Gray asymptotic equivalence of Toeplitz and circulant matrices (summable autocovariance) (Grenander & Szegő, 1958) makes the off-diagonal part vanish only as $H \to \infty$. (i) is why we do not say per-step MSE is biased, and (ii) is why a useful frequency term must be non-quadratic or non-uniformly weighted. Our term is neither a multiple of the time loss nor the covariance-weighted (GLS) objective, whose weights are the inverse spectral density.