MARGINAL SCORING HIDES WHAT GENERATIVE FORECASTERS LEARN

Anonymous authors
Paper under double-blind review

ABSTRACT

Probabilistic multivariate time-series forecasting is dominated by a family of models that pair a deterministic backbone with a generative component, and that report CRPS and QICE as headline numbers. We show that both metrics, as implemented throughout this literature, are computed independently at every (time step, channel) element: they score the predictive *marginals* and are blind, by construction, to dependence across the forecast horizon. We make the consequence concrete. Taking a recent state-of-the-art model and holding its backbone, normalisation, training objective, sampling budget and evaluation code fixed, we replace only the generative sampler with a diagonal Gaussian head that *cannot represent any temporal dependence*. This deliberately dependence-free floor beats the published model on CRPS in 10 of 14 dataset$\times$horizon cells (by 0.8–5.2%, at seed standard deviations of at most 0.0013) and is generally better calibrated; the four exceptions are the two highest-dimensional benchmarks, where its deficit is 1.1–1.4%, and a lookback sweep shows that which side wins CRPS is a regime question, decided jointly by dimensionality and context length. Under a dependence-sensitive diagnostic there is no regime: the lag-1 autocorrelation of the sampler's *residual* paths — what it adds beyond the conditional mean, which is zero for the floor by construction — is positive for the generative model in every one of the 20 cells we ran, across seven benchmarks, two horizons and three lookbacks. It is also small: the true forecast-error autocorrelation is 0.6–0.9 everywhere, and the generative model recovers 10–29% of it on the 7- and 21-series benchmarks, and all of it only on the two high-dimensional ones — exactly the cells where it also wins CRPS. The two metric families therefore rank the same pair of models in opposite directions, and the literature reports only one of them. We then isolate the cause. Across three datasets, giving the sampler an explicit horizon covariance buys 43–$287\times$ less autocorrelation than changing the objective does, so the *objective*, not the architecture, is the binding constraint. Replacing it with a joint proper scoring rule recovers 71–94% of the ground-truth autocorrelation (versus 36–46% for the published model), and an energy-score/KDE hybrid dominates the published model on CRPS, QICE, energy score and variogram score simultaneously on all three.

1 INTRODUCTION

A large and fast-moving line of work on probabilistic multivariate time-series (MTS) forecasting has converged on a common recipe: a deterministic forecaster supplies a conditional mean (and, more recently, a conditional scale), and a generative component—a diffusion model, a flow, or a learned push-forward—models what is left over (Li et al., 2024; Ye et al., 2025; Li et al., 2025; 2026). This lineage sits inside a larger and fast-growing literature on diffusion and flow models for time-series forecasting (Shen et al., 2024; Fan et al., 2024; Lee et al., 2024; Liu et al., 2024; Guo et al., 2025; Wang et al., 2025; Wu et al., 2025; Solano-Carrillo et al., 2026; Wu et al., 2026; Jiang & Xiong, 2026), and successive papers report steady gains on a fixed benchmark suite, measured almost exclusively by the Continuous Ranked Probability Score (CRPS) and the Quantile Interval Coverage Error (QICE).

This paper asks a question that the benchmark itself cannot answer: *what is the generative component contributing?* We ask it about one specific kind of structure — *temporal dependence across*

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.

the forecast horizon within each series. That is the scope of every dependence measurement below, and it is forced on us: the audited architecture is channel-factorised and cannot express cross-series dependence at all (Section 3), so a cross-series analogue would measure a quantity the model cannot represent. Where we write "dependence" without qualification we mean this temporal kind; we make no claim about the full multivariate joint law. Our starting point is a property of the evaluation code rather than of any model. In every implementation in this lineage that we examined, CRPS is obtained by flattening the sampled tensor of shape (batch, samples, horizon, channels) into independent univariate problems, one per (time step, channel), and averaging; QICE is a coverage histogram accumulated the same way. Both are therefore functions of the predictive *marginals* alone. A forecaster that gets every marginal exactly right and the joint law entirely wrong scores perfectly.

That observation would be pedantic if the gap were empirically small. It is not. We construct a **matched parametric floor**: a recent state-of-the-art model with its encoder, normalisation, loss, budget, sample count and evaluation code left exactly as published, and only the mechanism producing the predictive distribution replaced by a diagonal Gaussian over the horizon — independent noise at every step, hence no temporal dependence at all. It nonetheless beats the published model on CRPS in 10 of the 14 dataset$\times$horizon cells we test, at seed standard deviations one to two orders of magnitude smaller than the gaps (Figure 1a). The four exceptions are informative rather than fatal: they are the two highest-dimensional benchmarks (Electricity, 321 series; Traffic, 862), and a lookback sweep shows the same boundary moving with context length — on Weather the floor leads at $L{=}96$ and trails at $L{=}336$. Which side wins CRPS is a regime question.

The generative model is not, however, doing nothing. We measure what its sampler adds beyond the conditional mean by the lag-1 autocorrelation of its *residual* paths (sample minus sample mean), a quantity that is zero for the floor by construction and 0.6–0.9 for the actual forecast errors. On this diagnostic the generative model is ahead in all 14 cells (Figure 1b) and in all six further lookback cells — but on the 7- and 21-series benchmarks it recovers only 10–29% of the true residual dependence, and it reaches the truth only on Electricity and Traffic, which are precisely the cells where it also wins CRPS. *The two metric families rank the same pair of models in opposite directions.* The field reports only the family under which the trivial baseline wins. And no model is close: on ETTh1/192 the true forecast errors have lag-1 autocorrelation 0.74, the published model's residuals 0.16, the floor's 0; NsDiff (Ye et al., 2025), an actual diffusion model from the same lineage that we also ran end to end, produces sampled paths with less autocorrelation than the floor, and less than the plain Gaussian readout of its *own* mean and variance networks.

Finally we ask what would fix it, and separate two candidate explanations with a 2×2 design over sampler *structure* (diagonal versus low-rank plus diagonal horizon covariance) and training *objective* (the published point-wise likelihood versus joint proper scoring rules). The structure axis is inert on its own: adding a horizon covariance under the published objective changes CRPS, QICE, variogram score and autocorrelation by less than one seed standard deviation. The objective axis moves everything. Our contributions:

- We document, at the level of the evaluation code, that the headline metrics of this literature score marginals only, and we distinguish this blind spot from the previously reported failures of CRPS-Sum (Koochali et al., 2022; Marcotte et al., 2023), which are a different and in fact opposite phenomenon (Section 2).
- We introduce a matched parametric floor and show a rank reversal between marginal and dependence-sensitive scoring across seven benchmarks, three seeds per cell, together with the dimensional boundary at which it stops (Sections 5–6).
- We isolate the cause with a structure$\times$objective design replicated on three datasets: the objective binds, the architecture does not (Section 7).
- We show a cheap fix. Training the published architecture with a variogram score recovers 71–94% of the ground-truth autocorrelation; an energy-score/KDE hybrid dominates the published model on CRPS, QICE, energy score *and* variogram score at once, on every dataset tested (Section 8).

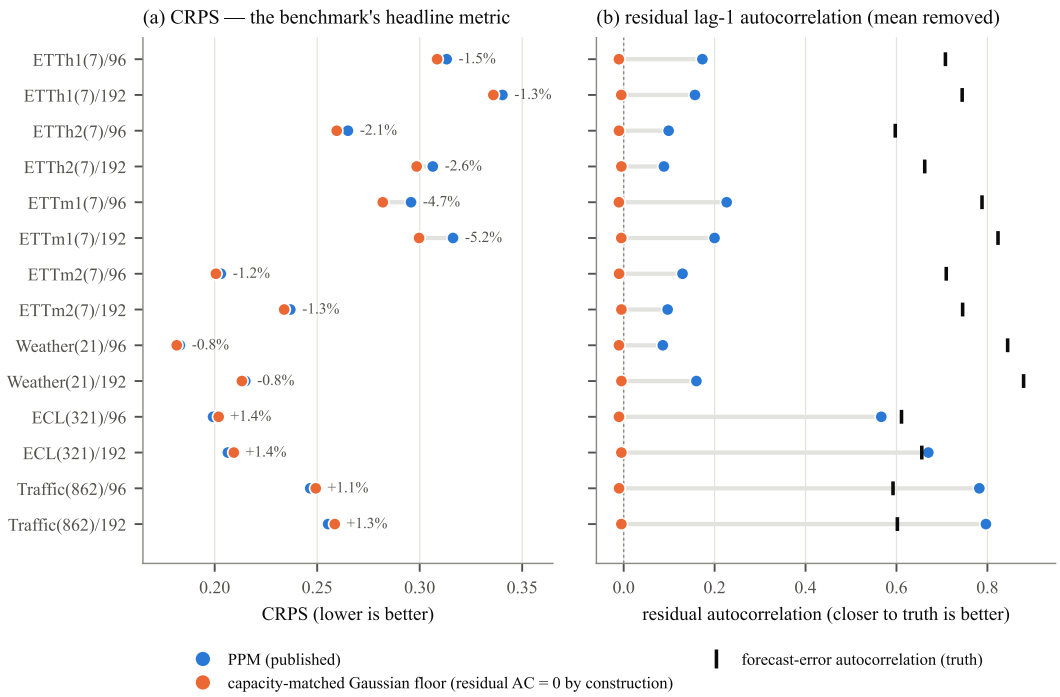


Figure 1: **The two metric families rank the models in opposite directions.** Each row is one dataset/horizon cell (series count in parentheses); markers are means over three seeds. **(a)** On CRPS—the score this literature reports—the dependence-free capacity-matched Gaussian floor is better in 10 of 14 cells (percentages are the floor's relative advantage); it loses only on the two highest-dimensional benchmarks. **(b)** On the lag-1 autocorrelation of the sampled paths the ordering flips in *all* 14 cells, and both models fall short of the ground truth.

2 Related Work

Critiques of multivariate forecast scoring. Koochali et al. (2022) show that a noise-like dummy model can obtain a better CRPS-Sum than strong forecasters, and Marcotte et al. (2023) give a finite-sample power analysis showing that CRPS-Sum and the energy score fail to detect simple distributional deviations at realistic test-set sizes. Both target *CRPS-Sum*, whose flaw is that summing across dimensions hides per-dimension error. Our blind spot is the opposite one: the per-element CRPS of equation 1 resolves every dimension separately and therefore cannot see dependence *between* them. The distinction is not academic—PPM and NsDiff report per-element CRPS, not CRPS-Sum, so the earlier critiques do not apply to their headline numbers, whereas ours does. Scheuerer & Hamill (2015) and Pinson & Tastu (2013) supply the diagnostics we use.

Strong simple baselines. Zeng et al. (2023) showed a linear model matching transformers on point forecasting; ours is the probabilistic analogue with a twist — the simple baseline wins the reported metric while being worse on the property the elaborate model was built to capture.

Dependence-aware forecasters. Copula-based (Salinas et al., 2019; Drouin et al., 2022) and GP-prior (Kollovieh et al., 2025) forecasters model dependence explicitly; Wang et al. (2025) give the diffusion noise a moving-average structure for the same reason. Work on the design space of time-series diffusion — schedules (Lee et al., 2024), elucidated parameterisations (Solano-Carrillo et al., 2026), multi-scale decompositions (Shen et al., 2024; Fan et al., 2024), second-order flows (Jiang & Xiong, 2026) — is scored with the same marginal suite, so our diagnostics apply to it directly. Our point is not that dependence modelling is unsolved in general, but that within the benchmark that this specific lineage optimises, dependence is invisible, and the lineage's own objective does not ask for it.

3 THE EVALUATION PROTOCOL UNDER EXAMINATION

The model family. We study the lineage TMDM (Li et al., 2024) → NsDiff (Ye et al., 2025) → D^{3}U (Li et al., 2025) → PPM (Li et al., 2026). All share a template: a deterministic network maps the context $\boldsymbol{x}_{1:T}$ to the parameters of a simple conditional distribution, which the generative component then transports to the predictive law of $\boldsymbol{y}_{1:H}$. We take PPM, the most recent and strongest member, as the primary object of study; its released implementation is compact and its parametric prior is explicit, which makes the floor comparison unusually clean.

The metrics. Given S sampled futures $\{\boldsymbol{y}^{(s)}\}_{s=1}^{S} \subset \mathbb{R}^{H\times D}$ and an observed future $\boldsymbol{y}$, the reported CRPS is

$$\mathrm{CRPS}_{\text{reported}} = \frac{1}{HD}\sum_{h=1}^{H}\sum_{d=1}^{D}\mathrm{crps}\Big(\{y_{h,d}^{(s)}\}_{s=1}^{S},\, y_{h,d}\Big), \tag{1}$$

i.e. the univariate CRPS of each element's marginal, averaged. QICE is a coverage histogram over the same flattened elements. Both are minimised by *any* joint law with the correct marginals; a sampler drawing each $y_{h,d}$ independently is not penalised at all, however wrong its dependence structure. This is a property of equation 1, not of finite samples.

Two further code-level facts. PPM's encoder operates on the transposed context, so all linear maps act along time and channels are processed independently with shared weights: the model has no mechanism for cross-series dependence, although it is evaluated as a *multivariate* method. This is why every dependence diagnostic in this paper is temporal (within a series, across the horizon); the cross-series axis is not merely unmeasured by the benchmark but unrepresentable by the model. Its training loss is a kernel density estimate evaluated per element, which the paper itself notes "does not take into account the joint distribution modeling of the labels".

Dependence-sensitive statistics. We add two standard diagnostics computed on the same samples. The variogram score of order p (Scheuerer & Hamill, 2015) over a lag set $\mathcal{L}$,

$$\mathrm{VS} = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\sum_{h,d}\Big(|y_{h+\ell,d}-y_{h,d}|^{p} - \tfrac{1}{S}\textstyle\sum_{s}|y_{h+\ell,d}^{(s)}-y_{h,d}^{(s)}|^{p}\Big)^{2}, \tag{2}$$

with $p = 1$ and $\mathcal{L} = \{1, 2, 4, 8, 16, 32\}$, and a *residual* autocorrelation. The autocorrelation of a sampled path conflates the conditional mean's structure, which any point forecaster carries, with the dependence the sampler adds around it. We therefore subtract the sample mean $\bar{\boldsymbol{y}} = \frac{1}{S}\sum_s \boldsymbol{y}^{(s)}$ and compare the lag-1 autocorrelation of the residual paths $\boldsymbol{y}^{(s)} - \bar{\boldsymbol{y}}$ with that of the forecast error $\boldsymbol{y} - \bar{\boldsymbol{y}}$; for an independent-noise sampler the former is zero by construction, so this isolates what the generative component adds (Appendix D). We also report the energy score. The variogram score was designed precisely because the energy score discriminates dependence structure poorly (Pinson & Tastu, 2013); our results reproduce that asymmetry.

4 WHAT THE REPORTED SCORE CAN AND CANNOT IDENTIFY

The empirical results below are all consequences of two facts about the objective and the metric used in this literature. We state them here; proofs are in Appendix A.

Write $\mathcal{P}(\mathbb{R}^n)$ for distributions on $\mathbb{R}^n$ with finite first moment, $n = HD$, and for $P \in \mathcal{P}(\mathbb{R}^n)$ let P_i denote its i-th one-dimensional marginal. Call a score *marginal* if it has the form $S(P, \boldsymbol{y}) = \sum_{i=1}^{n} s_i(P_i, y_i)$.

Proposition 1 (The reported score is proper but not strictly proper)**.** *Let each s_i be a strictly proper scoring rule for $\mathcal{P}(\mathbb{R})$. Then the marginal score S is proper for $\mathcal{P}(\mathbb{R}^n)$, and its expected-score minimiser set under any true law Q is exactly the Fréchet class $\mathcal{F}(Q) = \{P : P_i = Q_i\ \forall i\}$. In particular S assigns the same expected score to Q and to the independence coupling $\bigotimes_i Q_i$, and cannot separate any two elements of $\mathcal{F}(Q)$.*

Equation 1 is of exactly this form, with s_i the univariate CRPS, which is strictly proper (Gneiting & Raftery, 2007); QICE is a coverage statistic of the same marginals. So the entire reported metric

suite is constant on $\mathcal{F}(Q)$. The next statement says the indifference set is not a technicality but spans the whole range of dependence.

Corollary 2 (The optimum set spans all dependence structures)**.** *Fix Gaussian marginals $Q_i = \mathcal{N}(\mu_i, \sigma_i^2)$. Then for every correlation matrix $\boldsymbol{R}$ in the elliptope $\mathcal{E}_n = \{\boldsymbol{R} \succeq 0,\ R_{ii} = 1\}$ there is a $P \in \mathcal{F}(Q)$ with correlation $\boldsymbol{R}$. Consequently the lag-ℓ autocorrelation of a score-optimal forecaster is unconstrained on $[-1, 1]$ for every ℓ.*

The second fact concerns the *objective*, and explains why enlarging the sampler does not help. Consider the two families used in Sections 5 and 7, for a fixed context,

$$\mathcal{F}_{\mathrm{diag}} = \{\mathcal{N}(\boldsymbol{m}, \mathrm{diag}(\boldsymbol{v}))\}, \qquad \mathcal{F}_{\mathrm{lr}} = \{\mathcal{N}(\boldsymbol{m}, \boldsymbol{L}\boldsymbol{L}^\top + \mathrm{diag}(\boldsymbol{s}^2))\},\ \ \boldsymbol{L} \in \mathbb{R}^{H \times r}, \tag{3}$$

and let $\mathcal{L}(P) = \sum_h \ell_h(P_h)$ be any marginal objective.

Proposition 3 (A marginal objective is blind to, and gives no gradient to, the covariance)**.** *For $P = \mathcal{N}(\boldsymbol{m}, \boldsymbol{L}\boldsymbol{L}^\top + \mathrm{diag}(\boldsymbol{s}^2))$ one has $P_h = \mathcal{N}(m_h, \|\boldsymbol{L}_{h\cdot}\|^2 + s_h^2)$, and therefore*

- *(i) $\mathcal{F}_{\mathrm{diag}}$ and $\mathcal{F}_{\mathrm{lr}}$ induce the same set of marginal collections, hence $\inf_{\mathcal{F}_{\mathrm{lr}}} \mathcal{L} = \inf_{\mathcal{F}_{\mathrm{diag}}} \mathcal{L}$: the extra parameters cannot lower the objective;*
- *(ii) $\mathcal{L}$ is constant on $\{\boldsymbol{L}' : \|\boldsymbol{L}'_{h\cdot}\| = \|\boldsymbol{L}_{h\cdot}\|\ \forall h\}$, a set containing loadings that realise different covariances (e.g. any row sign flip $\boldsymbol{L}_{h\cdot} \mapsto -\boldsymbol{L}_{h\cdot}$);*
- *(iii) $\nabla_{\boldsymbol{L}_{h\cdot}} \mathcal{L} = 2\,\ell'_h(v_h)\,\boldsymbol{L}_{h\cdot}$ with $v_h = \|\boldsymbol{L}_{h\cdot}\|^2 + s_h^2$, so gradient flow rescales each loading row but never rotates it. The dependence structure of the trained sampler is the one present at initialisation.*

PPM's training loss is a per-element kernel density likelihood plus a mean-squared term on the sample mean; both are marginal in the sense above (in expectation over the sampling noise), so Proposition 3 applies. It predicts the null result of Section 7 quantitatively: with standard random initialisation the rows $\boldsymbol{L}_{h\cdot}$ are near-orthogonal, the induced autocorrelation is near zero, and training cannot move it.

Proposition 4 (What the joint objectives identify)**.** *The energy score is strictly proper on $\mathcal{P}(\mathbb{R}^n)$ (Gneiting & Raftery, 2007), so its minimiser is Q itself. The variogram score of order p in equation 2 is proper but* not *strictly proper: it depends on P only through the variogram $\gamma_{ij}^{(p)}(P) = \mathbb{E}_P|X_i - X_j|^p$, and is minimised by every P with $\gamma^{(p)}(P) = \gamma^{(p)}(Q)$.*

Remark 5. This is why we report both: the energy score repairs the gap of Proposition 1 in principle but discriminates dependence weakly in high dimension (Pinson & Tastu, 2013); the variogram score discriminates sharply but pins down only $\gamma^{(p)}$. Section 8 shows both effects.

5 A MATCHED PARAMETRIC FLOOR

We hold everything fixed except the origin of the predictive distribution. PPM encodes the normalised context to $\boldsymbol{h}$, samples a latent $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{h}), \mathrm{diag}\,\boldsymbol{\sigma}^2(\boldsymbol{h}))$ in a d_{model}-dimensional space, and pushes $(\boldsymbol{z}, \boldsymbol{h})$ through an MLP decoder to obtain a sampled future. Our floors keep the same encoder and an equal-depth head, and emit the predictive law directly:

$$\boldsymbol{y}_{:,d}^{(s)} = \boldsymbol{m}_d(\boldsymbol{h}) + \boldsymbol{s}_d(\boldsymbol{h}) \odot \boldsymbol{\epsilon}^{(s)}, \qquad \boldsymbol{\epsilon}^{(s)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_H), \tag{4}$$

with $\boldsymbol{m}_d, \boldsymbol{s}_d \in \mathbb{R}^H$. Because $\boldsymbol{\epsilon}^{(s)}$ is white, the sampled paths have zero temporal correlation beyond that of the conditional mean. We use three variants: GAUSS (equal head width, 140k parameters versus PPM's 337k), $\text{GAUSS}_{\text{cap}}$ (widened to match PPM's parameter count), and a Student-t version. We also run NO-LATENT: PPM's own decoder with the latent noise switched off and a learned output scale, which isolates the stochastic push-forward itself. Every variant is trained with PPM's exact loss (a weighted point-wise KDE likelihood plus a mean-squared term), for the same 30 epochs, with the same optimiser settings, and evaluated with PPM's own code.

6 THE RANK REVERSAL

Setup. We use all seven benchmarks of Li et al. (2026) — the four ETT datasets, Weather (21 series), Electricity (321) and Traffic (862) — with their protocol: context 96, horizons $H \in \{96, 192\}$

Table 1: **The two metric families rank the models in opposite directions.** Means over three seeds; "floor" is the capacity-matched diagonal Gaussian, which cannot represent temporal dependence. "Residual" autocorrelation is computed after subtracting the sample mean from each path (Section 3); it is zero for the floor by construction, and "truth" is the autocorrelation of the forecast error $\boldsymbol{y} - \bar{\boldsymbol{y}}$. Bold marks the better of the two (closest to truth, for autocorrelation). The floor wins CRPS in 10 of 14 cells and QICE in 6; PPM's residual autocorrelation is positive in all 14, and it wins the variogram score in 12. Per-seed standard deviations, the two further floors and the sampled-path autocorrelation are in Tables 6–4.

	CRPS ↓		QICE (%) ↓		residual lag-1 autocorr.			variogram score ↓	
dataset (D) / H	PPM	floor	PPM	floor	PPM	floor	truth	PPM	floor
ETTh1 (7) / 96	0.3132	**0.3085**	1.22	**1.05**	**0.173**	-0.010	0.708	**298**	301
ETTh1 (7) / 192	0.3404	**0.3359**	1.47	**1.37**	**0.157**	-0.005	0.744	**785**	798
ETTh2 (7) / 96	0.2650	**0.2595**	2.24	**1.45**	**0.099**	-0.010	0.597	**84**	135
ETTh2 (7) / 192	0.3064	**0.2985**	1.94	**1.76**	**0.089**	-0.005	0.662	**260**	396
ETTm1 (7) / 96	0.2958	**0.2820**	1.57	**1.46**	**0.227**	-0.011	0.788	239	**234**
ETTm1 (7) / 192	0.3163	**0.2997**	1.76	**1.33**	**0.200**	-0.005	0.823	626	**555**
ETTm2 (7) / 96	0.2030	**0.2006**	**1.48**	1.58	**0.129**	-0.010	0.709	**51**	63
ETTm2 (7) / 192	0.2369	**0.2339**	**1.41**	2.06	**0.097**	-0.005	0.746	**157**	209
Weather (21) / 96	0.1830	**0.1815**	**2.43**	2.70	**0.086**	-0.010	0.845	**223**	314
Weather (21) / 192	0.2149	**0.2133**	**2.69**	2.70	**0.160**	-0.005	0.880	**578**	853
Electricity (321) / 96	**0.1992**	0.2019	**2.56**	2.78	**0.567**	-0.010	0.611	**4141**	4638
Electricity (321) / 192	**0.2064**	0.2093	**2.43**	2.67	**0.670**	-0.005	0.656	**9241**	9925
Traffic (862) / 96	**0.2466**	0.2493	**2.51**	3.06	**0.782**	-0.010	0.592	**49138**	52540
Traffic (862) / 192	**0.2553**	0.2586	**2.74**	3.33	**0.797**	-0.005	0.602	**106778**	114030

(H=192 is the setting of their main table), S=100 samples, per-dataset hyperparameters taken from their released scripts, and three seeds for every cell reported. Our reproduction of PPM at its published setting gives CRPS 0.3404 ± 0.0001 on ETTh1/192 against 0.337 reported, a 0.9% discrepancy.

Marginal metrics: the floor wins. Table 1 and Figure 1a. The capacity-matched Gaussian floor improves CRPS over PPM by 0.8–5.2% in 10 of the 14 cells, over three seeds per cell; the largest seed standard deviation anywhere in the comparison is 0.0013, and the floors' is at most 0.0006. The exceptions are Electricity and Traffic at both horizons, where PPM leads by 1.1–1.4%. At the default lookback the pattern is ordered by series count (1.2–5.2% at D=7, 0.8% at D=21, negative at D=321 and 862), but a sweep over lookback $L \in \{96, 192, 336\}$ on three datasets (Table 5) shows the boundary is not a function of D alone: on the two 7-series datasets the floor's lead grows or holds with L, while on Weather it shrinks from 0.8% to 0.4% and turns into a 0.8% deficit at L=336. Which sampler wins the marginal metric is a regime question decided jointly by dimensionality and context length; we do not claim a universal ordering. The Student-t floor is dataset-dependent: it cuts QICE by 2.7–3.2× on ETTh2 (0.71 versus 2.24 at H=96) and helps on ETTh1/192, but is worse than PPM on the wider datasets, winning 5 of 14 cells overall. NO-LATENT — PPM's own decoder with the latent noise switched off — also beats PPM on CRPS in the same 10 of 14 cells as the Gaussian floor and loses the same four, so the stochastic push-forward contributes nothing that the marginal metrics can detect. QICE is split (6 of 14 cells to the floor), but PPM is markedly less stable across seeds than the floors: on ETTh2/192 its QICE has standard deviation ± 0.248 against the floor's ± 0.009, and its variogram score ± 19 against ± 3.

Dependence-sensitive metrics: the ordering flips. Table 1. The forecast errors are strongly autocorrelated on every benchmark: 0.59–0.88 at lag 1. The floor's residual paths have autocorrelation 0 by construction (the table shows -0.005 to -0.010, the finite-sample bias of subtracting a sample mean). PPM's residual paths are positively autocorrelated in *all* 14 cells — including the four where it also wins CRPS — and in all six lookback cells of Appendix D, 20 of 20; on the raw variogram score it wins 12 of 14. The sampler is adding real temporal structure that the benchmark cannot see.

How much, and where. The same diagnostic also shows how little. On the four ETT sets and Weather, PPM's residual autocorrelation is 0.09–0.23 against a true 0.60–0.88: it recovers 10–29% of the dependence in the forecast errors. On Electricity and Traffic it recovers 93–102% and 132% — on Traffic the residuals are *over*-correlated. These are exactly the four cells in which PPM also

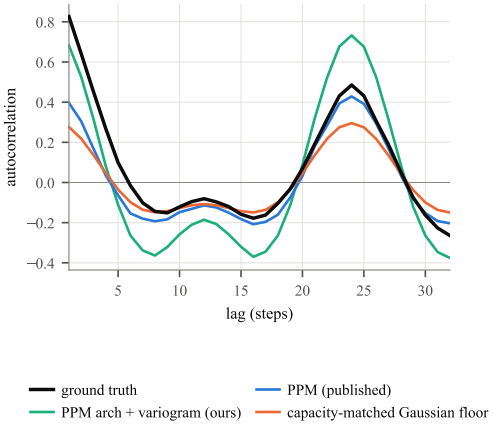


Figure 2: Autocorrelation of generated paths, ETTh1/192, computed with the same estimator as the reported metric on a 256-window test subset. The floor is a damped copy of the truth; the published model reproduces the shape with reduced amplitude; variogram training overshoots the diurnal peak.

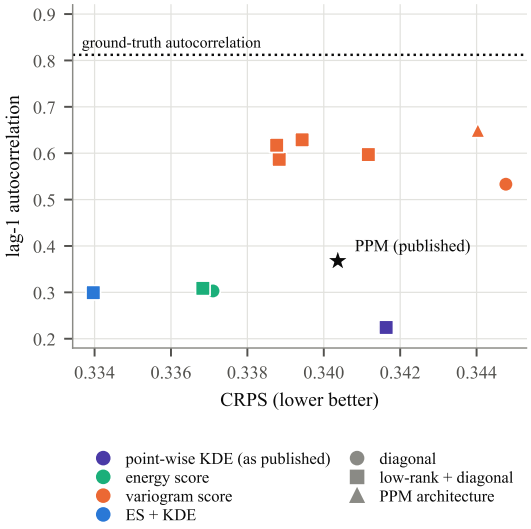


Figure 3: CRPS against lag-1 autocorrelation over all sampler/objective configurations, ETTh1/192. The published model is dominated: several configurations are better on both axes simultaneously.

wins CRPS. We claim no causal link to a marginal metric, but the coincidence is complete: the sampler wins the benchmark precisely where it has learned the structure the benchmark cannot see. The generative component is learning real temporal structure. None of it is visible in the left half of the table.

Sampled paths, for comparison. With the mean not removed the floor is far from zero (0.24 on ETTh1/192, 0.77 on Electricity), because every path inherits the conditional mean's structure — which is why the residual version is the one to read. Figure 2 extends the sampled-path autocorrelation to lag 32 and Figure 5 (Appendix D) shows the paths. Figure 2 extends this to lags up to 32: the floor is a uniformly damped copy of the truth, while PPM reproduces the shape—including the diurnal peak at lag 24—but with attenuated amplitude. Figure 5 shows the same fact in the sample paths.

A second model in the lineage. A natural objection is that PPM's sampler is a one-shot push-forward of a latent Gaussian, so its weak dependence might be an artefact of that particular design rather than of the objective. We therefore ran NsDiff (Ye et al., 2025) — an actual denoising diffusion model, and the predecessor PPM compares against — end to end from its release, using its own script, its own lookback (168) and its own data pipeline, with our diagnostics added to its evaluation loop. Over three seeds on ETTh1/192 it gives CRPS 0.4037 ± 0.0112 (against 0.417 as tabulated by Li et al., 2026), and lag-1 autocorrelation 0.214 ± 0.009 against a ground-truth value of 0.814 computed inside its own pipeline. Its sampled paths therefore carry *less* temporal dependence than our dependence-free floor (0.238). The two pipelines agree on the ground-truth autocorrelation to three decimals (0.812 versus 0.814), which is what makes this axis comparable across frameworks; the CRPS values are not, since the lookback and scaler differ, and we do not compare them.

An exact isolation of the diffusion sampler. NsDiff also permits a sharper experiment than the one we could run against PPM. It already contains a mean network F and a variance network G; its reverse diffusion is what turns those into samples. We can therefore hold the *trained model completely fixed* — same weights, same training, same checkpoint — and change only the read-out, replacing the reverse diffusion with direct sampling from $\mathcal{N}(F(\boldsymbol{x}), \operatorname{diag} G(\boldsymbol{x}))$, the diagonal Gaussian its own two networks already define. There is no separately trained baseline and hence no confound. Table 3 gives the result over the same three seeds. Running the diffusion changes CRPS

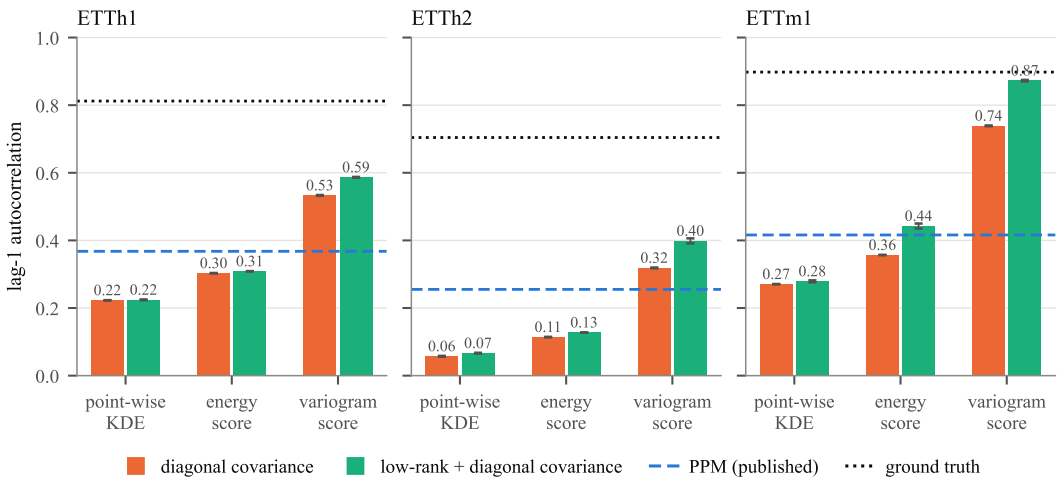


Figure 4: **The objective binds; the architecture does not.** Lag-1 autocorrelation at $H{=}192$, mean $\pm$ std over three seeds, on three datasets independently. Under the published point-wise objective (left pair in each panel) adding a horizon covariance is worth $+0.001$ to $+0.009$; changing the objective is worth $+0.393$ to $+0.608$. Once the objective scores the joint law, the covariance starts to pay.

by 0.6%, which is within the seed spread and not consistent in sign (2 of 3 seeds), and improves QICE slightly. It makes the variogram score *worse* in all three seeds (881 versus 839, $+5.0\%$) and lowers the sampled autocorrelation in all three (0.214 versus 0.235), moving it further from the ground-truth 0.814. On this model, the reverse diffusion buys no detectable marginal accuracy and actively degrades the joint structure relative to reading off the Gaussian the model has already fitted.

7 Structure or objective?

Two explanations are available for the missing dependence: the samplers may be unable to *represent* it, or the objective may not *reward* it. We separate them with a $2{\times}2$ design crossing sampler structure with training objective, replicated independently on ETTh1, ETTh2 and ETTm1 at $H{=}192$. The structure axis adds a low-rank plus diagonal horizon covariance,

$$\boldsymbol{y}^{(s)}_{:,d} = \boldsymbol{m}_d(\boldsymbol{h}) + \boldsymbol{L}_d(\boldsymbol{h})\,\boldsymbol{z}^{(s)}_r + \boldsymbol{s}_d(\boldsymbol{h}) \odot \boldsymbol{\epsilon}^{(s)}, \qquad \boldsymbol{L}_d(\boldsymbol{h}) \in \mathbb{R}^{H\times r}, \tag{5}$$

which can express temporal correlation. The objective axis replaces the point-wise KDE likelihood with the energy score or the variogram score of equation 2. Results are in Table 2 and Figure 4.

Structure alone is nearly inert, as predicted. Proposition 3 says this measurement should come out null. Under the published point-wise objective, adding the covariance of equation 5 moves lag-1 autocorrelation by $+0.001$ on ETTh1 ($0.223 \to 0.224$), $+0.009$ on ETTh2 ($0.057 \to 0.066$) and $+0.008$ on ETTm1 ($0.271 \to 0.279$). On ETTh1 that is within one seed standard deviation; on ETTh2 and ETTm1 it is a few standard deviations, so the idealised statement is not exactly attained — as the scope remark after Proposition 3 anticipates, our $\boldsymbol{L}$ is produced by a shared network rather than parameterised directly. The magnitude is nevertheless the point: changing the *objective* instead moves the same quantity by $+0.393$, $+0.401$ and $+0.608$ on the three datasets, i.e. $287\times$, $43\times$ and $73\times$ more than the structure does.

The objective moves everything. Swapping the point-wise likelihood for the variogram score raises autocorrelation at fixed architecture from 0.223 to 0.533 on ETTh1, 0.057 to 0.319 on ETTh2 and 0.271 to 0.739 on ETTm1 — past the published model's 0.368, 0.255 and 0.416 in every case — and lowers the variogram score by 32%, 59% and 58%.

They interact. Once the objective rewards dependence, the structure starts to matter, and by much more than it did under the point-wise objective: under variogram training the low-rank covariance improves autocorrelation from 0.533 to 0.587 on ETTh1 ($+0.054$ against $+0.001$), 0.319 to 0.399

Table 2: **Sampler structure × training objective** at $H=192$, on three datasets independently; means over three seeds. VS is the variogram score of equation 2 and AC the lag-1 autocorrelation of the sampled paths.

		ETTh1				ETTh2				ETTm1			
structure	objective	CRPS	QICE	VS	AC	CRPS	QICE	VS	AC	CRPS	QICE	VS	AC
PPM as published		0.3404	1.47	785	0.368	0.3064	1.94	260	0.255	0.3163	1.76	626	0.416
diagonal	point-wise KDE (as published)	0.3416	1.38	851	0.223	0.2998	1.76	397	0.057	0.3041	1.41	598	0.271
low-rank+diag.	point-wise KDE (as published)	0.3416	1.38	855	0.224	0.3000	1.76	399	0.066	0.3035	1.36	595	0.279
diagonal	energy score	0.3371	1.13	688	0.303	0.3002	1.43	213	0.114	0.3019	1.42	402	0.357
low-rank+diag.	energy score	0.3368	1.10	685	0.309	0.3001	1.42	210	0.128	0.3000	1.32	340	0.443
diagonal	variogram score	0.3448	3.56	580	0.533	0.3240	3.91	163	0.319	0.3194	4.62	253	0.739
low-rank+diag.	variogram score	0.3388	2.96	579	0.587	0.3208	3.70	161	0.399	0.3035	2.89	248	0.873
low-rank+diag.	variogram score (λ=5)	0.3388	2.77	574	0.617	0.3196	3.49	160	0.468	0.3024	2.66	246	0.887
low-rank+diag.	energy score + KDE	0.3340	0.98	682	0.299	0.2957	1.11	229	0.133	0.2963	1.20	320	0.484
PPM decoder	variogram score	0.3440	3.48	570	0.649	0.3279	4.51	164	0.501	0.3109	3.67	254	0.848
ground truth		–	–	–	0.812	–	–	–	0.704	–	–	–	0.898

on ETTh2 ($+0.080$ against $+0.009$) and 0.739 to 0.873 on ETTm1 ($+0.134$ against $+0.008$), while also improving CRPS on all three. This interaction is the signature of a representational constraint that only binds after the objective is fixed, and it is why the two axes must be crossed rather than tested separately.

8 A CHEAP FIX, AND ITS PRICE

Table 2 gives two operating points; Figure 3 shows the published model dominated by both.

Joint fidelity. Training PPM's own architecture with the variogram score recovers 80%, 71% and 94% of the ground-truth lag-1 autocorrelation on ETTh1, ETTh2 and ETTm1 (versus 45%, 36% and 46% as published) and improves the variogram score by 27–59%. The price is marginal calibration: QICE roughly doubles. That trade-off is real and we report it as such.

Marginal metrics. A hybrid that adds the energy score to the published likelihood dominates PPM on *all four* reported quantities at once on every dataset we ran it on (Table 2); on ETTh1, CRPS 0.3340 versus 0.3404, QICE 0.980 versus 1.469, energy score 18.33 versus 18.52, variogram score 682 versus 785. It does so with a diagonal-plus-low-rank Gaussian and no iterative sampling.

Reading these together. The variogram-trained model is not uniformly better, and Proposition 4 says why: the variogram score identifies only $\gamma^{(p)}$, so a minimiser can match lag-wise increments while distorting the correlation itself (Figure 2: it overshoots the diurnal peak, 0.73 against 0.48). What the two operating points establish is that the dependence deficit in this family is a choice of objective, not a capacity limit, and is reachable at no architectural cost.

9 LIMITATIONS AND CONCLUSION

Limitations. Seven benchmarks, two horizons, three seeds per cell, a lookback sweep on three; the mechanism study uses three datasets. The floor is run against PPM and NsDiff; D^3U cannot be reproduced from its release. Autocorrelation and the variogram score summarise the joint law rather than characterise it (ours mismatches at lag 24), and the residual diagnostic takes the sample mean as the conditional mean, so it absorbs any error in that mean.

Conclusion. The literature we examined scores marginals only. Under that score a sampler with no temporal dependence beats a state-of-the-art generative model in 10 of 14 cells; under a diagnostic isolating what the sampler adds beyond the mean, the ordering reverses in all 20 — yet the sampler recovers only a fraction of the dependence in the forecast errors, except on the two benchmarks where it also wins CRPS. The generative component does work its benchmark cannot see, and is not asked to do more: a covariance changes nothing until the objective scores the joint law. Report a dependence-sensitive score alongside CRPS, with a dependence-free baseline, so that a gain can be attributed to the component that produced it.

REPRODUCIBILITY STATEMENT

All experiments are run on publicly available benchmarks with the released code of the audited models, unmodified in any quantity that is reported. The protocol, per-dataset hyperparameters, seeds, and compute are given in Appendix B; the two code changes we made to the PPM release — a vectorised CRPS verified identical to the original to 1.3×10^{-15}, and removal of an unused $\sim$30 GB accumulation in the evaluation loop — are documented there together with our reproduction of the published number (0.9% discrepancy). The capacity-matching rule for the floor, including the horizon dependence that an earlier version of this comparison got wrong, is stated explicitly. Every table and figure is generated by script from the run logs, so all numbers can be regenerated from the log directory alone. The dependence diagnostics are defined in Section 3 and Appendix B with their exact estimator settings. Complete proofs of all propositions are in Appendix A. The reproducibility barriers we encountered in the other two releases of the lineage are recorded in Appendix C so that others do not repeat the same attempts. Code and run logs will be released.

REFERENCES

Alexandre Drouin, Étienne Marcotte, and Nicolas Chapados. TACTiS: Transformer-attentional copulas for time series. In *International Conference on Machine Learning*, 2022.

Xinyao Fan, Yueying Wu, Chang Xu, Yuhao Huang, Weiqing Liu, and Jiang Bian. MG-TSD: Multi-granularity time series diffusion models with guided learning process. In *International Conference on Learning Representations*, 2024.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Xingzhuo Guo, Yu Zhang, Baixu Chen, Haoran Xu, Jianmin Wang, and Mingsheng Long. Dynamical diffusion: Learning temporal dynamics with diffusion models. In *International Conference on Learning Representations*, 2025.

Haiqi Jiang and Hui Xiong. KineFlow: Kinematic second-order flow matching for time-series forecasting. In *International Conference on Machine Learning*, 2026.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with gaussian process priors for probabilistic time series forecasting. In *International Conference on Learning Representations*, 2025.

Alireza Koochali, Peter Schichtel, Andreas Dengel, and Sheraz Ahmed. Random noise vs. state-of-the-art probabilistic forecasting methods: A case study on CRPS-Sum discrimination ability. *Applied Sciences*, 12(10):5104, 2022.

Seunghan Lee, Kibok Lee, and Taeyoung Park. ANT: Adaptive noise schedule for time series diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Jinglin Li, Jun Tan, Qi Fang, and Ning Gui. Parametric prior mapping framework for non-stationary probabilistic time series forecasting. In *International Conference on Machine Learning*, 2026.

Qi Li, Zhenyu Zhang, Lei Yao, Zhaoxia Li, Tianyi Zhong, and Yong Zhang. Diffusion-based decoupled deterministic and uncertain framework for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, 2025.

Yuxin Li, Wenchao Chen, Xinyue Hu, Bo Chen, Baolin Sun, and Mingyuan Zhou. Transformer-modulated diffusion models for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, 2024.

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

Étienne Marcotte, Valentina Zantedeschi, Alexandre Drouin, and Nicolas Chapados. Regions of reliability in the evaluation of multivariate probabilistic forecasts. In *International Conference on Machine Learning*, 2023.

Pierre Pinson and Julija Tastu. Discrimination ability of the energy score. Technical Report 2013-15, DTU Compute, Technical University of Denmark, 2013.

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

Michael Scheuerer and Thomas M Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.

Lifeng Shen, Weiyu Chen, and James T Kwok. Multi-resolution diffusion models for time series forecasting. In *International Conference on Learning Representations*, 2024.

Edgardo Solano-Carrillo, Sreerag V Naveenachandran, and Julia Niebling. TEDM: Time series forecasting with elucidated diffusion models. In *International Conference on Learning Representations*, 2026.

Chenxi Wang, Linxiao Yang, Zhixian Wang, Liang Sun, and Yi Wang. A non-isotropic time series diffusion model with moving average transitions. In *International Conference on Machine Learning*, 2025.

Xingjian Wu, Xiangfei Qiu, Hongfan Gao, Jilin Hu, Bin Yang, and Chenjuan Guo. K^2VAE: A Koopman-Kalman enhanced variational AutoEncoder for probabilistic time series forecasting. In *International Conference on Machine Learning*, 2025.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. In *International Conference on Learning Representations*, 2026.

Weiwei Ye, Zhuopeng Xu, and Ning Gui. Non-stationary diffusion for probabilistic time series forecasting. In *International Conference on Machine Learning*, 2025.

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

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, 2021.

A Proofs

Proof of Proposition 1

Let $Q \in \mathcal{P}(\mathbb{R}^n)$ be the true law and P a forecast. Since $S(P, \boldsymbol{y}) = \sum_i s_i(P_i, y_i)$ and each summand depends on $\boldsymbol{y}$ only through y_i, linearity of expectation gives

$$\mathbb{E}_{\boldsymbol{y}\sim Q}\, S(P, \boldsymbol{y}) = \sum_{i=1}^{n} \mathbb{E}_{y_i \sim Q_i}\, s_i(P_i, y_i). \tag{6}$$

The right-hand side depends on P only through $(P_1, \ldots, P_n)$. Each term is minimised over P_i uniquely at $P_i = Q_i$ because s_i is strictly proper, and the terms are decoupled, so the sum is minimised exactly when $P_i = Q_i$ for all i, i.e. on $\mathcal{F}(Q)$. This establishes properness (the minimum is attained at $P = Q \in \mathcal{F}(Q)$) and shows the minimiser set is all of $\mathcal{F}(Q)$, which contains $\bigotimes_i Q_i$ and every other coupling of the Q_i. Since equation 6 is constant on $\mathcal{F}(Q)$, no two of its elements are separated. □

Remark 6. The argument uses only that the score decomposes over coordinates; it is exact in expectation and does not depend on the number of samples S. The finite-sample deficiencies of multivariate scores studied by Marcotte et al. (2023) are a separate phenomenon: they concern the *power* of a strictly proper score at finite test-set size, whereas Proposition 1 is an identifiability statement that survives infinite data.

Proof of Corollary 2

Given $\boldsymbol{R} \in \mathcal{E}_n$ let $\boldsymbol{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ and $P = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{DRD})$. Then P is Gaussian with $\mathrm{Var}(X_i) = \sigma_i^2$ and mean μ_i, so $P_i = Q_i$ and $P \in \mathcal{F}(Q)$, while $\mathrm{Corr}(P) = \boldsymbol{R}$ by construction. Taking $\boldsymbol{R}$ with $R_{i,i+\ell} = \rho$ for any $\rho \in [-1, 1]$ admissible in $\mathcal{E}_n$ — for instance the AR(1) correlation matrix $R_{ij} = \rho^{|i-j|}$, which is positive definite for all $\rho \in (-1, 1)$ — gives forecasters with lag-ℓ autocorrelation ρ^ℓ, all with identical expected marginal score. □

Proof of Proposition 3

Let $\boldsymbol{y} = \boldsymbol{m} + \boldsymbol{L}\boldsymbol{z}_r + \boldsymbol{s} \odot \boldsymbol{\epsilon}$ with $\boldsymbol{z}_r \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_H)$ independent. Then $\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{L}\boldsymbol{L}^\top + \mathrm{diag}(\boldsymbol{s}^2)$ and the h-th marginal is $\mathcal{N}(m_h,\ \|\boldsymbol{L}_{h\cdot}\|^2 + s_h^2)$, where $\boldsymbol{L}_{h\cdot} \in \mathbb{R}^r$ is row h. Write $v_h = \|\boldsymbol{L}_{h\cdot}\|^2 + s_h^2$.

(i) The marginal collection induced by $(\boldsymbol{m}, \boldsymbol{L}, \boldsymbol{s})$ is $(\mathcal{N}(m_h, v_h))_h$. As $(\boldsymbol{L}, \boldsymbol{s})$ ranges over $\mathbb{R}^{H\times r} \times \mathbb{R}_+^H$, v_h ranges over all of $\mathbb{R}_+$ independently across h (take $\boldsymbol{L} = \boldsymbol{0}$ and $s_h^2 = v_h$). Conversely every $\mathcal{F}_{\mathrm{diag}}$ element has v_h realisable in $\mathcal{F}_{\mathrm{lr}}$. The two induced sets of marginal collections therefore coincide, and since $\mathcal{L}(P) = \sum_h \ell_h(P_h)$ depends on P only through them, the infima coincide.

(ii) $\mathcal{L}$ is a function of $(\boldsymbol{m}, v_1, \ldots, v_H)$ alone, and v_h depends on $\boldsymbol{L}$ only through the row norms $\|\boldsymbol{L}_{h\cdot}\|$. Hence any $\boldsymbol{L}'$ with the same row norms gives the same objective value. Row sign flips $\boldsymbol{L}_{h\cdot} \mapsto -\boldsymbol{L}_{h\cdot}$ preserve row norms; for $H \geq 2$ and $r \geq 1$ they change $(\boldsymbol{L}\boldsymbol{L}^\top)_{hk} = \langle \boldsymbol{L}_{h\cdot}, \boldsymbol{L}_{k\cdot} \rangle$ in sign for $k \neq h$ whenever that inner product is nonzero, so the set contains distinct covariances.

(iii) By the chain rule, $\nabla_{\boldsymbol{L}_{h\cdot}} \mathcal{L} = \ell_h'(v_h)\, \nabla_{\boldsymbol{L}_{h\cdot}} v_h = 2\, \ell_h'(v_h)\, \boldsymbol{L}_{h\cdot}$, which is parallel to $\boldsymbol{L}_{h\cdot}$. The gradient flow $\dot{\boldsymbol{L}}_{h\cdot} = -2\ell_h'(v_h)\boldsymbol{L}_{h\cdot}$ is a time-dependent scalar rescaling, so $\boldsymbol{L}_{h\cdot}(t) = c_h(t)\, \boldsymbol{L}_{h\cdot}(0)$ with $c_h(t) > 0$ as long as c_h does not cross zero. Consequently $\mathrm{Corr}(y_h, y_k)$ depends on t only through the scalars c_h, c_k and the fixed initial directions; in particular $\mathrm{sign}\,\mathrm{Cov}(y_h, y_k)$ and the angles $\angle(\boldsymbol{L}_{h\cdot}(0), \boldsymbol{L}_{k\cdot}(0))$ are invariant. □

Remark 7 (Scope). Proposition 3 is stated for a fixed context and a directly parameterised $(\boldsymbol{m}, \boldsymbol{L}, \boldsymbol{s})$. In our implementation $\boldsymbol{L}$ is the output of a shared network, so rows for different contexts are coupled through the weights and the invariance holds only approximately. The measured effect (Section 7) is nevertheless null to within one seed standard deviation on all four metrics, which is what the idealised statement predicts.

Proof of Proposition 4

Strict properness of the energy score $\mathrm{ES}(P, \boldsymbol{y}) = \mathbb{E}_P\|\boldsymbol{X} - \boldsymbol{y}\| - \frac{1}{2}\mathbb{E}_P\|\boldsymbol{X} - \boldsymbol{X}'\|$ on distributions with finite first moment is standard (Gneiting & Raftery, 2007, Sec. 4.3). For the variogram score, write $\gamma_{ij}^{(p)}(P) = \mathbb{E}_P|X_i - X_j|^p$. From equation 2, $\mathrm{VS}(P, \boldsymbol{y}) = \sum_{ij} w_{ij}\big(|y_i - y_j|^p - \gamma_{ij}^{(p)}(P)\big)^2$, so $\mathbb{E}_{\boldsymbol{y}\sim Q}\mathrm{VS}(P, \boldsymbol{y}) = \sum_{ij} w_{ij}\big[\mathrm{Var}_Q(|Y_i - Y_j|^p) + (\gamma_{ij}^{(p)}(Q) - \gamma_{ij}^{(p)}(P))^2\big]$, a sum of a P-independent term and a squared bias. It is therefore minimised precisely by those P with $\gamma^{(p)}(P) = \gamma^{(p)}(Q)$: the score is proper, and its minimiser set is the (generally large) set of laws sharing the variogram, so it is not strictly proper. □

B Experimental details

Data and protocol. We use all seven benchmarks of Li et al. (2026) — ETTh1, ETTh2, ETTm1, ETTm2 (Zhou et al., 2021), Weather, Electricity and Traffic — with the split, instance normalisation and window construction of the released PPM code, unchanged. As in that code, the three large-channel datasets use a test-window stride equal to the prediction length. Context length is 96; horizons are $H \in \{96, 192\}$, where H=192 is the setting of the main table of Li et al. (2026). All seven channels are forecast jointly (`features=M`). Test windows use stride 1. The predictive distribution is represented by $S = 100$ samples at both training and evaluation time, as in the released configuration.

Optimisation. Every run trains for 30 epochs with Adam, `lradj=type1`, and early stopping on validation loss with patience 5. Learning rate, batch size, d_{model} and d_{ff} are taken per dataset

from the released PPM scripts (10^{-4}/64 for ETTh, 10^{-4}/128 for ETTm and Weather, $5 \cdot 10^{-4}$/16 for Electricity, 10^{-3}/16 for Traffic; d_{model}=512, d_{ff}=1024 on Weather and 256/256 elsewhere). The loss is $\tau \cdot \text{KDE} + \text{MSE}$ with $\tau = 0.1$ and KDE bandwidth 0.3, the values hard-coded in the released scripts. Floors and the structure/objective grid use identical settings; only the term named in the table changes.

Architectures. PPM: encoder $\text{Linear}(96 \to 256)$ with GELU, two heads producing a 256-dimensional latent mean and log-variance, a reparameterised sample, and a decoder $\text{Linear}(512 \to 256) \to \text{GELU} \to \text{Linear}(256 \to H)$ applied to the latent concatenated with a context embedding (337k parameters). GAUSS: the same encoder and a two-layer head of equal depth emitting $(\boldsymbol{m}, \log \boldsymbol{s})$ (140k parameters at d_{ff}=256). $\text{GAUSS}_{\text{cap}}$: identical but with d_{ff} solved per (L, H, d_{model}) so that the floor's parameter count matches PPM's to within 0.1%. Because PPM's decoder output width is H while the floor's is $2H$, the matching width depends on the horizon: 695 at H=96 and 525 at H=192 for the 256/256 configuration, 2444 and 2030 for Weather. Using a single width for both horizons, as an earlier version of this comparison did, over-parameterises the floor by 31% at H=192 and inflates its apparent advantage by roughly 0.7 percentage points. STUDENT-t: as GAUSS with unit-variance-rescaled t_5 noise. NO-LATENT: PPM's own decoder with the latent noise set to zero plus a learned per-horizon scale head. JOINT: the low-rank+diagonal sampler of equation 5, rank $r = 8$ unless stated.

Training objectives. The energy score loss uses a random subsample of 16 of the 100 paths per step, computed per channel over the horizon and normalised by $\sqrt{H}$; the variogram loss uses all paths with $p = 1$ and $\mathcal{L} = \{1, 2, 4, 8, 16, 32\}$. The joint-score weight λ (τ_j in the code) defaults to 1; Table 2 reports $\lambda \in \{1, 5, 20\}$ and rank $\in \{8, 32\}$.

Metric implementations. CRPS and QICE are computed by the released PPM evaluation code, unmodified in definition. We replaced its per-element Python loop over a third-party CRPS implementation with an equivalent vectorised closed form, having first verified agreement to 1.3×10^{-15} (maximum absolute difference over 400 randomly generated ensembles at $S \in \{10, 100\}$); this was necessary because the original performs $\sim$ 3.6M scalar calls per evaluation. The energy score is estimated from 20 of the 100 paths (the $O(S^2)$ term dominates evaluation cost); the variogram score and autocorrelation use 50. All models are compared under identical estimator settings.

Autocorrelation estimator. For samples $\boldsymbol{X}$ of shape (B, S, H, D) we centre along the horizon axis and report $\frac{1}{BSD} \sum_{b,s,d} \hat{a}_\ell / \hat{v}$ with $\hat{a}_\ell = \text{mean}_h \, \tilde{X}_{h+\ell} \tilde{X}_h$ and $\hat{v} = \text{mean}_h \, \tilde{X}_h^2$, i.e. a per-series ratio averaged last. The ground-truth column applies the same estimator to the observed futures. Figure 2 uses this estimator on a 256-window test subset, which is why its lag-1 values differ slightly from the full-test-set numbers in Table 7.

Compute and seeds. All runs use a single RTX A6000 with `OMP_NUM_THREADS=2`. A run costs 4–11 minutes on ETTh and 20–40 minutes on ETTm. All PPM/floor cells in Table 1 and every configuration in Table 2 use three seeds $\{2021, 2022, 2023\}$; a few secondary architectures (STUDENT-t, NO-LATENT) use fewer on some cells, which each table reports implicitly through the presence or absence of a standard deviation. Runs that failed from a transient out-of-memory condition caused by other users on the shared GPU were retried automatically.

Reproduction fidelity and one code fix. Our PPM reproduction gives CRPS 0.3404 ± 0.0001 on ETTh1/192 against 0.337 reported, a 0.9% discrepancy at the published single-seed setting. The released evaluation loop accumulates every test prediction in memory before discarding the arrays unused; on ETTm this costs $\sim$ 30 GB per process and prevents parallel evaluation. We removed the accumulation, which changes no reported quantity.

C THE REPORTED METRIC ACROSS THE LINEAGE

We inspected the released evaluation code of three members of the lineage. In each, the reported CRPS is a function of the predictive marginals alone, in the sense of Proposition 1.

model	how the reported CRPS is computed	marginal?
PPM (Li et al., 2026)	`(B,S,L,N)` flattened to `(-1,S)`, univariate CRPS per row	yes
NsDiff (Ye et al., 2025)	`pred.view(-1, pred.shape[3])`, univariate CRPS per row	yes
D^{3}U (Li et al., 2025)	per-element quantiles over the sample axis, pinball loss summed	yes

D^{3}U additionally reports CRPS-Sum, the statistic whose separate deficiencies are documented by Koochali et al. (2022) and Marcotte et al. (2023); PPM and NsDiff report only the per-element form, to which those critiques do not apply and ours does. QICE is a coverage histogram over the same flattened elements in all three.

Reproducibility barriers encountered. We attempted to extend the floor comparison to the other two models and record what stopped us, since both issues are properties of the releases rather than of the methods.

- **D^{3}U** requires a pre-trained point-forecasting checkpoint that is not distributed: the repository's `pretrain_checkpoints` entry is a text file reading "Place the checkpoint of the already trained point prediction model into this folder," and no script for training that model is included. The framework cannot be run end to end from the release.
- **NsDiff** ships a pretraining script for its F network whose default seeds are $\{10, 2, 3\}$, while the main experiment loads that network from a hard-coded seed directory `1`; and its G (variance) pretrainer defaults to `rolling_length=24` while the main experiment defaults to 96, so the pretrained G has input width $168-24 = 144$ where the main model expects $168-96 = 72$ and the checkpoint cannot be loaded. The released `scripts/NSDiff/*.sh` sidestep both by passing `--load_pretrain=False`, which is the path we use.

Table 3: **Isolating the reverse diffusion in NsDiff.** ETTh1/H=192, mean ± std over three seeds. Both rows use *the same trained weights*; they differ only in whether the predictive samples come from the reverse diffusion or directly from the Gaussian defined by NsDiff's own mean and variance networks. Arrows give the direction of improvement; the last column is the number of seeds in which the diffusion is better.

predictive distribution	CRPS ↓	QICE (%) ↓	ES ↓	VS ↓	lag-1 AC
reverse diffusion (as published)	**0.4037** ±0.0112	**2.055** ±0.449	**21.74** ±0.83	881 ±48	0.214 ±0.009
$\mathcal{N}(F(\boldsymbol{x}), \operatorname{diag} G(\boldsymbol{x}))$, same weights	0.4061 ±0.0074	2.199 ±0.209	21.82 ±0.65	**839** ±37	**0.235** ±0.009
seeds where diffusion wins	2/3	3/3	3/3	0/3	0/3
ground truth	–	–	–	–	0.814

D RESIDUAL DEPENDENCE, ALL CELLS

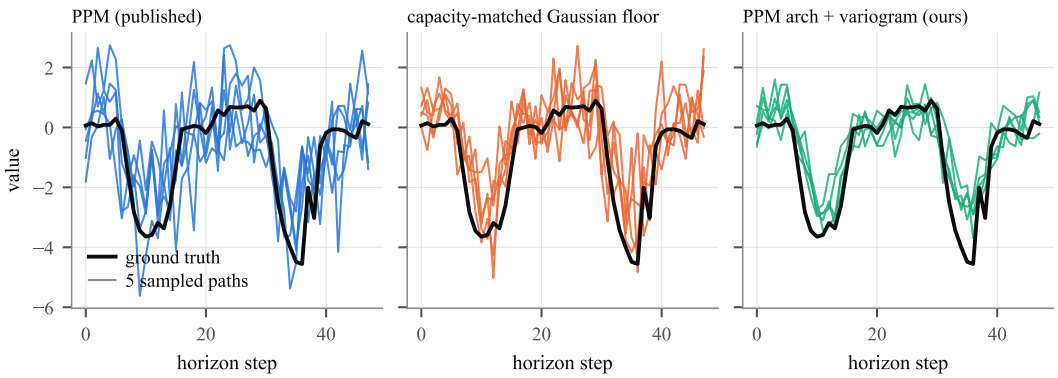


Figure 5: **Five sampled futures per model** (first 48 of 192 steps, one channel, ETTh1). The floor's paths oscillate around the conditional mean with white noise; the published model is wider but similarly rough; training the same architecture with a variogram score produces temporally coherent paths.

Table 4 gives the residual (mean-removed) lag-1 autocorrelation for every cell in the paper, next to the sampled-path version, so the effect of removing the conditional mean can be read directly. The residual statistic is computed from the same $S{=}100$ samples as the reported metrics: for each test window, $\bar{\boldsymbol{y}} = \frac{1}{S}\sum_s \boldsymbol{y}^{(s)}$, residual paths are $\boldsymbol{y}^{(s)} - \bar{\boldsymbol{y}}$, the forecast error is $\boldsymbol{y} - \bar{\boldsymbol{y}}$, and the lag-1 autocorrelation estimator of Appendix B is applied to each. Using the model's own sample mean as the conditional-mean proxy is the standard residual diagnostic and treats both samplers identically. Subtracting a sample mean of S draws induces a small negative bias of order $-1/S$ in the residual autocorrelation of an independent sampler; the floor's values of -0.005 to -0.010 are that bias, and are the estimator's zero point.

Table 4: **Sampled-path versus residual autocorrelation**, all 20 cells, means over three seeds. The residual column for the floor is zero by construction (up to the $-1/S$ estimator bias). PPM's residual autocorrelation is positive in every row; it approaches the true forecast-error autocorrelation only on Electricity and Traffic.

		sampled-path lag-1 AC			residual lag-1 AC			resid. VS: PPM / floor
dataset (D) / H	L	PPM	floor	truth	PPM	floor	truth	
ETTh1 (7) / 96	96	0.433	0.300	0.793	**0.173**	-0.010	0.708	204 / 204
ETTh1 (7) / 192	96	0.368	0.238	0.812	**0.157**	-0.005	0.744	555 / 558
ETTh2 (7) / 96	96	0.317	0.085	0.650	**0.099**	-0.010	0.597	73 / 119
ETTh2 (7) / 192	96	0.255	0.064	0.704	**0.089**	-0.005	0.662	230 / 353
ETTm1 (7) / 96	96	0.465	0.313	0.874	**0.227**	-0.011	0.788	212 / 211
ETTm1 (7) / 192	96	0.416	0.284	0.898	**0.200**	-0.005	0.823	553 / 491
ETTm2 (7) / 96	96	0.295	0.145	0.746	**0.129**	-0.010	0.709	45 / 55
ETTm2 (7) / 192	96	0.202	0.085	0.786	**0.097**	-0.005	0.746	138 / 184
Weather (21) / 96	96	0.290	0.211	0.851	**0.086**	-0.010	0.845	195 / 268
Weather (21) / 192	96	0.301	0.177	0.884	**0.160**	-0.005	0.880	496 / 723
Electricity (321) / 96	96	0.865	0.763	0.857	**0.567**	-0.010	0.611	2846 / 3145
Electricity (321) / 192	96	0.877	0.769	0.866	**0.670**	-0.005	0.656	6325 / 6751
Traffic (862) / 96	96	0.915	0.676	0.846	**0.782**	-0.010	0.592	37720 / 37702
Traffic (862) / 192	96	0.912	0.685	0.840	**0.797**	-0.005	0.602	84305 / 82984
ETTh1 (7) / 192	192	0.368	0.272	0.812	**0.113**	-0.005	0.740	501 / 464
ETTh1 (7) / 192	336	0.385	0.301	0.812	**0.103**	-0.005	0.736	447 / 385
ETTh2 (7) / 192	192	0.269	0.084	0.704	**0.082**	-0.005	0.656	322 / 350
ETTh2 (7) / 192	336	0.257	0.095	0.704	**0.076**	-0.005	0.654	274 / 322
Weather (21) / 192	192	0.307	0.173	0.884	**0.143**	-0.005	0.874	468 / 708
Weather (21) / 192	336	0.342	0.203	0.884	**0.156**	-0.005	0.877	448 / 710

E LOOKBACK SENSITIVITY

Table 5 repeats the PPM-versus-floor comparison at $H{=}192$ for lookbacks $L \in \{96, 192, 336\}$ on three datasets, three seeds each, with the floor's head width re-solved for each (L, H) so that its parameter count matches PPM's. The CRPS ordering moves with L and differs across datasets; the autocorrelation ordering does not.

F FULL RESULT TABLES

Table 6 gives the marginal metrics for all five samplers with per-seed standard deviations, and Table 7 the dependence-sensitive metrics.

G ADDITIONAL OBSERVATIONS

Seed stability. The floors are markedly more stable across seeds than PPM on ETTh2/192: QICE ± 0.009 versus ± 0.248, variogram score ± 3 versus ± 19. Single-seed reporting, as in the released scripts, is therefore more fragile for the generative model than for the baselines it is compared against.

Channel independence. PPM's encoder acts along the time axis with weights shared across channels, so the model factorises over channels given the context and cannot represent cross-series de-

Table 5: **Lookback sweep**, H=192, means over three seeds. Bold marks the better CRPS and the autocorrelation closer to the truth. PPM's autocorrelation is higher in every row; its CRPS is better only on Weather at L=336.

		CRPS ↓			lag-1 autocorrelation			
dataset (D)	L	PPM	floor	floor adv.	PPM	floor	truth	VS: PPM / floor
ETTh1 (7)	96	0.3404	**0.3359**	−1.3%	**0.368**	0.238	0.812	785 / 798
	192	0.3365	**0.3283**	−2.4%	**0.368**	0.272	0.812	755 / 724
	336	0.3328	**0.3204**	−3.7%	**0.385**	0.301	0.812	705 / 646
ETTh2 (7)	96	0.3064	**0.2985**	−2.6%	**0.255**	0.064	0.704	260 / 396
	192	0.3090	**0.2974**	−3.8%	**0.269**	0.084	0.704	365 / 397
	336	0.3016	**0.2948**	−2.2%	**0.257**	0.095	0.704	314 / 367
Weather (21)	96	0.2149	**0.2133**	−0.8%	**0.301**	0.177	0.884	578 / 853
	192	0.1969	**0.1962**	−0.4%	**0.307**	0.173	0.884	533 / 825
	336	**0.1868**	0.1883	+0.8%	**0.342**	0.203	0.884	509 / 833

Table 6: Marginal metrics, all samplers. Mean ± std over three seeds; best per row in bold.

metric	dataset (D) / H	PPM	Gauss	Gauss$_{\text{cap}}$	Student-t	no-latent
CRPS	ETTh1 (7) / 96	0.3132 ±0.0001	0.3143 ±0.0005	0.3085 ±0.0003	0.3069 ±0.0002	**0.3060 ±0.0002**
	ETTh1 (7) / 192	0.3404 ±0.0001	0.3416 ±0.0002	0.3359 ±0.0002	0.3330 ±0.0002	**0.3330 ±0.0003**
	ETTh2 (7) / 96	0.2650 ±0.0002	0.2603 ±0.0002	0.2595 ±0.0001	**0.2583 ±0.0001**	0.2590 ±0.0001
	ETTh2 (7) / 192	0.3064 ±0.0013	0.2998 ±0.0002	0.2985 ±0.0002	**0.2968 ±0.0002**	0.2982 ±0.0004
	ETTm1 (7) / 96	0.2958 ±0.0001	0.2864 ±0.0002	0.2820 ±0.0002	**0.2796 ±0.0002**	0.2819 ±0.0003
	ETTm1 (7) / 192	0.3163 ±0.0006	0.3041 ±0.0001	0.2997 ±0.0006	**0.2970 ±0.0005**	0.2975 ±0.0001
	ETTm2 (7) / 96	0.2030 ±0.0003	0.2014 ±0.0002	0.2006 ±0.0003	0.2007 ±0.0004	**0.1996 ±0.0004**
	ETTm2 (7) / 192	0.2369 ±0.0004	0.2353 ±0.0005	0.2339 ±0.0000	**0.2333 ±0.0002**	0.2336 ±0.0001
	Weather (21) / 96	0.1830 ±0.0002	0.1818 ±0.0002	0.1815 ±0.0004	0.1815 ±0.0004	**0.1792 ±0.0002**
	Weather (21) / 192	0.2149 ±0.0004	0.2132 ±0.0002	0.2133 ±0.0003	0.2131 ±0.0003	**0.2110 ±0.0001**
	Electricity (321) / 96	**0.1992 ±0.0001**	0.2039 ±0.0002	0.2019 ±0.0001	0.2017 ±0.0001	0.2028 ±0.0002
	Electricity (321) / 192	**0.2064 ±0.0002**	0.2103 ±0.0001	0.2093 ±0.0002	0.2091 ±0.0002	0.2099 ±0.0002
	Traffic (862) / 96	**0.2466 ±0.0002**	0.2553 ±0.0002	0.2493 ±0.0001	0.2467 ±0.0001	0.2506 ±0.0001
	Traffic (862) / 192	**0.2553 ±0.0002**	0.2623 ±0.0000	0.2586 ±0.0004	0.2559 ±0.0004	0.2581 ±0.0004
QICE (%)	ETTh1 (7) / 96	1.216 ±0.011	**1.006 ±0.016**	1.049 ±0.011	1.452 ±0.012	1.051 ±0.009
	ETTh1 (7) / 192	1.469 ±0.007	1.376 ±0.012	1.369 ±0.012	**1.030 ±0.003**	1.427 ±0.007
	ETTh2 (7) / 96	2.236 ±0.072	1.432 ±0.021	1.454 ±0.051	**0.706 ±0.017**	1.529 ±0.005
	ETTh2 (7) / 192	1.943 ±0.248	1.761 ±0.016	1.759 ±0.009	**0.715 ±0.020**	1.880 ±0.026
	ETTm1 (7) / 96	1.573 ±0.005	**1.343 ±0.014**	1.456 ±0.011	2.003 ±0.011	1.416 ±0.039
	ETTm1 (7) / 192	1.756 ±0.041	1.414 ±0.008	1.328 ±0.023	1.556 ±0.025	**1.270 ±0.005**
	ETTm2 (7) / 96	**1.480 ±0.036**	1.538 ±0.024	1.579 ±0.051	1.788 ±0.074	1.623 ±0.040
	ETTm2 (7) / 192	1.407 ±0.093	1.918 ±0.090	2.062 ±0.023	**1.333 ±0.016**	2.183 ±0.026
	Weather (21) / 96	**2.425 ±0.061**	2.576 ±0.031	2.696 ±0.034	3.459 ±0.061	2.454 ±0.040
	Weather (21) / 192	2.690 ±0.014	2.631 ±0.030	2.703 ±0.006	3.150 ±0.023	**2.589 ±0.020**
	Electricity (321) / 96	**2.556 ±0.080**	2.607 ±0.022	2.780 ±0.021	3.601 ±0.020	3.147 ±0.023
	Electricity (321) / 192	**2.432 ±0.029**	2.539 ±0.010	2.675 ±0.051	3.504 ±0.054	3.009 ±0.020
	Traffic (862) / 96	**2.510 ±0.050**	3.109 ±0.030	3.059 ±0.053	2.771 ±0.050	3.349 ±0.040
	Traffic (862) / 192	**2.740 ±0.049**	3.137 ±0.029	3.329 ±0.045	2.911 ±0.046	3.489 ±0.021

pendence, despite being evaluated as a multivariate method. All dependence diagnostics we report are therefore temporal; a cross-channel analogue would be measuring a quantity the architecture cannot express.

Why the Student-t floor calibrates so well. The largest QICE improvements from the Student-t head occur on ETTh2, the dataset with the heaviest-tailed residuals in the suite. This is consistent with QICE measuring quantile coverage: a Gaussian head misallocates mass between the shoulders and the tails, and the coverage histogram registers it, while CRPS — an integral of squared CDF error — is comparatively insensitive to that reallocation.

Table 7: Dependence-sensitive metrics, same runs.

metric	dataset (D) / H	PPM	$\text{Gauss}_{\text{cap}}$	no-latent	ground truth
lag-1 autocorr.	ETTh1 (7) / 96	**0.433** ±0.009	0.300 ±0.001	0.306 ±0.000	0.793
	ETTh1 (7) / 192	**0.368** ±0.003	0.238 ±0.000	0.248 ±0.001	0.812
	ETTh2 (7) / 96	**0.317** ±0.009	0.085 ±0.002	0.089 ±0.000	0.650
	ETTh2 (7) / 192	**0.255** ±0.013	0.064 ±0.001	0.068 ±0.001	0.704
	ETTm1 (7) / 96	**0.465** ±0.013	0.313 ±0.001	0.311 ±0.001	0.874
	ETTm1 (7) / 192	**0.416** ±0.009	0.284 ±0.002	0.288 ±0.000	0.898
	ETTm2 (7) / 96	**0.295** ±0.013	0.145 ±0.010	0.149 ±0.011	0.746
	ETTm2 (7) / 192	**0.202** ±0.008	0.085 ±0.001	0.084 ±0.001	0.786
	Weather (21) / 96	**0.290** ±0.006	0.211 ±0.002	0.217 ±0.000	0.851
	Weather (21) / 192	**0.301** ±0.007	0.177 ±0.002	0.183 ±0.003	0.884
	Electricity (321) / 96	**0.865** ±0.002	0.763 ±0.001	0.762 ±0.001	0.857
	Electricity (321) / 192	**0.877** ±0.002	0.769 ±0.001	0.767 ±0.001	0.866
	Traffic (862) / 96	**0.915** ±0.000	0.676 ±0.002	0.670 ±0.001	0.846
	Traffic (862) / 192	**0.912** ±0.001	0.685 ±0.002	0.678 ±0.000	0.840
variogram score	ETTh1 (7) / 96	298	301	**289**	–
	ETTh1 (7) / 192	785	798	**748**	–
	ETTh2 (7) / 96	**84**	135	135	–
	ETTh2 (7) / 192	**260**	396	406	–
	ETTm1 (7) / 96	239	234	**233**	–
	ETTm1 (7) / 192	626	555	**527**	–
	ETTm2 (7) / 96	**51**	63	64	–
	ETTm2 (7) / 192	**157**	209	212	–
	Weather (21) / 96	**223**	314	319	–
	Weather (21) / 192	**578**	853	876	–
	Electricity (321) / 96	**4141**	4638	4601	–
	Electricity (321) / 192	**9241**	9925	9848	–
	Traffic (862) / 96	**49138**	52540	52559	–
	Traffic (862) / 192	**106778**	114030	113501	–

H AI USE STATEMENT

Parts of this paper's text, code and analysis were produced with the assistance of generative AI. The authors take full responsibility for the final content of this paper, including any text produced with the assistance of generative AI.