NEURAL CONDITIONAL SIMULATION FOR PDE REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Refining a cheap coarse PDE solution to high fidelity is a conditional simulation problem, not a regression one. Under squared-error loss the optimal predictor is the conditional mean, which by the law of total variance discards all conditional variability and so cannot serve as a realization of the fine field. For Gaussian fields the answer is classical. Kriging supplies the conditional mean, and Matheron's construction adds an innovation drawn from the Schur-complement covariance to produce an exact sample in one pass. The fields that matter in practice, such as turbulence, stochastically forced dynamics, and precipitation, break these assumptions with innovations that are input-dependent, heteroscedastic, and heavy-tailed. We introduce *Neural Conditional Simulation* (NCS), which keeps the classical recipe but learns the innovation. NCS is initialized at the exact Gaussian simulator, trained end to end with strictly proper scoring rules, and samples through a coarse-to-fine cascade at one network pass per level. Because the Gaussian simulator is embedded as a zero-correction fallback and reported as a baseline, every result isolates the *nonlinear conditional-simulation gain*, which highlights the improvement that learning adds beyond exact Gaussian conditioning. On the near-Gaussian Allen–Cahn and Navier–Stokes benchmarks, this gain is small, as expected, and NCS matches the calibration of the best classical baseline to within a few percent. On the GARD-LENS precipitation ensemble, where the Gaussian assumptions break, NCS is the only method that is globally variance-calibrated at competitive CRPS and spectral error (Fortin-corrected $\sigma/s = 0.96$): across 132 noise-schedule settings of the diffusion baselines, every setting with $\sigma/s \geq 0.9$ has higher CRPS and more than twice the spectral error. The heavy precipitation tail, however, is under-reproduced by NCS trained with the energy score ($\kappa/\kappa_{\text{data}} = 0.17$); tuned consistency models reach it only while underdispersed ($\sigma/s \leq 0.57$). Our supply-versus-reach diagnostic attributes this deficit to the objective rather than the source. A scale-free flatness term closes 93% of an analogous deficit on a controlled system. On GARD-LENS, where the deficit depends on amplitude, an amplitude-conditional objective recovers most of the tail while keeping calibration and spectrum within their guards. With the same backbone and no more parameters than CorrDiff, NCS matches multi-step diffusion baselines such as EDM and CorrDiff in accuracy and proper score while using two network passes instead of 35 which reduces per-member wall-clock time by 17–36×.

1 INTRODUCTION

High-fidelity simulation is often too expensive at the resolution the science demands. Multi-fidelity learning runs a cheap coarse solver $u^{\text{LF}} = S^{\text{LF}}(\xi)$ and learns the correction $\delta = u^{\text{HF}} - u^{\text{LF}}$ that restores the fine structure (Kennedy & O'Hagan, 2000; Peherstorfer et al., 2018). The difficulty is that the correction is not one field. When the fine dynamics are driven by random forcing ω such as subgrid turbulence, stochastic nucleation, or convective variability, many fine fields are consistent with one coarse input. A refiner trained with squared error returns their mean, which by the law of total variance omits exactly $\mathbb{E}[\text{Cov}(\delta \mid u^{\text{LF}})]$, the within-condition spread, and therefore erases the extremes that matter for risk. A useful refiner must be a conditional *sampler*. Neural operators such as the FNO (Li et al., 2021) regress the correction and inherit this problem while diffusion samplers

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.

such as CorrDiff (Mardani et al., 2025) or GenCast (Price et al., 2025) recover the conditional law but pay tens of network passes per sample. Neither reports how much of its result a linear Gaussian baseline would have produced for free.

For Gaussian fields, conditional simulation was solved decades ago by *conditioning by kriging* (Matheron, 1963; Journel, 1974). The idea is to draw an unconditional realization u_{sim} with the right covariance by Rice's spectral synthesis (Rice, 1944) and then, correct it linearly toward the data,

$$u_{\mathrm{cond}} = u_{\mathrm{sim}} + \mathcal{K}\big(\mathrm{data} - \mathcal{O}(u_{\mathrm{sim}})\big). \tag{1}$$

$\mathcal{K}$ denotes the kriging operator, which is an *exact* sample from the conditional law in one pass. But turbulent residuals are intermittent, with conditional variance that grows with local amplitude (Kolmogorov, 1962), and precipitation is a point mass at zero with a heavy tail. No linear correction of a Gaussian draw honors either.

Neural Conditional Simulation. We extend the classical program by transforming the linear kriging operator into a learned nonlinear map which we call **Neural Conditional Simulation** (NCS). The unconditional simulator (Rice synthesis, composed with the Nataf transform (Nataf, 1962) for non-Gaussian marginals) and the nested multigrid hierarchy (Fedorenko, 1964) stay classical. At level ℓ of the cascade,

$$\hat{\delta}_\ell = \underbrace{\mathcal{K}_\ell\, c_\ell}_{\text{kriging mean}} + \underbrace{g_{\theta_\ell}\big(\varepsilon,\, c_\ell\big)}_{\text{learned innovation}}, \qquad \varepsilon \sim q_{\delta_\ell}, \tag{2}$$

where $\mathcal{K}_\ell$ is a closed-form Wiener filter estimated from training residuals, ε is the classical innovation draw, and g_{θ_ℓ} is a network initialized at the identity in ε. At initialization Eq. 2 is Matheron's sampler. We train NCS with the energy score (Gneiting & Raftery, 2007) which the untrained model already minimizes if the field is Gaussian (§4). This design makes the benefit of learning directly measurable. Since NCS starts from the exact Gaussian sampler, every table includes that sampler as a zero-learning baseline and we report each trained model relative to it. We call this difference the *nonlinear conditional-simulation gain* which tells the reader whether the network justifies its cost. Thanks to the modular design of NCS, we are able to identify the cause of departures from the Gaussian model, which we classify as supply-limited or reach-limited (§5.3). Supply-limited structure calls for a richer source or transport while a reach-limited structure calls for a different objective, which is how we close the tail deficits on SHE–PAM and GARD-LENS.

The zero-learning baseline and this taxonomy together organize our results. Where the Gaussian model nearly holds the gain is small, as it should be, and NCS matches both the calibrated classical floor and 35-step diffusion at 17–36$\times$ lower cost. On GARD-LENS precipitation the Gaussian assumptions fail, and NCS is the only method that is globally variance-calibrated at competitive CRPS and spectral error, even against diffusion baselines swept over 132 noise-schedule settings ($\sigma/s = 0.96$; Appendix F.6). Controlled experiments then show where learning is needed. *Supply-limited* departures, e.g. a nonlinear conditional mean, a condition-dependent spectrum, or nonstationarity, are recovered only by the learned transport which closes the full gap left by the stationary filter in a controlled nonstationary test. *Reach-limited* departures, such as heavy increment tails, are within the sampler's capacity but are not learned under the energy score and must be closed by changing the objective. A flatness term closes 93% of the tail deficit on the controlled SHE–PAM system, and an amplitude-conditional objective closes 78% on GARD-LENS precipitation (§5.4).

Contributions.

- **Neural conditional simulation.** We extend conditioning by kriging to non-Gaussian fields with an identity-initialized learned innovation (Eq. 2), which makes every learned gain measurable against the exact Gaussian sampler.
- **Diffusion-level quality in one pass per level.** NCS matches multi-step diffusion on stochastic PDEs and real precipitation at 17–36$\times$ lower cost and no more parameters than CorrDiff.
- **A diagnosis of where learning is needed.** We find that each component has its own job. The source shapes the spectrum, the transport captures supply-limited structure, and the objective governs the tail.

- **Cascade and evaluation.** The cascade doubles cross-level coherence ($0.46 \to 0.94$) and we evaluate all models with proper scores, calibration, spectra, and tail statistics against four families of baselines.

2 Background and related work

Conditioning by kriging. Kriging (Krige, 1951; Matheron, 1963) is the best affine predictor of a field from partial data; conditioning by kriging (Journel, 1974) turns it into a sampler by drawing an unconditional realization and kriging the mismatch between the data and what the realization implies for it (Eq. 1). In innovation form the sample is the kriging mean plus a draw from the Schur-complement covariance, exact for a jointly Gaussian pair (Prop. 2). The unconditional draw is Rice's spectral synthesis (Rice, 1944); the Nataf transform (Nataf, 1962) pushes it through an inverse marginal CDF. One classical fact is easy to get wrong in a learned pipeline: the source must be fitted to the *innovation*, not the raw residual, or it double-counts the variance the kriging mean explains (Lem. 2). Multigrid (Fedorenko, 1964; Bakhvalov, 1966) supplies the cascade; we evaluate with the proper scoring rules of BRIER (1950), Matheson & Winkler (1976), and Gneiting & Raftery (2007) and the spread/skill correction of Fortin et al. (2014).

Learned conditional simulation. Each ingredient of NCS exists in some line of work; we claim the combination and the measurement it permits (Appendix B gives the longer comparison). *Pathwise conditioning* of Gaussian processes (Wilson et al., 2020; 2021) is Matheron's rule at the level of random variables and the direct ancestor of our decomposition; it is linear and Gaussian, and NCS keeps its update and learns the remainder. NARGP (Perdikaris et al., 2017) is its multi-fidelity analogue. *Scoring-rule generators* train a one-pass conditional generator by minimizing the energy score rather than an adversarial loss (Gritsenko et al., 2020; Pacchiardi et al., 2024; Chen et al., 2024; Shen & Meinshausen, 2025), including for neural operators (PNO; Bülte et al., 2025). NCS belongs to this class; what differs is the base and the initialization, so that the zero-learning sampler is a row in every table, which none of these works reports. *Diffusion and GAN refiners* recover the conditional law at one resolution with many passes (Mardani et al., 2025; Price et al., 2025), as geostatistical simulators (Laloy et al., 2018; Mosser et al., 2020; Song et al., 2021), and for precipitation downscaling (Leinonen et al., 2021; Harris et al., 2022; Price & Rasp, 2022), which defines our GARD-LENS benchmark family; all replace the classical simulator rather than extend it. PDE-Refiner (Lippe et al., 2023) and DiffusionPDE (Huang et al., 2024) are the nearest iterative refiners. *One-step and multi-resolution generators*—consistency models (Song et al., 2023), MeanFlow (Geng et al., 2025), drifting (Deng et al., 2026; Esteban-Casadevall et al., 2026), Laplacian-pyramid and cascaded generators (Denton et al., 2015; Ho et al., 2022)—reach one step by distillation or compose resolutions as our cascade does; we use drifting only to initialize a sampler that is one-step because Eq. 1 is, and our cascade anchors each level to a classical simulator. The multi-fidelity surrogate lineage (Kennedy & O'Hagan, 2000; Lu et al., 2022; Howard et al., 2023; Li et al., 2020; Geneva & Zabaras, 2020) models the coarse-to-fine correction deterministically or with multi-step surrogates; none returns a calibrated conditional law in one pass per level.

3 Neural Conditional Simulation

Setup. Let $u^{\mathrm{LF}} = S^{\mathrm{LF}}(\xi)$ be the coarse solve and $u^{\mathrm{HF}} = S^{\mathrm{HF}}(\xi, \omega)$ the fine solve. Because u^{LF} does not encode the forcing ω, many fine fields are consistent with the same coarse field, so u^{HF} given u^{LF} is random. We work on nested grids $G_0 \subset \cdots \subset G_L$ with bilinear prolongation $\mathcal{I}$. At level ℓ the condition is $c_\ell = \mathcal{I}_{\ell\to\ell+1}(\hat{u}_\ell)$ and the target is the residual $\delta_\ell = u_{\ell+1} - c_\ell$ with law $p_{\mathrm{data}}(\delta_\ell \mid c_\ell)$. Each training data sample contains K fine realizations per condition. Figure 1 summarizes the construction.

The classical sampler The zero-learning sampler has two parts, both estimated once from training residuals and frozen. The *kriging mean* assumes second-order stationarity within a level, so the best affine predictor of δ_ℓ from c_ℓ is a per-wavenumber Wiener filter, $K(k) = S_{\delta c}(k) S_{cc}(k)^{-1}$, applied as a diagonal multiply in the Fourier domain. We use the Tikhonov form $\widehat{K}_\lambda(k) = \widehat{S}_{\delta c}(k)\big(\widehat{S}_{cc}(k) + \lambda I\big)^{-1}$ at $\lambda_{\mathrm{rel}} = 10^{-6}$ due to the finite data regime. The *innovation source* is Rice synthesis (Rice,

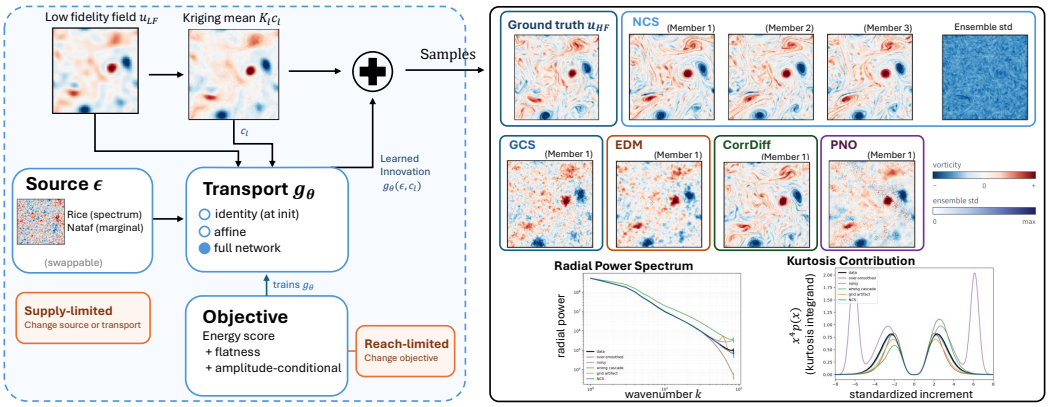


Figure 1: Neural Conditional Simulation. Left: One cascade level - closed-form kriging mean $\mathcal{K}_l c_l$ plus learned transport g_θ of a source draw ϵ. Source, transport and objective are swappable. Right: NCS samples vs. truth and baselines. Bottom: schematic spectra and PDFs of typical failure modes.

1944),

$$\varepsilon_\ell(k) = \sqrt{\widehat{S}_{r,\lambda}(k)}\, \xi_\ell(k), \qquad \xi_\ell(k) \sim \mathcal{CN}(0,1), \tag{3}$$

with Hermitian symmetry enforced and the DC mode zeroed. The spectrum $\widehat{S}_{r,\lambda}$ is that of the innovation $r = \delta - \widehat{K}_\lambda c$, not the raw residual: a source fitted to the residual double-counts the variance the kriging mean explains and is over-dispersed by exactly $\mathrm{Cov}(\widehat{K}_\lambda c)$ (Lem. 2). $\widehat{S}_{r,\lambda}$ is the exact residual spectrum of the regularized filter (Eq. 22; Appendix E.3). For zero-inflated or heavy-tailed marginals we compose Eq. 3 with the Nataf transform (Nataf, 1962), $F_P^{-1} \circ \Phi$, which sets the marginal independently of the spatial dependence. Kriging mean plus a draw from either source is the classical conditional simulator, reported in every table as GCS (Gaussian source) or TGCS (transformed source).

The learned innovation NCS keeps the kriging mean and replaces the innovation draw by a learned map of it, Eq. 2. The network sees the classical draw and the condition, so it can reshape the innovation's mean, scale, spectrum, and marginal as functions of c_ℓ. It is initialized such that $g_\theta(\varepsilon, c) = \varepsilon$, which makes the untrained NCS identical to GCS. Details on the networks are available in (Appendix G).

Training The cascade is trained end to end against the energy score (Gneiting & Raftery, 2007),

$$\mathrm{ES}\big(P_\theta(\cdot \mid c), y\big) = \mathbb{E}\|X - y\| - \tfrac{1}{2}\mathbb{E}\|X - X'\|, \qquad X, X' \overset{\text{iid}}{\sim} P_\theta(\cdot \mid c), \tag{4}$$

estimated with M cascade rollouts per condition, the unbiased leave-one-out repulsion term, and the data term averaged over the K members (Eq. 28). The energy score is strictly proper and weakly sensitive to tails; for zero-inflated fields we also train a variant with an added threshold-weighted score (Allen et al., 2023), reported as NCS+tw.

From the identity initialization, energy-score training is well posed. We initialize g_θ by *drifting* (Deng et al., 2026), which in our ablations mainly improves spatial structure (§5.2). For each condition, with generated members $\{x_m\}_{m=1}^M$, data members $\{y_k\}_{k=1}^K$, and a radial kernel w_h,

$$V(x_m \mid c) = \frac{\sum_k w_h(x_m, y_k)\,(y_k - x_m)}{\sum_k w_h(x_m, y_k)} - \frac{\sum_{m' \neq m} w_h(x_m, x_{m'})\,(x_{m'} - x_m)}{\sum_{m' \neq m} w_h(x_m, x_{m'})} \tag{5}$$

is an attraction toward the data ensemble minus a leave-one-out repulsion within the generated one. V is the negative gradient of a kernelized energy distance between the two ensembles (Esteban-Casadevall et al., 2026), so regressing g_θ onto V matches distributions, not means. We pretrain for 100 epochs with $M = 16$, then fine-tune the cascade for 50 epochs under Eq. 4 with $M = 8$. Every NCS result uses this recipe. The identity-initialized, energy-score-only arm is reported as the ablation NCS (no pretrain), with a matched-budget comparison in Appendix F.4.

4 WHAT THE THEORY GUARANTEES

The five results below, each proved in Appendix C, establish what NCS guarantees at initialization and in population, and where these guarantees end. Each builds on classical results, and together they show which properties of Matheron's sampler survive when the innovation is learned. Let $m_{\mathrm{lin}}(c)$ be the affine (kriging) predictor and $\Sigma_r = \Sigma_{\delta\delta} - \Sigma_{\delta c}\Sigma_{cc}^{-1}\Sigma_{c\delta}$ the innovation covariance.

1. **Point prediction under-disperses** (Cor. 1). A refiner trained with squared error returns $\mathbb{E}[\delta \mid c]$ and omits exactly $\mathbb{E}[\mathrm{Cov}(\delta \mid c)]$, so a refiner must sample.
2. **The source is the innovation** (Lem. 1, 2). $r = \delta - m_{\mathrm{lin}}(c)$ is uncorrelated with c and has covariance Σ_r. A source fitted to $\Sigma_{\delta\delta}$ instead over-disperses by exactly $\mathrm{Cov}(m_{\mathrm{lin}}(c))$ (measured 1.98 ± 0.02 vs. 2.00, Appendix E.3).
3. **Exact for Gaussian fields** (Prop. 2). If (δ, c) are jointly Gaussian, $m_{\mathrm{lin}}(c) + r_0$ with $r_0 \sim \mathcal{N}(0, \Sigma_r)$ independent of c is an exact draw from $\mathcal{L}(\delta \mid c)$: the untrained NCS *is* Matheron's sampler.
4. **No loss of generality** (Lem. 3). Any conditional law can be written as $T(r_0, c)$ for a measurable T, so transporting the Gaussian innovation can represent any target. Remaining deficits stem from finite capacity or the objective.
5. **Training preserves a correct answer** (Prop. 3, Cor. 3). A strictly proper score is minimized in population only by the true law, so for Gaussian fields NCS at initialization is already optimal.

We explicitly test two limits: (i) per-level exactness does not imply exactness of the composed cascade without a Markov factorization in resolution (Eq. 29, §5.5), and (ii) strict propriety says nothing about a finite-sample score's sensitivity to tails (§5.4).

5 EXPERIMENTS

5.1 SETUP AND REPORTING POLICY

Benchmarks. We perform the full benchmark ladder on three systems: stochastic Allen–Cahn (near-Gaussian, $16 \to 64$), turbulent Navier–Stokes (intermittent, $64 \to 256$, large-scale vorticity shared across members by nudging), and the GARD-LENS precipitation ensemble (real, zero-inflated, heavy-tailed, $32 \to 128$ over CONUS at 12 km). Each condition has $K = 8$ fine members and splits are by condition. Additionally, five single-axis synthetic controls and two Gaussian controls with closed-form conditional laws serve the diagnostic experiments. We use two Navier–Stokes datasets. The large dataset described above is used for the main results, and a smaller one ($32 \to 128$) for the factorial, cascade, warm-start, and sampling-budget studies. Navier–Stokes scores from these studies are therefore not directly comparable to Table 1. We refer the reader to Appendix D.1 for details.

Baseline ladder. The baselines comprise three categories: First, we use the following point predictors: (1) bilinear prolongation, (2) the kriging mean (LIN), and (3) a squared-error network (NN-L_2). Second, we present classical gaussian conditional simulators: (4) GCS which is the Gaussian innovation sampler of §3 and *classical floor*, (5) HET-GCS using a per-condition innovation scale, (6) TGCS adding the Nataf marginal and (7) MARG which implements the marginal-residual source of Lemma 2. Additionally, the third category deploys learned samplers: (8) PNO (Bülte et al., 2025), (9) a one-pass operator trained with the same energy score, (10) consistency distillation, (11) PDE-Refiner, (12) EDM and (13) CorrDiff as 35-step conditional diffusion. NCS is the drift-initialized model of §3, while NCS (no pretrain) is used the identity-initialized ablation and NCS+tw adds the threshold-weighted score. All learned arms share the same backbone class, conditioning channels, and split (Appendix D.2).

Metrics and policy. We report RMSE of the ensemble mean, energy score (ES), pixelwise CRPS, Fortin-corrected spread/skill σ/s (Fortin et al., 2014) (a score of 1.0 indicates perfect calibration), radial log-power-spectrum error (logPSD), the standardized-increment kurtosis ratio $\kappa/\kappa_{\mathrm{data}}$ (1.0 means matching tail structure), and network evaluations per member (NFE), all at $M = 32$ members

(Appendix D.3). Our evaluation is based on three rules: (i) GCS is always scored on the same conditions and metrics as every learned arm to obtain a paired difference from the Gaussian floor, (ii) we report the condition-bootstrap standard deviation ($\pm$ one s.d.) or the 95% condition-bootstrap percentile intervals, and (iii) Appendix H presents the criteria for each claim in our paper.

Supply versus reach. We split deviations from the Gaussian model in two categories. We deem a system *supply-limited* if no classical innovation model can supply the structure, potentially because it requires a nonlinear conditional mean, a condition-dependent spectrum, or a nonstationarity. We call a system *reach-limited* when the sampler could represent the structure but the objective does not reach it which could be caused by heavy increment tails. The distinction decides where a fix must act and is further explained in §5.3 and §5.4.

5.2 Results

Table 1 reports the ladder on the three principal systems (GARD-LENS members in Fig. 8, Appendix F).

Allen–Cahn: Allen–Cahn is a nearly Gaussian system. Unsurprisingly, the classical samplers show a good calibration (GCS $\sigma/s = 0.85$, TGCS 0.95) and their energy score, 32.8–33.0, is among the best values any arm reaches. NCS ties it at 32.8 with $\sigma/s = 0.91$, as do the 35-step diffusion samplers. The gain over the floor is small, as the Gaussian anchor predicts, and shows only in the spectrum. NCS's logPSD (0.13) is lower than that of EDM and CorrDiff (0.78–0.79) whose sources do not carry the residual spectrum.

Navier–Stokes: The classical samplers are again calibrated in spread ($\sigma/s = 1.00$) and match bilinear prolongation in terms of RMSE. Every learned sampler shows a correct center. NCS reaches RMSE 0.85, ES 312.6, and $\sigma/s = 0.97$, matching CorrDiff within noise. Plain NCS has a weaker spectrum than EDM (logPSD 0.36 vs 0.05), which an additional variogram term reduces to 0.24 (§5.4).

GARD-LENS: No classical sampler is useful on precipitation: GCS, HET-GCS, and MARG are calibrated in spread ($\sigma/s = 0.92$–0.93) but far off on CRPS (2.46–2.50), and TGCS reaches CRPS 1.90 at $\sigma/s = 0.83$, because a stationary Gaussian innovation cannot place wet cells. NCS has the best energy score of all sixteen arms (473.7; PNO 476.2) and is the only method that is calibrated ($\sigma/s = 0.96$) at competitive CRPS (1.59) and spectral error (0.15). At their default schedules EDM and CorrDiff reach lower CRPS (1.54–1.55) but are strongly under-dispersed (0.43, 0.51), and tuning their sampling-time noise schedule does not change the picture (Table 19): across 132 settings, every setting with $\sigma/s \geq 0.9$ has higher CRPS (≥ 1.64) and more than twice the spectral error of NCS (≥ 0.37), and the settings that beat NCS on CRPS remain under-dispersed ($\sigma/s \leq 0.87$). PNO is calibrated too (1.02) but at worse CRPS (1.61) and spectrum (0.55). The tail is where energy-score NCS falls short: its $\kappa/\kappa_{\mathrm{data}} = 0.17$ is below Consistency's 0.21 at default, and multi-step consistency sampling reaches the data tail (up to 1.10) only while under-dispersed ($\sigma/s \leq 0.57$) and at more than twice the spectral error of NCS. In §5.4 we address this deficit with an amplitude-conditional term that recovers the tail at unchanged accuracy and spread, and with improved spectrum (Table 15).

Pretraining improves spatial structure: In a clean ablation that changes only the initialization, with the same source and training budget (Table 17), drift pretraining lowers the variogram score from 0.168 to 0.048 on Allen–Cahn and from 12.1 to 5.7 on GARD-LENS, improves logPSD on both, and lowers GARD-LENS CRPS from 1.92 to 1.66. Energy score and spread are improved modestly and all metrics are kept within their guards.

5.3 Supply-limited Systems

In this section, we run a series of controlled experiments to identify which property of the field produces the gain over the floor and how NCS addresses them. First, starting from a Gaussian control with known conditional law, we break (with ranging severities) one Gaussian assumption at a time,

Table 1: Held-out results (mean ± s.d. over three seeds or a condition bootstrap). For σ/s (Fortin-corrected) and $\kappa/\kappa_{\mathrm{data}}$, 1 is ideal; "–" means no spread. Bold is best per benchmark. Full results in Table 10.

Method	RMSE	ES	CRPS	σ/s	logPSD	$\kappa/\kappa_{\mathrm{data}}$	NFE
Allen–Cahn							
Bilinear	0.506	52.6	0.576	0.00	0.193	–	0
NN-L_2	0.326 ± 0.004	47.4 ± 0.1	0.625 ± 0.001	0.00	0.312 ± 0.034	–	1
GCS (floor)	0.305 ± 0.001	33.0 ± 0.1	0.425 ± 0.001	0.85 ± 0.01	0.026 ± 0.001	0.70 ± 0.00	0
TGCS	0.310 ± 0.001	32.8 ± 0.0	0.418 ± 0.001	0.95 ± 0.01	0.319 ± 0.001	0.97 ± 0.00	0
PNO	0.400 ± 0.000	37.9 ± 0.0	0.491 ± 0.004	0.50 ± 0.00	1.076 ± 0.090	0.98 ± 0.10	1
EDM	0.316 ± 0.022	33.1 ± 0.6	0.388 ± 0.008	0.87 ± 0.05	0.775 ± 0.024	5.78 ± 0.11	35
CorrDiff	0.306 ± 0.008	32.8 ± 0.1	0.384 ± 0.003	0.90 ± 0.02	0.793 ± 0.012	5.35 ± 0.06	35
NCS	0.304 ± 0.011	32.8 ± 0.2	0.409 ± 0.002	0.91 ± 0.00	0.134 ± 0.001	0.87 ± 0.00	2
Navier–Stokes							
Bilinear	0.862 ± 0.006	452.0 ± 1.0	1.395 ± 0.003	0.00	3.258 ± 0.015	–	0
NN-L_2	0.929 ± 0.004	458.2 ± 0.8	1.417 ± 0.002	0.00	1.226 ± 0.006	–	1
GCS (floor)	0.882 ± 0.005	314.8 ± 0.7	0.976 ± 0.002	0.998 ± 0.002	0.033 ± 0.001	0.75 ± 0.00	0
TGCS	0.881 ± 0.005	314.8 ± 0.7	0.976 ± 0.002	0.997 ± 0.002	0.037 ± 0.002	0.76 ± 0.00	0
PNO	1.113 ± 0.007	341.7 ± 0.9	1.203 ± 0.033	0.83 ± 0.00	0.326 ± 0.002	2.39 ± 0.90	1
EDM	0.859 ± 0.005	312.8 ± 0.7	0.969 ± 0.002	1.00 ± 0.00	0.045 ± 0.001	1.01 ± 0.00	35
CorrDiff	0.836 ± 0.005	310.9 ± 0.6	0.963 ± 0.002	1.00 ± 0.00	0.041 ± 0.001	1.01 ± 0.00	35
NCS	0.852 ± 0.005	312.6 ± 0.6	0.969 ± 0.002	0.97 ± 0.00	0.364 ± 0.012	0.90 ± 0.01	2
GARD-LENS							
Bilinear	1.213	662.5	2.429	0.00	1.529	–	0
NN-L_2	1.248 ± 0.056	645.9 ± 0.2	2.316 ± 0.004	0.00	1.206 ± 0.004	–	1
GCS (floor)	1.285 ± 0.010	505.9 ± 0.4	2.499 ± 0.004	0.92 ± 0.00	0.018 ± 0.000	0.01 ± 0.00	0
TGCS	1.474 ± 0.004	507.3 ± 0.2	1.895 ± 0.002	0.83 ± 0.00	0.230 ± 0.001	0.13 ± 0.00	0
PNO	1.290 ± 0.030	476.2 ± 0.6	1.612 ± 0.007	1.02 ± 0.03	0.550 ± 0.080	0.09 ± 0.00	1
EDM	1.445 ± 0.165	513.0 ± 5.9	1.535 ± 0.009	0.43 ± 0.04	0.329 ± 0.071	0.16 ± 0.03	35
CorrDiff	1.150 ± 0.117	508.0 ± 3.1	1.550 ± 0.017	0.51 ± 0.03	0.122 ± 0.040	0.12 ± 0.01	35
NCS	1.183 ± 0.035	473.7 ± 0.1	1.588 ± 0.010	0.96 ± 0.01	0.148 ± 0.013	0.17 ± 0.03	1

i.e. nonlinear conditional mean, condition-dependent innovation scale, skewed innovations, heavy-tailed innovations, a point mass at zero, condition-dependent spatial correlation and a cross-scale dependence between innovations, and score GCS, HET-GCS, TGCS, and NCS on the metric it targets (Fig. 2, Table 11). We observe that the classical specialists recover what they were built for: HET-GCS recovers condition-dependent innovation *scale* at the strongest severity ($\kappa/\kappa_{\mathrm{data}}$ 0.88, GCS 0.41) and TGCS is able to recover a *skewed* marginal (0.98, GCS 0.68). NCS is competitive at moderate severity and overshoots the innovation scale departure at the strongest severity (1.88), where a scalar per-condition variance is the tighter tool. We find that a *nonlinear conditional mean* leaves every classical arm at CRPS 1.13–1.16 while *only* NCS stays at 0.60 since no classical innovation model touches the mean. Moreover, since a *condition-dependent spectrum* is invisible to pooled metrics, we extend our evaluation to a condition-resolved tracking ratio (Appendix F) and find that classical arms track at ≤ 0.01 while NCS is much better at 0.78, 0.57, 0.24 across severities. On the *heavy tail* and *point mass* no arm recovers fully and NCS advances furthest (0.73 vs TGCS 0.50). *Cross-scale* dependence for single and multi-level arms is tested thoroughly in §5.5.

Nonstationarity against a known ceiling: The kriging mean assumes stationarity within each level. To quantify the effects of nonstationarity, we construct a Gaussian control whose conditional mean is a position-dependent mixture of two frequency bands (Eq. 38) which gives us the exact conditional mean in closed form. As a result, we can measure the gap between the stationary filter and the exact mean: it accounts for 78% of the local-variance error and we take this gap as the recoverable ceiling. Relative to this ceiling, a locally stationary filter recovers $0.54\times$, a per-pixel rescaling $0.51\times$, and a season-stratified filter $-0.15\times$, i.e. it widens the gap. The learned NCS correction recovers $1.00\times$, closing the gap entirely (Appendix E). On GARD-LENS there is no closed-form ceiling available and our experiments show that the learned arm improves on the global filter by 61–68% which we state as unnormalized improvement.

Three separable levers: Since NCS combines a spectrally fitted source, a non-Gaussian marginal, and a learned transport, its gains could come from any one of these components or from their interaction. To isolate each component's contribution, we run a factorial experiment on Allen–Cahn and Navier–Stokes, pairing each of four sources with each of three transports (identity, learned affine, and full network) (Table 12). The results show that each component governs a different metric. Fit-

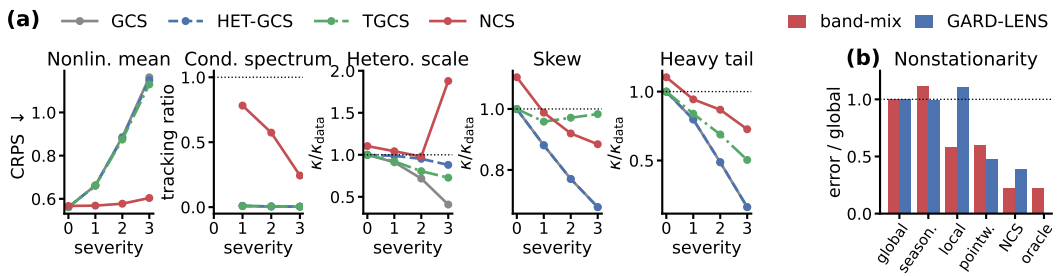


Figure 2: Where learning is needed. (a) One departure per panel at four severities, scored on its target metric (dotted: data). (b) Local-variance error relative to the global stationary filter. Remaining departures in Table 11.

Table 2: Objective ablation on Navier–Stokes: one generator and source, five objectives. Point estimates with 95% condition-bootstrap intervals. Tail columns are gaps to the data increment law (0 is the target). Full table in Appendix F.

Objective	ES	σ/s	logPSD	κ_{h2} gap	P99 gap
L_2	367.0 [364.0, 369.6]	0.375 [0.372, 0.378]	0.706	-0.924	-1.713
Energy score (ES)	312.6 [311.3, 313.8]	0.973 [0.969, 0.977]	0.364	-0.386	-0.900
ES + variogram	312.7 [311.4, 313.9]	0.964 [0.960, 0.968]	0.236	-0.953	-0.740
ES + threshold-weighted	312.8 [311.5, 314.1]	0.973 [0.969, 0.977]	**0.208**	$\mathbf{+0.027}$	$\mathbf{+0.143}$
Drift pretraining only	312.1 [310.8, 313.4]	0.951 [0.947, 0.954]	0.493	-0.402	-0.917

ting the source to the residual spectrum reduces logPSD from 1.3–1.7 to 0.1–0.5, whereas replacing the affine transport with the full network changes it by less than 0.2 indicating that the *source spectrum* affects logPSD. Furthermore, holding the source fixed, the full network lowers ES relative to the affine transport by 12% on Allen–Cahn (37.5 → 33.0) and 11% on Navier–Stokes (573 → 513) without affecting logPSD or the tail which shows that the *transport* controls the proper score. On Allen–Cahn, only the Nataf source reaches the tail (1.1, versus 0.7–0.8 for the other sources) regardless of spectrum or transport, so the *source marginal* controls the tail. On Navier–Stokes, by contrast, no combination exceeds 0.65. Since no source or transport reaches the tail on this turbulent system, we attribute the shortfall to the training objective rather than to the model. We investigate this further for GARD-LENS in §5.4.

5.4 Reach-limited Systems

To test whether the remaining gap lies in the objective, we hold the Navier–Stokes architecture and source fixed and vary the training objective (Table 2). All probabilistic objectives reach the same accuracy (ES 312–313), whereas squared error collapses the ensemble to a spread–skill ratio of $\sigma/s = 0.375$. Trained with the energy score alone, NCS calibrates spread but leaves a spectrum error of 0.36 and an increment-kurtosis gap of -0.39. Adding a variogram score (Appendix F) reduces logPSD by 35% without any loss in accuracy but widens the kurtosis gap to -0.95. Adding a threshold-weighted score is the only change that closes the tail (κ_{h2} gap $+0.03$, P99 gap $+0.14$), and it also lowers logPSD to 0.21, though this may partly reflect its smaller training batch (Appendix F). The energy score alone thus captures neither the spectrum nor the tail, and only a tail-sensitive term reaches the tail. Drift pretraining without the energy-score stage matches this accuracy but falls short on both spread and spectrum.

Restoring a tail the energy score cannot reach: Following our definition of a reach-limited system, the sampler can represent the increment tail but is not learned under the energy score. Since changing the source or transport leaves the tail unchanged (§5.3), the fix must act on the objective. We therefore add a scale-free term that matches the increment flatness $F(h) = S_4(h)/S_2(h)^2$. This term targets the fourth-order structure directly and since it is a ratio, it leaves the calibrated

second moment untouched. On our controlled reach-limited system SHE–PAM, the data shows an increment kurtosis of 38.3 and the pure energy-score reaches only 3.99. Introducing the flatness term raises the kurtosis to 35.9, closing 93% of the deficit, while nRMSE, σ/s, and logPSD remain within their tolerances (Table 14, Appendix F).

The precipitation tail: On GARD-LENS precipitation, the scale-free flatness term recovers only about half of the kurtosis deficit at the cost of degrading the spectrum and spread, while a threshold-weighted score leaves the tail unchanged. The GARD-LENS deficit is not a uniform kurtosis gap but depends on amplitude. When we bin wet cells by local intensity, the ensemble is over-dispersed by $3.8\times$ at low amplitudes and under-dispersed to $0.87\times$ in the heavy tail and the within-bin variance ratio crosses one as amplitude increases. By construction, the flatness term cannot correct a spread ratio that changes with amplitude and forcing the pooled statistic toward the data corrupts the spectrum instead. Matching instead how the increment variance scales with amplitude, the K62 relation (Kolmogorov, 1962), closes 78% of the deficit, raising the lag-1 kurtosis ratio from 0.14 to 0.81 while halving the spectral error (Table 15): increments whose variance varies with amplitude pool into a heavy-tailed mixture. Only the objective changed, so the tail was within the sampler's reach. Merely reweighting the energy score by amplitude is not enough and worsens every metric.

5.5 Cascade and sampling cost

The cascade buys cross-scale consistency, not aggregate score: Exactness at each level does not imply exactness of the composed cascade (§4), so we compare the hierarchy with a capacity-matched single level (Table 18, Appendix F.5). On Navier–Stokes the aggregate scores barely move (ES 509.6 vs 511.6), but the finest-level logPSD halves ($1.46 \rightarrow 0.70$), the cross-level coherence rises from 0.46 to 0.94 (data 0.91), and the restriction inconsistency falls from 0.43 to 0.10. The end-to-end stage matters: without it, logPSD on Navier–Stokes rises from 0.70 to 0.87 and nRMSE on Allen–Cahn from 0.61 to 0.79. Correlated latents change nothing beyond seed variance.

Sampling budget and cost. Re-sampling the same checkpoints across their step budgets (Fig. 9, Appendix F.6) shows that few-step diffusion is unstable and that both multi-step methods need 8–11 evaluations to reach their best scores. NCS, at two evaluations, lies on or below that frontier on every metric shown. On one GH200 (Appendix F.7) this is 0.74–0.95 ms per member against 15.0–26.4 ms for EDM with the same backbone, a 17–36× speedup. CorrDiff is as slow as EDM. One-pass operators and the distilled consistency model are faster, but none is calibrated on all three benchmarks.

6 Limitations and conclusion

Limitations: The kriging mean assumes stationarity within each level. The learned correction recovers the resulting cost in our experiments (§5.3), but we have no guarantee that it always does. On GARD-LENS, the amplitude-conditional objective closes most of the tail deficit, but the kurtosis ratio stays below one; baselines that reach the tail do so only while under-dispersed. NCS is one to two orders of magnitude slower per member than one-pass operators such as PNO. Finally, the Gaussian anchor gives guarantees only near Gaussianity, i.e. far from it, the evidence is empirical.

Conclusion. Matheron's construction gives an exact one-pass sampler for Gaussian fields. NCS extends it to non-Gaussian multi-fidelity refinement by learning only the innovation, starting from the classical sampler. This makes the gain from learning directly measurable: it is small where the Gaussian model nearly holds and decisive where it does not. NCS delivers it at one pass per level while matching 35-step diffusion. Our taxonomy shows where to intervene: the transport for structure the Gaussian model cannot supply, the source for the spectrum, and the objective for the tail. Following it, an amplitude-conditional objective recovers most of the precipitation tail.

AI use statement

We used generative AI (LLM-based) tools to help draft and edit prose in this manuscript, i.e. tightening sentences, restructuring paragraphs, and standardizing notation and terminology across sections.

We did not use generative AI tools to design experiments, generate or fabricate results, run analyses, or produce the numerical values reported in any table or figure. Every reported number was computed by the authors' own experimental pipeline and is traceable to a committed run. All AI-assisted text was reviewed and edited by the authors.

Ethics statement

This work raises no human-subjects, privacy, or dual-use concerns that we are aware of. All experiments use synthetic stochastic PDE simulations that we generate ourselves, plus the publicly available GARD-LENS precipitation ensemble. No personally identifiable or sensitive data is involved. We are not aware of a direct path from calibrated multi-fidelity refinement of PDE solutions to harmful applications, though as with any generative modeling technique, downstream use in decision-critical settings (e.g., flood or storm risk assessment) should not treat model output as a substitute for domain-expert validation.

Reproducibility statement

We aim for every claim in this paper to be traceable to a specific, reported measurement rather than to prose alone. Section H states the evidentiary criterion required before each class of claim is made, and Appendix H.1 maps each theoretical step to the experiment that would falsify or weaken it (Table 21). Full theory statements, proofs, and a discussion of their scope are given in Appendix C; the assumptions under which each proposition holds (joint Gaussianity only where stated) are given alongside each result rather than left implicit. Benchmark definitions, data splits, and the GARD-LENS preprocessing pipeline are described in Appendix D.1. The full baseline ladder (affine, nonlinear, classical Gaussian/heteroscedastic/ transformed-Gaussian conditional simulation, generative, and NCS variants) is specified in Appendix D.2 and every metric used (proper scores, Fortin-corrected spread/skill, spectral, variogram, and tail diagnostics) is defined precisely in Appendix D.3. All main and appendix tables report multi-seed uncertainty. We intend to release the code upon acceptance.

References

Sam Allen, David Ginsbourger, and Johanna Ziegel. Evaluating forecasts for high-impact events using transformed kernel scores. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):906–940, 2023. doi: 10.1137/22M1532184.

N.S. Bakhvalov. On the convergence of a relaxation method with natural constraints on the elliptic operator. *USSR Computational Mathematics and Mathematical Physics*, 6(5):101–135, 1966. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(66)90118-2.

GLENN W. BRIER. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Christopher Bülte, Philipp Scholl, and Gitta Kutyniok. Probabilistic neural operators for functional uncertainty quantification. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Jieyu Chen, Tim Janke, Florian Steinke, and Sebastian Lerch. Generative machine learning methods for multivariate ensemble postprocessing. *The Annals of Applied Statistics*, 18(1):159–183, 2024. doi: 10.1214/23-AOAS1784.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting, 2026.

Emily L. Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1486–1494, 2015.

Maria Esteban-Casadevall, Jorge Carrasco-Pollo, Max Welling, Jan-Willem van de Meent, Erik J. Bekkers, and Floor Eijkelboom. Kernel-gradient drifting models, 2026.

R.P. Fedorenko. The speed of convergence of one iterative process. *USSR Computational Mathematics and Mathematical Physics*, 4(3):227–235, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90253-8.

V. Fortin, M. Abaza, F. Anctil, and R. Turcotte. Why should ensemble spread match the rmse of the ensemble mean? *Journal of Hydrometeorology*, 15(4):1708 – 1713, 2014. doi: 10.1175/JHM-D-14-0008.1.

Nicholas Geneva and Nicholas Zabaras. Multi-fidelity generative deep learning turbulent flows. *Foundations of Data Science*, 2(4):391428, 2020. ISSN 2639-8001. doi: 10.3934/fods.2020019.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Alexey A. Gritsenko, Tim Salimans, Rianne van den Berg, Jasper Snoek, and Nal Kalchbrenner. A spectral energy distance for parallel speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13062–13072, 2020.

Lucy Harris, Andrew T. T. McRae, Matthew Chantry, Peter D. Dueben, and Tim N. Palmer. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003120, 2022. doi: 10.1029/2022MS003120.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Amanda A. Howard, Mauro Perego, George Em Karniadakis, and Panos Stinis. Multifidelity deep operator networks for data-driven and physics-informed problems. *Journal of Computational Physics*, 493:112462, 2023. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2023.112462.

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. DiffusionPDE: Generative PDE-solving under partial observation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

A. G. Journel. Geostatistics for conditional simulation of ore bodies. *Economic Geology*, 69(5):673–687, 1974. doi: 10.2113/gsecongeo.69.5.673.

Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer Nature Switzerland AG, Cham, 3rd edition, 2021. ISBN 978-3-030-61870-4. doi: 10.1007/978-3-030-61871-1.

M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. ISSN 00063444, 14643510.

A. N. Kolmogorov. A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number. *Journal of Fluid Mechanics*, 13:82–85, January 1962. doi: 10.1017/S0022112062000518.

D.G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951. doi: 10.10520/AJA0038223X_4792.

Eric Laloy, Romain Hérault, Diederik Jacques, and Niklas Linde. Training-image based geostatistical inversion using a spatial generative adversarial neural network. *Water Resources Research*, 54(1):381–406, 2018. doi: 10.1002/2017WR022148.

Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, 2021. doi: 10.1109/TGRS.2020.3032790.

Shibo Li, Wei Xing, Mike Kirby, and Shandian Zhe. Multi-fidelity bayesian optimization via deep neural networks. 2020.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E Turner, and Johannes Brandstetter. PDE-refiner: Achieving accurate long rollouts with neural PDE solvers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lu Lu, Raphaël Pestourie, Steven G. Johnson, and Giuseppe Romano. Multifidelity deep neural operators for efficient learning of partial differential equations with application to fast inverse design of nanoscale heat transport. *Phys. Rev. Res.*, 4:023210, Jun 2022. doi: 10.1103/PhysRevResearch.4.023210.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, Feb 2025. ISSN 2662-4435. doi: 10.1038/s43247-025-02042-5.

Georges Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976. ISSN 00251909, 15265501.

Lukas Mosser, Olivier Dubrule, and Martin J. Blunt. Stochastic seismic waveform inversion using generative adversarial networks as a geological prior. *Mathematical Geosciences*, 52(1):53–79, 2020. doi: 10.1007/s11004-019-09832-6.

André Nataf. Étude graphique de détermination de distributions de probabilités planes dont les marges sont données. *Annales de l'ISUP*, XI(3):[257]–260, 1962.

Lorenzo Pacchiardi, Rilwan A. Adewoyin, Peter Dueben, and Ritabrata Dutta. Probabilistic forecasting with generative networks via scoring rule minimization. *Journal of Machine Learning Research*, 25(45):1–64, 2024.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018. doi: 10.1137/16M1082469.

Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D. Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017. doi: 10.1098/rspa.2016.0751.

Pierre Pinson and Julija Tastu. Discrimination ability of the energy score. Technical Report 15, 2013.

Ilan Price and Stephan Rasp. Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10555–10571. PMLR, 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, Jan 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08252-9.

S. O. Rice. Mathematical analysis of random noise. *The Bell System Technical Journal*, 23(3): 282–332, 1944. doi: 10.1002/j.1538-7305.1944.tb00874.x.

Xinwei Shen and Nicolai Meinshausen. Engression: Extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):653–677, 2025. doi: 10.1093/jrsssb/qkae108.

Suihong Song, Tapan Mukerji, and Jiagen Hou. GANSim: Conditional facies simulation using an improved progressive growing of generative adversarial networks (GANs). *Mathematical Geosciences*, 53(7):1413–1444, 2021. doi: 10.1007/s11004-021-09934-0.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023.

Gábor J. Székely and Maria L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2003.12.002.

James T. Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. Efficiently sampling functions from Gaussian process posteriors. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10292–10302. PMLR, 2020.

James T. Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. Pathwise conditioning of Gaussian processes. *Journal of Machine Learning Research*, 22(105):1–47, 2021.

A Appendix

The appendix contains the extended comparison with prior work (Appendix B), full theory statements and proofs (Appendix C), benchmark, baseline, and metric definitions (Appendix D), the diagnostic experiments on the classical machinery that the main text summarizes in one paragraph each (Appendix E), full result tables and the remaining studies (Appendix F), architecture and training details (Appendix G), and the claim criteria and theory-to-experiment map (Appendix H).

B Relation to prior work

Section 2 states where each ingredient of NCS comes from. This appendix records, for each line of work, what NCS shares with it and where it departs, so that the claim of novelty can be checked against the specific combination we make rather than against any single component. We do not claim to be the first to condition a Gaussian draw by kriging, the first to train a one-pass generator with a proper scoring rule, the first to apply a learned sampler to geostatistical or precipitation fields, or the first to compose generators across resolutions. The claim is the combination in Eq. 2, its identity initialization at the exact classical sampler, and the evaluation this permits.

Pathwise conditioning and Matheron's rule. Wilson et al. (2020; 2021) restate Gaussian-process conditioning at the level of random variables: a posterior sample is a prior sample plus a kriging update on the residual between the data and the prior sample's own observations. This is Matheron's substitution rule (Matheron, 1963; Journel, 1974), Eq. 1, and it is the same decomposition NCS uses at each level. The differences are what is held fixed and what is learned. Pathwise conditioning is linear and exact for Gaussian processes, and its contribution is computational: the update scales with the number of observations rather than the number of test locations. NCS keeps the linear update as an explicit, closed-form spectral Wiener filter (Cor. 2) and adds a nonlinear map on the innovation source that is zero at initialization. Where pathwise conditioning approximates the prior draw (random Fourier features) to make Gaussian sampling cheap, NCS draws the prior exactly by Rice synthesis and spends its learning budget on the non-Gaussian remainder. Pathwise conditioning has no analogue of the Nataf marginal transform, the proper-score objective, or the cascade.

Nonlinear multi-fidelity Gaussian processes. NARGP (Perdikaris et al., 2017) replaces the linear cross-fidelity term of Kennedy & O'Hagan (2000) with a Gaussian process whose inputs include the low-fidelity prediction, so that the high-fidelity output can be a nonlinear function of the low-fidelity one. This is the closest classical statement of "kriging plus a learned nonlinear correction" in the multi-fidelity literature. It differs from NCS in three ways: the conditional it returns is Gaussian, so it cannot represent the heteroscedastic or heavy-tailed innovations that motivate us; it is a pointwise regression model, not a field sampler with a prescribed spatial covariance; and its nonlinearity is fitted by marginal likelihood, not by a proper score on samples. NCS can be read as the field-scale, non-Gaussian, sample-based analogue.

One-pass generators trained by scoring rules. Minimizing the energy score, or a spectral or weighted variant of it, over a minibatch of generated samples gives a stable, likelihood-free training signal for an implicit conditional generator. Gritsenko et al. (2020) use a spectral energy distance for parallel waveform synthesis; Pacchiardi et al. (2024) train generative networks by prequential scoring-rule minimization for chaotic systems and a global weather benchmark; Chen et al. (2024) do so for multivariate ensemble post-processing; Shen & Meinshausen (2025) study the same estimator as a general distributional-regression method and analyze its extrapolation behavior; and PNO (Bülte et al., 2025) applies it to neural operators. Our NCS-ES arm is a member of this class: a network maps a latent draw and a conditioning field to a sample, and the energy score is the loss. Three things are different. First, the latent is not white noise but a residual-calibrated Gaussian (or transformed-Gaussian) innovation with the empirical Schur-complement spectrum (Eq. 22), so the source alone already has the right second-order geometry. Second, the network is a correction on top of an explicit kriging term and is initialized at the identity, so the untrained model is the exact classical sampler and the trained model's departure from it can be reported as a paired difference on every metric (Eqs. 39–43). Third, because of the first two, the classical sampler appears as a row in every table; none of the works above reports the zero-learning conditional sampler as a baseline, and PNO in particular is compared with NCS on that footing in Table 1. Lemma 2 is also directly relevant to this class: a generator whose additive source is fitted to the marginal residual covariance, rather than the innovation covariance, is over-dispersed by a computable factor before training begins.

Learned conditional simulation in geostatistics. GANs have been used as geological priors for inversion (Laloy et al., 2018; Mosser et al., 2020) and, with progressive growing and condition-based losses, as direct conditional facies simulators (Song et al., 2021). These are conditional simulators in the geostatistical sense and share our problem statement. They replace the classical two-point simulator wholesale with a learned one, are trained adversarially, and are evaluated against training images and well data rather than against the Gaussian conditional sampler. NCS is the complementary strategy: retain the two-point simulator, learn only what it cannot represent, and report the retained simulator's performance as the floor.

Stochastic precipitation downscaling. Leinonen et al. (2021) introduce a recurrent stochastic super-resolution GAN for time-evolving atmospheric fields; Harris et al. (2022) extend it to forecast-to-radar downscaling over the UK with a CRPS-augmented loss and evaluate with CRPS, rank histograms, and spectra; Price & Rasp (2022) train a corrector GAN for global ensemble precipitation over CONUS; and CorrDiff (Mardani et al., 2025) replaces the GAN with a residual diffusion model. These papers define the benchmark family and much of the evaluation protocol we adopt for GARD-LENS. They differ from NCS in the generator (adversarial or diffusion rather than scoring-rule-trained on a classical source), in the absence of a zero-learning conditional-simulation baseline, and in their conditioning field, which is a coarse forecast rather than a coarsening of the target; we flag the consequences of our own choice in Appendix E.2.

One-step samplers and multi-resolution generators. Consistency models (Song et al., 2023), MeanFlow (Geng et al., 2025), and drifting models (Deng et al., 2026; Esteban-Casadevall et al., 2026) obtain one-step sampling by distillation or by evolving the pushforward during training. NCS is one pass per level because Eq. 1 is; the optional drifting warm-start (§3) is used as an optimizer for g_θ, and whether it changes the trained model at a matched training budget is the question the protocol of Appendix F.4 is designed to answer. Laplacian-pyramid GANs (Denton et al., 2015) and cascaded diffusion (Ho et al., 2022) compose generators across resolutions, conditioning each level on an upsampled version of the previous one. Our cascade shares that structure. It differs in that

each level is anchored to a classical simulator with its own fitted innovation spectrum, and in the measured fact (Appendix E.1.1) that, unless the coarse field is tightly pinned, most of the conditional variance a level must supply lies at wavenumbers the coarser level already represents, which a pure resolution-doubling reading of the cascade would not predict. The cross-level factorization needed for exactness of the composed law is stated in Appendix C; Appendix F.5 gives the protocol that tests it.

C FULL THEORY DERIVATIONS

This appendix gives the full statements, proofs, and scope discussion for the five-step theoretical spine summarized in Section 4. Fields are discretized: δ (the fine-scale residual) and c (the coarse conditioning state) are random vectors in $\mathbb{R}^{n_\delta}$ and $\mathbb{R}^{n_c}$ with finite second moments on a common probability space. Joint Gaussianity is assumed *only* in Step 3. Conditional statements hold for P_c-almost every c.

STEP 1: THE CONDITIONAL MEAN IS A PROJECTION—AND WHY IT UNDERDISPERSES

Proposition 1 (L_2 projection)**.** *Among all $\sigma(c)$-measurable, square-integrable predictors $g(c)$, the risk*

$$\mathbb{E}\|\delta - g(c)\|^2$$

is uniquely minimized, up to P_c-almost-sure equality, by the conditional mean

$$m(c) = \mathbb{E}[\delta \mid c],$$

the orthogonal projection of δ onto $L^2(\sigma(c))$.

Corollary 1 (Underdispersion of point prediction)**.** *The law of total covariance gives*

$$\mathrm{Cov}(\delta) = \mathrm{Cov}\big(m(c)\big) + \mathbb{E}[\mathrm{Cov}(\delta \mid c)]\,. \tag{6}$$

A deterministic refiner that outputs $m(c)$ has zero conditional spread and reproduces only the first term in Eq. equation 6. Its marginal covariance is therefore deficient by exactly

$$\mathbb{E}[\mathrm{Cov}(\delta \mid c)]\,,$$

which is positive semidefinite and is nonzero whenever δ is not $\sigma(c)$-measurable.

Corollary 1 is the quantitative form of the familiar observation that L_2-trained refiners are too smooth to serve as conditional simulations: the omitted uncertainty is not an architectural artifact but the conditional covariance discarded by point prediction. The scale distribution of this deficit is determined by $\mathbb{E}[\mathrm{Cov}(\delta \mid c)]$; whether it is concentrated at unresolved fine scales is a property of the underlying problem, which we verify empirically in §5.

STEP 2: AFFINE PROJECTION AND THE INNOVATION

Restrict the predictor class to affine maps $g(c) = a + Bc$. Let

$$\mu_\delta = \mathbb{E}\delta, \qquad \mu_c = \mathbb{E}c,$$

and denote the covariance blocks by $\Sigma_{\delta\delta}$, $\Sigma_{\delta c}$, $\Sigma_{c\delta}$, and Σ_{cc}. Assuming $\Sigma_{cc} \succ 0$, the affine risk minimizer is the Wiener–kriging predictor

$$m_{\mathrm{lin}}(c) = \mu_\delta + K(c - \mu_c), \qquad K = \Sigma_{\delta c}\Sigma_{cc}^{-1}. \tag{7}$$

If Σ_{cc} is singular, the same statement holds on its effective range using the Moore–Penrose pseudoinverse, subject to the usual covariance range conditions.

Definition 1 (Linear innovation)**.** *The innovation relative to the affine predictor is*

$$r := \delta - m_{\mathrm{lin}}(c). \tag{8}$$

Lemma 1 (Innovation properties)**.** *The innovation satisfies*

$$\mathbb{E}r = 0, \tag{9}$$

$$\mathrm{Cov}(r, c) = 0, \tag{10}$$

$$\mathrm{Cov}(r) = \Sigma_r := \Sigma_{\delta\delta} - \Sigma_{\delta c}\Sigma_{cc}^{-1}\Sigma_{c\delta}. \tag{11}$$

Its conditional mean is

$$\mathbb{E}[r \mid c] = m(c) - m_{\mathrm{lin}}(c). \tag{12}$$

Thus, outside the jointly Gaussian case, r is generally only uncorrelated *with c: its conditional law may have a nonzero condition-dependent mean, condition-dependent covariance, skewness, tails, and spatial dependence.*

Lemma 1 marks the frontier of the linear-Gaussian conditional simulator. It is exact when the innovation is independent of c with a fixed Gaussian Schur-complement law. NCS learns departures from that baseline, including both the nonlinear mean correction $m(c) - m_{\mathrm{lin}}(c)$ and higher-order, condition-dependent innovation structure.

STEP 3: EXACT GAUSSIAN CONDITIONAL SIMULATION

Proposition 2 (Exactness of Gaussian conditioning by kriging)**.** *Let (δ, c) be jointly Gaussian with $\Sigma_{cc} \succ 0$, and let K and Σ_r be defined by Eqs. equation 7–equation 11. Then the following two constructions produce exact conditional samples.*

(i) ***Innovation form.*** *If*

$$r_0 \sim \mathcal{N}(0, \Sigma_r)$$

independently of c, then

$$\hat{\delta} = m_{\mathrm{lin}}(c) + r_0 \tag{13}$$

satisfies

$$\mathcal{L}(\hat{\delta} \mid c) = \mathcal{L}(\delta \mid c)$$

for P_c-almost every c.

(ii) ***Substitution form.*** *If (δ_0, c_0) is an independent draw from the joint law of (δ, c), independent of the conditioning value c, then*

$$\hat{\delta}_{\mathrm{M}} = \delta_0 + K(c - c_0) \tag{14}$$

has the same conditional law. The equivalence with the innovation form follows from

$$\delta_0 + K(c - c_0) = m_{\mathrm{lin}}(c) + \big[\delta_0 - m_{\mathrm{lin}}(c_0)\big]. \tag{15}$$

Proof sketch. Under joint Gaussianity,

$$\mathcal{L}(\delta \mid c) = \mathcal{N}(m_{\mathrm{lin}}(c), \Sigma_r)\,.$$

By Lemma 1 and joint Gaussianity, the innovation is independent of c and has law $\mathcal{N}(0, \Sigma_r)$, which proves part (i). For part (ii), Eq. equation 15 rewrites the substitution sampler in innovation form, and $\delta_0 - m_{\mathrm{lin}}(c_0)$ has law $\mathcal{N}(0, \Sigma_r)$ independently of the target conditioning value c. □

Equation equation 14 is Matheron's substitution rule, generalized here to conditioning on a jointly modeled auxiliary low-fidelity field. When c is not literally an observation operator applied to δ, the construction may equivalently be viewed as Gaussian conditional co-simulation. Both forms are one-pass samplers: the innovation form requires simulation from the Schur-complement covariance, whereas the substitution form requires a joint unconditional simulator for (δ, c).

Lemma 2 (Overdispersion of an additive marginal-residual source)**.** *Under the assumptions of Proposition 2, consider the specific additive sampler*

$$\hat{\delta}_{\mathrm{marg}} = m_{\mathrm{lin}}(c) + r_0, \qquad r_0 \sim \mathcal{N}(0, \Sigma_{\delta\delta}) \quad \textit{independently of } c. \tag{16}$$

Then

$$\mathrm{Cov}(\hat{\delta}_{\mathrm{marg}} \mid c) = \Sigma_{\delta\delta} \tag{17}$$

$$= \Sigma_r + K\Sigma_{cc}K^\top \succeq \Sigma_r. \tag{18}$$

Thus its conditional covariance exceeds the Gaussian target covariance by exactly

$$K\Sigma_{cc}K^\top = \mathrm{Cov}\big(m_{\mathrm{lin}}(c)\big),$$

the covariance already explained by the affine conditional predictor.

Lemma 2 applies specifically when a source fitted to the marginal residual covariance is added unchanged and independently to the conditional predictor. A learned transport can in principle shrink or otherwise transform such a source, so the lemma is not a claim about every pipeline initialized from marginal residual noise. It instead shows that the additive identity-source baseline is overdispersed by construction and that a learned model starting from this source must first undo variability already explained by the condition. We therefore fit the default source to empirical innovations

$$r = \delta - m_{\mathrm{lin}}(c)$$

and report the linearly explained covariance fraction, including its spectral distribution, for each benchmark.

Corollary 2 (Stationary spectral representation)**.** *Suppose (δ, c) are jointly second-order stationary, mean-zero fields represented on a common periodic grid, with c prolonged to the target resolution before the joint spectra are estimated. Let $S_{\delta\delta}(k)$, $S_{\delta c}(k)$, $S_{c\delta}(k)$, and $S_{cc}(k)$ denote their spectral-density matrices. At each wavenumber k for which $S_{cc}(k)$ is invertible, the population Wiener filter and innovation spectrum are*

$$K(k) = S_{\delta c}(k)S_{cc}(k)^{-1}, \tag{19}$$

$$S_r(k) = S_{\delta\delta}(k) - S_{\delta c}(k)S_{cc}(k)^{-1}S_{c\delta}(k). \tag{20}$$

These quantities are matrix-valued for multichannel fields. A Gaussian innovation sample is obtained by Fourier synthesis using a matrix square root of $S_r(k)$ with Hermitian symmetry enforced.

The identities in Corollary 2 are population formulas. In finite data, small eigenvalues of $\widehat{S}_{cc}(k)$ require regularization. For the Tikhonov-regularized filter

$$\widehat{K}_\lambda(k) = \widehat{S}_{\delta c}(k)\big(\widehat{S}_{cc}(k) + \lambda I\big)^{-1}, \tag{21}$$

the covariance of the corresponding empirical innovation

$$\hat{r}_\lambda = \delta - \widehat{K}_\lambda c$$

is not, in general, given by replacing S_{cc}^{-1} with $(S_{cc} + \lambda I)^{-1}$ in the Schur complement. The correct residual spectrum is

$$\widehat{S}_{r,\lambda}(k) = \widehat{S}_{\delta\delta}(k) - \widehat{K}_\lambda(k)\widehat{S}_{c\delta}(k) - \widehat{S}_{\delta c}(k)\widehat{K}_\lambda(k)^* + \widehat{K}_\lambda(k)\widehat{S}_{cc}(k)\widehat{K}_\lambda(k)^*. \tag{22}$$

Here * denotes conjugate transpose. We estimate the joint block spectrum with Hermitian positive-semidefinite regularization and construct the source from Eq. equation 22, followed only by numerical symmetrization and clipping of negligible negative eigenvalues caused by estimation or floating-point error. Consequently, exactness belongs to the population unregularized construction; the regularized spectral implementation is an estimator whose approximation error is assessed empirically in Appendix E.3.

STEP 4: CONDITIONAL TRANSPORT GENERALIZES THE INNOVATION LAW

Outside joint Gaussianity, $\mathcal{L}(r \mid c)$ need not be independent of c or Gaussian. The general object is therefore a conditional transport.

Lemma 3 (Noise outsourcing; Kallenberg, 2021, transfer theorem)**.** *Let (δ, c) take values in finite-dimensional Euclidean spaces, hence standard Borel spaces. There exist a measurable map*

$$T : \mathbb{R}^{n_c} \times \mathbb{R}^{n_\delta} \to \mathbb{R}^{n_\delta}$$

and a random variable

$$Z \sim \mathcal{N}(0, I_{n_\delta})$$

independent of c such that

$$(c, \delta) \overset{d}{=} \big(c, T(c, Z)\big). \tag{23}$$

The dimension n_δ in Lemma 3 is a convenient choice rather than a minimality claim: the standard transfer theorem may be formulated using a single nonatomic uniform random variable, which can be constructed from one coordinate of Z.

NCS parameterizes the conditional law in innovation coordinates:

$$\hat{\delta}_\theta = m_{\text{lin}}(c) + T_\theta(r_0, c), \qquad r_0 \sim \mathcal{N}(0, \widehat{\Sigma}_r^\eta) \quad \text{independently of } c, \tag{24}$$

where

$$\widehat{\Sigma}_r^\eta = \widehat{\Sigma}_r + \eta I \succ 0$$

is a regularized empirical innovation covariance and $T_\theta(r_0, c) = r_0$ at initialization.

Because $\widehat{\Sigma}_r^\eta \succ 0$, r_0 is an invertible affine reparameterization of $Z \sim \mathcal{N}(0, I_{n_\delta})$. Thus, at the level of arbitrary measurable transports, replacing Z by the full-rank Gaussian innovation source loses no representational generality. This is a measurable existence statement only: it does not imply that a finite continuous neural network can approximate every such transport under a specified topology or with finite data. Approximation by the chosen network architecture remains a model-capacity and optimization question assessed empirically.

In the population Gaussian case, if $\widehat{\Sigma}_r^\eta$ is replaced by the exact Σ_r and $T_\theta(r_0, c) = r_0$, Eq. equation 24 reduces to the exact innovation sampler of Proposition 2. With empirical covariance estimation and positive regularization η, the identity transport is instead an approximation to that Gaussian sampler.

Remark 1 (Non-identifiability of the transport map)**.** *The transport in Lemma 3 is not unique. For any measurable, possibly condition-dependent transformation $\varphi(\cdot; c)$ preserving the source law,*

$$\varphi(r_0; c) \overset{d}{=} r_0,$$

the maps

$$T(r_0, c) \quad \textit{and} \quad T(\varphi(r_0; c), c)$$

induce the same conditional output law. Geometric properties of a trained T_θ, such as its distance from the identity or from its best affine approximation, are therefore not intrinsic functionals of the data distribution.

We consequently make no claim that the network's apparent nonlinearity measures non-Gaussianity. The value of learning is measured distributionally as the improvement over the fitted Gaussian conditional baseline, reported separately for proper scores, conditional-variance error, tail ratios, and extreme-quantile reliability in §5.

A transformed-Gaussian or Nataf innovation source is treated as a distinct classical non-Gaussian baseline. For such a source, the identity transport is not the Gaussian sampler of Proposition 2; the exact Gaussian fallback applies specifically to the Gaussian Schur-complement source.

Step 5: Proper scoring rules identify the conditional law

Proposition 3 (Population identification)**.** *Let S be a strictly proper scoring rule on a class $\mathcal{P}$ of probability laws on $\mathbb{R}^{n_\delta}$. Suppose that:*

1. *the regular conditional law $p(\cdot \mid c)$ and model law $Q_\theta(\cdot \mid c)$ belong to $\mathcal{P}$ for P_c-almost every c;*

2. *the conditional score risks are measurable and integrable, so that the population risk below is well-defined; and*

3. *there exists at least one parameter value $\theta^\star$ satisfying*

$$Q_{\theta^\star}(\cdot \mid c) = p(\cdot \mid c) \qquad P_c\text{-a.e.}$$

Then the population risk

$$\mathcal{R}(\theta) = \mathbb{E}_c \mathbb{E}_{y\sim p(\cdot|c)} \left[S\big(Q_\theta(\cdot \mid c), y\big) \right] \tag{25}$$

is minimized precisely by parameters whose induced conditional distributions satisfy

$$Q_\theta(\cdot \mid c) = p(\cdot \mid c) \qquad P_c\text{-a.e.} \tag{26}$$

The minimizing parameter values and transport maps need not be unique.

Proof sketch. For every fixed c, strict propriety gives

$$\mathbb{E}_{y\sim p(\cdot|c)} S\big(Q(\cdot \mid c), y\big) \geq \mathbb{E}_{y\sim p(\cdot|c)} S\big(p(\cdot \mid c), y\big),$$

with equality if and only if $Q(\cdot \mid c) = p(\cdot \mid c)$. Integrating this nonnegative pointwise difference over P_c yields the population inequality. Under the representability assumption, the lower bound is attained; almost-everywhere uniqueness of the minimizing distribution follows from strict propriety. This identifies the induced conditional law, not the parameterization or transport map. □

We instantiate S as the energy score, which is strictly proper on laws with finite first moment on $\mathbb{R}^{n_\delta}$ (Székely & Rizzo, 2005; Gneiting & Raftery, 2007), with an extension to suitable classes of laws on separable Hilbert spaces given by Bülte et al. (2025). For a fixed observation y and i.i.d. ensemble members $X_1, \ldots, X_M \sim Q_\theta(\cdot \mid c)$, the estimator

$$\widehat{\mathrm{ES}} = \frac{1}{M} \sum_{m=1}^{M} \|X_m - y\| - \frac{1}{2M(M-1)} \sum_{m\neq n} \|X_m - X_n\| \tag{27}$$

is unbiased for $\mathrm{ES}(Q_\theta(\cdot \mid c), y)$. This is a statement about the estimated score value; unbiasedness of gradients additionally requires the usual conditions permitting differentiation under the expectation.

When K conditionally independent observations $y^{(1)}, \ldots, y^{(K)}$ are available for the same condition, we use

$$\widehat{\mathrm{ES}}_K = \frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} \|X_m - y^{(k)}\| - \frac{1}{2M(M-1)} \sum_{m\neq n} \|X_m - X_n\|. \tag{28}$$

With $K = 1$, the population risk estimate remains unbiased when (c, y) is sampled from the joint data law; larger K reduces the conditional Monte Carlo variance but is not required for validity.

Corollary 3 (The Gaussian conditional law attains the optimum)**.** *Under the population assumptions of Proposition 2, let the NCS source have the exact covariance Σ_r and let*

$$T_\theta(r_0, c) = r_0.$$

Then the induced law satisfies

$$Q_\theta(\cdot \mid c) = p(\cdot \mid c) \qquad P_c\text{-a.e.},$$

and therefore attains the population minimum of every strictly proper scoring rule for which the required expectations exist. This is a containment statement: the minimizing conditional distribution is unique by strict propriety, whereas the minimizing map is not unique by Remark 1.

Strict propriety is a population identification property and does not imply that a finite-sample score is equally sensitive to every aspect of a high-dimensional spatial distribution. In particular, the energy score can have weak practical discrimination of dependence and tail errors in high dimensions. We therefore supplement it with spectral, variogram, amplitude-conditional variance, occurrence, and extreme-value diagnostics. For zero-inflated fields we additionally consider a threshold-weighted score. The transformed score is proper for the induced feature distribution and is strictly proper only up to distinctions retained by its chaining transformation (Allen et al., 2023); we do not interpret it as identifying features that the transformation discards.

SCOPE OF THE SINGLE-LEVEL RESULT AND MULTILEVEL COMPOSITION

Propositions 2 and 3 concern one conditional transition. Exactness of every level separately does not by itself imply that independently composing the levelwise samplers yields the correct finest-level conditional law. Such a conclusion additionally requires an appropriate factorization, for example

$$p(u_1, \dots, u_L \mid u_0) = \prod_{\ell=0}^{L-1} p(u_{\ell+1} \mid u_\ell), \tag{29}$$

which holds when the current level is sufficient for the coarser history, together with consistency of generated states under the restriction and prolongation operators. If cross-level innovations remain dependent after conditioning on the current state, drawing fresh independent innovations at each level need not reproduce the correct joint cross-scale law.

We therefore use the theory above as a per-level Gaussian anchor, not as an unconditional proof of cascade exactness. The final composed distribution is trained end-to-end with a proper score, and the effects of fresh versus shared cross-level latent variables and of restriction consistency are evaluated empirically in §5.

WHAT THIS THEORY DOES AND DOES NOT ESTABLISH

Our theoretical results form a single argument, running from why a refiner must sample to why training preserves a correct Gaussian answer. Together, they establish five properties of NCS:

1. A deterministic L_2 refiner outputs the conditional mean and omits exactly the mean conditional covariance (Corollary 1).
2. Joint Gaussian conditioning admits an exact embedded one-pass sampler through either Schur-complement innovation simulation or Matheron's substitution rule (Proposition 2).
3. Adding an independent source fitted to the raw marginal residual covariance without transforming it double-counts the linearly explained covariance (Lemma 2).
4. At the level of measurable maps, conditional transport can represent the non-Gaussian, condition-dependent innovation law (Lemma 3).
5. When the target conditional law is representable, a strictly proper score identifies its induced distribution at population level (Proposition 3).

The theory does *not* establish:

1. uniqueness or intrinsic geometric meaning of the learned transport;
2. approximation guarantees for the chosen finite neural architecture;
3. convergence of stochastic optimization with finite ensembles and data;
4. finite-sample consistency of the estimated covariance and spectral operators;
5. exactness of the regularized or stationarity-based empirical Gaussian baseline on nonstationary fields;
6. exactness of a multilevel cascade without the required cross-resolution factorization and consistency conditions;
7. generalization to conditions outside the training distribution; or
8. strong finite-sample sensitivity of the energy score to dependence, tails, and extremes.

These gaps are assessed empirically through held-out innovation diagnostics, global and localized spectral estimates, matched-budget ablations, multiple training seeds, cross-level latent experiments, proper scores, and dedicated conditional-variance and tail-calibration metrics.

D BENCHMARKS, BASELINES, AND METRICS

D.1 BENCHMARKS AND DATA SPLITS

Synthetic stochastic PDEs. We evaluate on seven synthetic systems: stochastic Burgers, Allen–Cahn, turbulent Navier–Stokes, the stochastic heat equation / parabolic Anderson model (SHE–PAM, β=1.0), a shot-noise storm-cell precipitation surrogate, a linear damped-wave system with colored forcing, and a passive scalar advected by Kolmogorov turbulence against an imposed mean gradient. For each condition ξ, the low-fidelity state is a noise-free coarse solve, held fixed while K_{data} independent high-fidelity realizations are generated by resampling the stochastic forcing on the fine grid, sharing the same initial condition as the low-fidelity solve. This provides direct finite-sample access to the conditional law

$$p(u_{\mathrm{HF}} \mid u_{\mathrm{LF}})$$

rather than only one realization per condition. Table 3 reports the domain, boundary conditions, grid hierarchy, forcing law, refinement ratio, number of conditions, and number of realizations per condition. SHE–PAM and storm cells are the taxonomy's reach-limited exemplars (concentrated intermittency islands / heavy-tailed convective cells on a dry background); damped wave and passive scalar are the supply-limited exemplars (linear filtering and ramp–cliff advective transport, respectively, of an otherwise well-behaved source).

GARD–LENS precipitation. GARD–LENS supplies a native ensemble of high-resolution precipitation fields. We use K=8 members per condition as the high-fidelity realizations, with the low-fidelity field the same fields coarse-grained to the 32^2 grid. Each daily field is one condition (1024 conditions total), and we split by *condition*, not by ensemble member, so no condition contributes members to more than one of training, validation, and test. All covariance, spectral, marginal, threshold, and source statistics are estimated from the training split only. The extract is in physical mm day^{-1} (wet-pixel mean 5.46, maximum 520). 20.5% of grid cells are zero in every condition and member (the off-domain ocean mask, which the loader fills with 0), and 0.596 of the remaining pixel-slots are exactly zero.

Gaussian control. To test Proposition 2 independently of neural optimization, we construct a jointly Gaussian pair (δ, c) with known covariance and known conditional law. We consider both:

1. a stationary periodic case whose covariance is exactly diagonalized by the DFT; and
2. a controlled nonstationary case in which the global spectral estimator is intentionally misspecified.

The first tests exact implementation; the second isolates the approximation introduced by stationarity.

Controlled non-Gaussian controls. Starting from the Gaussian control, we introduce one departure at a time:

1. a nonlinear conditional mean;
2. condition-dependent innovation scale;
3. skewed innovations;
4. heavy-tailed innovations;
5. a point mass at zero;
6. condition-dependent spatial correlation; and
7. cross-scale dependence between innovations.

The severity of each departure is varied continuously. These controls identify which baseline first fails and whether NCS recovers the corresponding conditional structure.

Table 3: Benchmark summary. All splits are made at the condition level. K_{data} is the number of observed HF realizations per condition. *Main results (Tables 1 and 2); the large-scale vorticity trajectory ($|k| \leq 8$) is shared across members. †Factorial, cascade, warm-start, and sampling-budget studies.

Benchmark	Domain	Grids	Refinement	Train/val/test conditions	K_{data}	Boundary	Stochastic forcing
Gaussian stationary	$[0,1)^2$	$32, 128$	$4\times$	$204/26/26$	8	periodic	known
Gaussian nonstationary	$[0,1)^2$	$32, 128$	$4\times$	$102/13/13$	8	periodic	known
Burgers	$[0,1)$, t_f=1	$32, 64, 128$	$4\times$	$192/32/32$	8	periodic	additive colored, σ=0.12
Allen–Cahn	$[0,1)^2$	$16, 32, 64$	$4\times$	$96/16/16$	8	periodic	additive colored, σ=0.25
Turbulent Navier–Stokes (large)*	$[0,1)^2$	$64, 128, 256$	$4\times$	$819/102/103$	16	periodic	fine-band forcing resampled
Turbulent Navier–Stokes (small)†	$[0,1)^2$	$32, 64, 128$	$4\times$	$192/32/32$	8	periodic	additive colored, σ=1.0
SHE–PAM (β=1.0)	$[0,1)^2$	$32, 64, 128$	$4\times$	$192/32/32$	8	periodic	multiplicative (PAM), β=1.0
Storm cells	$[0,1)^2$	$32, 64, 128$	$4\times$	$192/32/32$	8	periodic	Poisson shot-noise cells, rate 30
Damped wave	$[0,1)^2$	$32, 64, 128$	$4\times$	$192/32/32$	8	periodic	additive colored, σ=0.6
Passive scalar	$[0,1)^2$	$32, 64, 128$	$4\times$	$192/32/32$	8	periodic	additive colored (velocity), σ=1.0
GARD–LENS	CONUS, 12 km	$32, 128$	$4\times$	$768/128/128$	8	nonperiodic	native ensemble

D.2 BASELINE LADDER

Every principal table contains the following ladder. All learned models use the same backbone and conditioning interface where applicable. Architecture, training data, ensemble size, and optimization budget are held fixed unless the comparison explicitly studies one of them.

Point-prediction baselines.

1. **LIN: affine/Wiener predictor.**
$$\hat{\delta} = m_{\mathrm{lin}}(c).$$
This is the best fitted affine point predictor.
2. **NN-L_2: nonlinear conditional mean.** A deterministic neural refiner trained by squared error. This separates gains from a nonlinear conditional mean from gains due to probabilistic simulation.

Source-only and classical conditional-simulation baselines.

3. **MARG: marginal-residual additive source.**
$$\hat{\delta} = m_{\mathrm{lin}}(c) + \varepsilon_\delta, \qquad \varepsilon_\delta \sim \mathcal{N}(0, \widehat{\Sigma}_{\delta\delta}).$$
This directly tests Lemma 2.
4. **GCS: Gaussian conditional simulation.**
$$\hat{\delta} = m_{\mathrm{lin}}(c) + r_0, \qquad r_0 \sim \mathcal{N}(0, \widehat{\Sigma}_r).$$
This is the Gaussian innovation form of Matheron conditional simulation. Where a joint unconditional simulator is available, we additionally verify its numerical equivalence to the substitution form $\delta_0 + \widehat{K}(c - c_0)$.
5. **HET-GCS: heteroscedastic Gaussian conditional simulation.**
$$\hat{\delta} = m_{\mathrm{lin}}(c) + s_\phi(c) \odot r_0.$$
The scale model $s_\phi(c)$ depends only on variables available at inference. This baseline tests whether condition-dependent second moments, without a general neural transport, explain the NCS gain.
6. **TGCS: transformed-Gaussian conditional simulation.** A Gaussian-copula/Nataf transformation is applied to the innovation source to fit non-Gaussian marginal structure. For precipitation, occurrence and positive intensity are optionally modeled separately. This is the strongest classical non-Gaussian baseline.

Learned conditional transports.

7. **NCS-ES: direct proper-score training.** The Gaussian innovation source is passed through $T_\theta(r_0, c)$ and trained from the identity initialization using the energy score.

8. **NCS-TG-ES: transformed-source NCS.** The same transport is initialized from the transformed-Gaussian innovation source. This tests whether source-side marginal calibration remains useful after learned conditional transport.
9. **NCS-twES: tail-weighted training.** NCS is trained with the energy score plus a preregistered threshold-weighted score. This tests whether explicit tail emphasis improves extreme calibration without degrading bulk calibration.
10. **NCS-Drift+ES: drifting warm-start.** The same NCS model is pretrained using conditional drifting and then optimized with the same proper-score objective. This baseline is compared with NCS-ES at both equal fine-tuning steps and equal *total* training compute.

Modern generative baselines.

11. **Conditional diffusion.** A CorrDiff-style conditional mean plus stochastic residual model, evaluated at its default noise schedule in Table 1; its sensitivity to the noise range and step count on GARD-LENS is in Table 19.
12. **One-step distilled generator.** A consistency or equivalent one-step distilled baseline using the same conditioning information.
13. **PNO: probabilistic neural operator.** A generative neural operator (Bülte et al., 2025) with an FNO backbone, made stochastic by concatenating latent Gaussian-noise channels to the conditioning field and trained by the same energy score at a single forward pass per member. Because it shares NCS's objective but replaces the classical-source cascade with a pure operator, it isolates the contribution of the source-and-transport construction from that of proper-score training: any tail behavior it shares with NCS-ES is attributable to the objective rather than the architecture.

Table 4: Baseline ladder. "Passes" reports inference-time neural evaluations per ensemble member for a single level; measured compute is reported separately in §F.7.

Method	Conditional mean	Stochastic source	Learned transport	Proper-score trained	Classical fallback	Passes
LIN	affine	none	none	no	yes	0
NN-L_2	neural	none	none	no	no	1
MARG	affine	raw residual	identity	no	no	0
GCS	affine	Gaussian innovation	identity	no	exact if Gaussian	0
HET-GCS	affine	scaled Gaussian innovation	scale only	no	no	0
TGCS	affine	transformed innovation	identity	no	classical	0
NCS-ES	affine + learned	Gaussian innovation	neural	ES	contains GCS	1
NCS-TG-ES	affine + learned	transformed innovation	neural	ES	contains TGCS	1
NCS-twES	affine + learned	[best source]	neural	ES + twES	no	1
NCS-Drift+ES	affine + learned	[best source]	neural	ES	no	1
Diffusion	neural	Gaussian	iterative	diffusion	no	T
One-step distilled	neural	Gaussian	neural	distillation	no	1

The learned method is considered justified only if it improves over both GCS and the strongest source-only classical baseline. Drifting is retained as a method component only if it improves convergence, final quality, or seed-to-seed stability at matched total training compute.

D.3 Evaluation metrics

No single metric is treated as synonymous with calibration. We report accuracy, aggregate second-moment calibration, conditional calibration, dependence, spectra, tails, and sampling cost separately.

Point accuracy. We report the RMSE of the ensemble mean and its normalized form (nRMSE). These measure conditional-mean accuracy and do not assess ensemble calibration.

Proper scores. We report the multivariate energy score, pixelwise CRPS (aggregated only after computing it per coordinate), and the variogram score for spatial dependence. All scores are evaluated at the same ensemble size, $M = 32$.

Spread–skill. We report both raw and Fortin-corrected spread–skill ratios, with the exact finite-ensemble correction defined in Appendix D.4. A global ratio near one is interpreted only as aggregate second-moment calibration.

Conditional variance calibration. Conditions are binned by an inference-available coarse predictor, the local coarse amplitude, and in each bin b we compute

$$R_{\mathrm{var}}(b) = \frac{\mathbb{E}[\widehat{\mathrm{Var}}_{\mathrm{model}}(\delta \mid c) \mid c \in b]}{\mathbb{E}[\widehat{\mathrm{Var}}_{\mathrm{data}}(\delta \mid c) \mid c \in b]}. \tag{30}$$

This exposes amplitude-dependent miscalibration that a global scalar ratio can conceal.

Rank and coverage diagnostics. On the Gaussian controls, where the conditional law is known, we report randomized rank histograms and central 90% interval coverage (Appendix E.3).

Spectral and spatial dependence. We report the radially averaged log-power-spectrum error, the innovation–condition coherence (Appendix E.2), variogram scores, and, for multilevel samples, the cross-level residual coherence.

Tails. We report the increment-kurtosis ratio $\kappa/\kappa_{\mathrm{data}}$ together with the data's own κ, and the 99th percentile of the standardized increment pool.

Statistical uncertainty. All learned configurations carrying a substantive conclusion are trained with at least 3 independent seeds. Tables report mean $\pm$ s.d. over seeds or condition-bootstrap intervals, as stated in each caption. Hyperparameters are selected on validation conditions and frozen before test evaluation.

D.4 Metric conventions

This appendix fixes the two conventions that a reader cannot recover from a column name: the finite-ensemble spread–skill correction promised in Section D.3, and what the *NFE* column of Tables 1 and 10 counts. Throughout, M is the number of generated members and K the number of held-out data members per condition; every run in this paper uses $M = 32$ and $K = 8$.

Spread–skill, raw and corrected. Let $\{x_m\}_{m=1}^{M}$ be the generated members for one condition, $\bar{x} = M^{-1}\sum_m x_m$ their mean, and $\{y_k\}_{k=1}^{K}$ the held-out data members. We report

$$\sigma^2 = \frac{1}{M-1}\sum_{m=1}^{M}\|x_m - \bar{x}\|^2, \qquad \text{(unbiased ensemble variance)} \tag{31}$$

$$s^2 = \frac{1}{K}\sum_{k=1}^{K}\|\bar{x} - y_k\|^2, \qquad \text{(skill against } \textit{individual} \text{ members)} \tag{32}$$

averaged over coordinates and conditions, and then the two ratios

$$r_{\mathrm{raw}} = \frac{\sigma}{s}, \qquad r_{\mathrm{c}} = \sqrt{\frac{M+1}{M}}\,\frac{\sigma}{s}. \tag{33}$$

The inflation factor is the standard finite-ensemble verification relationship (Fortin et al., 2014): for a calibrated ensemble with no mean bias, $\mathbb{E}[s^2] = \sigma^2(M+1)/M$, so $r_{\mathrm{c}} \to 1$ while r_{raw} sits below it. At $M = 32$ the factor is $\sqrt{33/32} = 1.0155$, so the *only* difference between the two reported columns is 1.55% — they are printed together because the convention, not the gap, is what needs to be auditable.

Two neighbouring quantities are *not* what we report, and both read very differently:

1. Skill measured against the K-member *data mean*, $\|\bar{x} - \bar{y}\|$, rather than against individual members. That reference is itself a noisy estimate of the conditional mean, so a perfectly calibrated ensemble reads $(1/M + 1/K)^{-1/2} = 2.53$ at $M = 32, K = 8$, not 1. A ratio of that form must never be read as calibration, and in particular an arm at 2.75 there can be *under*-dispersed.

2. A deterministic arm has $\sigma = 0$ exactly, so both ratios are a genuine 0.0 rather than undefined. Maximal under-dispersion is a finding and is reported as such; the kurtosis and P99 ratios, which are $0/0$ for those arms, are the ones printed as "–".

Cost. The *NFE* column counts network function evaluations per generated member: the number of forward passes through a trained network needed to draw one member for one condition, summed over cascade levels for the multi-level arms. Zero is a real value — LIN, bilinear, and the four classical rows draw a member in closed form, with no forward pass at all.

We report NFE rather than wall-clock because wall-clock is not comparable across these rows: the classical arms sample in NUMPY on the CPU while the neural arms sample on a GPU, and the runs span two hosts and three months. Per-sample sampling times are carried in the released CSV for completeness.

Two caveats matter when reading a single NFE against another:

1. **Deterministic arms amortize.** An arm that broadcasts one prediction to all M members pays its NFE once per condition, not once per member, so its marginal member is free and its per-condition cost does not grow with ensemble size the way every probabilistic row's does. The released CSV reports the marginal and per-condition figures separately.
2. **One arm has a shared stage.** CorrDiff evaluates its regression network once per condition and reuses that mean across members, so its per-condition cost is $M \cdot \mathrm{NFE} + 1$.

NFE is a declared property of each architecture, so it is verified against the samplers rather than trusted: a test counts the batch elements actually pushed through each arm's networks during sampling and requires the declared cost to reproduce that count exactly for 13 of the 16 arms. The other three reuse a tested sampler (MARG and HET-GCS the closed-form GCS path, NCS-twES the NCS sampler). Counting batch *elements* rather than calls is necessary because one arm folds all M members into a single batched call, which a call count would credit as one evaluation.

E DIAGNOSTICS OF THE CLASSICAL MACHINERY

These three experiments test the classical stage of NCS before any learning: Experiment 1 measures the covariance a deterministic refiner omits and where in wavenumber it lives; Experiment 2 tests the innovation-versus-residual distinction of Lemma 2 on a Gaussian control and sizes the non-affine remainder on every benchmark; Experiment 3 verifies that the implemented Gaussian sampler recovers a known conditional law, prices finite data and regularization, and measures the cost of the stationarity assumption against a closed-form ceiling. The main text quotes one number from each; the protocols, falsification criteria, and full tables are here.

E.1 EXPERIMENT 1: PROJECTION AND THE DETERMINISTIC VARIANCE DEFICIT

Question. How much uncertainty is omitted by deterministic refinement, and is the deficit explained by Corollary 1?

Procedure. For each benchmark, we fit LIN and NN-L_2 using training conditions. Splits are contiguous condition blocks. We additionally report an oracle conditional mean, obtained from the between-condition term of the ANOVA decomposition, which upper-bounds what *any* deterministic refiner can reproduce and so separates "this predictor is too weak" from "the information is not in c". On held-out conditions, we estimate:

$$\widehat{\Sigma}_\delta = \widehat{\mathrm{Cov}}(\delta), \tag{34}$$

$$\widehat{\Sigma}_{\mathrm{mean}} = \widehat{\mathrm{Cov}}\big(\hat{m}(c)\big), \tag{35}$$

$$\widehat{\Sigma}_{\mathrm{cond}} = \widehat{\mathbb{E}}\Big[\widehat{\mathrm{Cov}}(\delta \mid c)\Big]. \tag{36}$$

We compare

$$\widehat{\Sigma}_\delta \quad \text{with} \quad \widehat{\Sigma}_{\mathrm{mean}} + \widehat{\Sigma}_{\mathrm{cond}}$$

using trace, Frobenius norm, leading eigenvalues, and spectral decomposition.

For the affine predictor, we report the explained-variance fraction

$$\rho_{\mathrm{lin}} = \frac{\mathrm{tr}\Big(K\widehat{\Sigma}_{cc}K^{\top}\Big)}{\mathrm{tr}\Big(\widehat{\Sigma}_{\delta\delta}\Big)}, \tag{37}$$

as well as the wavenumber-dependent explained fraction. For NN-L_2, we report the analogous empirical decomposition using its held-out predictions.

Equation 37 is computed on the innovation, which makes it sensitive to the prolongation baseline. If $\mathcal{P}c$ is poorly scaled then δ contains a large deterministic multiple of $\mathcal{P}c$, and the fitted filter can approach $K \to -I$ on the coarse band, scoring a high ratio purely by undoing the prolongation rather than by predicting anything. This is not hypothetical: on one benchmark δ carries more variance than the target itself, and the ratio reads 0.55 while the same predictor explains 0.0001 of the target. We therefore report alongside it the held-out residual reduction of y measured against y's own variance—which no baseline correction can inflate, since removing a baseline raises the residual—together with the filter gain, so that the degenerate case is visible in the table.

Expected theoretical signature. The deterministic predictors may improve RMSE, but their conditional spread is identically zero. Their marginal covariance should reproduce only between-condition variation in the conditional mean and omit the within-condition covariance.

A negative result must not be an artifact of the architecture. NN-L_2 is a spectral operator, and a spectral operator truncated at M modes cannot emit structure above wavenumber M. With the conventional $M = 12$ on our 128^2 grid at refinement 4, the truncation sits *below* the coarse Nyquist $k_{\mathrm{nyq}} = 16$: such a network places 99.9994% of its predicted residual power at $k \le 12$ and would score zero on fine-band structure even where that structure is predictable, manufacturing agreement with Corollary 1. This is a capability limit of the default and not of the operator—at $M \ge k_{\mathrm{nyq}}$ the same architecture fits a purely fine-band target from a coarse-band input to $R^2 = 1.00$—so we set $M = 2k_{\mathrm{nyq}}$ and report the fraction of each trained network's residual power that lands above k_{nyq}. A deterministic refiner's failure to supply fine-band spread is only evidence if the refiner could have supplied it.

Primary exhibit. Figure 3 decomposes target variance by wavenumber into predictable and conditional components. Rather than assume the missing uncertainty sits at unresolved scales, we *measure* where it lives, and report two distinct quantities that a single "conditional fraction" curve conflates: the conditional share of each band's own variance, and each band's share of the total deficit. The first is near unity above the coarse Nyquist on every benchmark—fine bands are almost entirely unpredictable from c—while the second is small there whenever the spectrum is red, because the unresolved band holds most of the modes but little of the power. Both are properties of the problem, not of the refinement geometry: we find that where the deficit lives tracks *conditioning strength*, and exhibit a matched pair of Navier–Stokes benchmarks that differ only in how tightly the coarse field is pinned (§E.1.1).

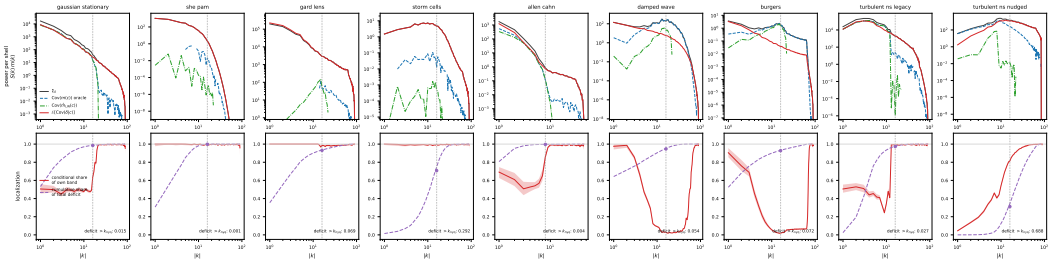


Figure 3: Projection decomposition on held-out conditions, per benchmark. The leftmost panel is the Gaussian stationary control, whose answers are known in closed form (Table 5). Upper row: radially binned spectra of the target innovation Σ_δ, of the *oracle* conditional mean $\mathrm{Cov}(m(c))$, of the fitted affine predictor $\mathrm{Cov}(\hat{m}_{\mathrm{LIN}}(c))$, and of the mean within-condition covariance $\mathbb{E}[\mathrm{Cov}(\delta \mid c)]$; the gap between the first and second is the deficit *no* deterministic refiner can close, and the gap between the second and third is mean structure that is present but not affinely accessible. LIN's power falls away at the edge of the coarse box, which is the support of K and not a fitting artifact. Lower row: the two localization curves that a single "conditional fraction" would conflate— each band's conditional share of *its own* variance (near unity above the coarse Nyquist k_{nyq} on the stochastic systems, but small on the hyperbolic ones where fine structure propagates deterministically), and each band's share of the *total* deficit (concentrated below k_{nyq} wherever the spectrum is red). Both are Parseval-exact: band densities times mode multiplicity sum to the spatial trace. Shaded regions show bootstrap uncertainty over conditions; the rightmost panels are the matched Navier–Stokes pair of §E.1.1, identical but for conditioning strength. The law of total covariance holds per Fourier mode without any stationarity assumption, so these panels are valid on the nonstationary benchmarks; only the Wiener filter of the estimated arms in Figure 5 invokes joint second-order stationarity, and §E.3.7 prices that assumption directly.

Table 5: Projection decomposition on held-out conditions. Closure is $\mathrm{tr}(\widehat{\Sigma}_{\mathrm{mean}} + \widehat{\Sigma}_{\mathrm{cond}})/\mathrm{tr}(\widehat{\Sigma}_\delta)$, which is 1 by Eq. 6 and is therefore an estimator check, not a result. The two localization columns are the distinct quantities of §E.1: the conditional share of each band's own variance, split at the coarse Nyquist, and the share of the *total* deficit lying above it. ρ_{lin} (Eq. 37) is reported beside the median in-box filter gain $1 + K$ and the baseline-free $R^2(y \mid c)$, because a gain near zero means the filter is undoing the prolongation rather than predicting.

			cond. share of own band		deficit location		affine predictor	
Benchmark	n_{test}	closure	$k \le k_{\mathrm{nyq}}$	$k > k_{\mathrm{nyq}}$	share $> k_{\mathrm{nyq}}$	$\bar{k}$	ρ_{lin} / gain	$R^2(y \mid c)$
Gaussian stationary (control)	26	1.017	0.496	0.866	0.015	2.6	0.495 / 2.00	0.821
SHE/PAM	26	1.000	1.000	0.992	0.001	3.2	0.000 / 0.99	0.005
GARD–LENS	103	0.999	1.000	0.985	0.069	5.1	0.001 / 0.73	0.109
Storm cells	26	1.000	0.992	0.994	0.292	12.5	0.000 / 0.82	0.005
Allen–Cahn	13	1.025	0.646	0.972	0.004	1.1	0.038 / 0.64	0.272
Damped wave	26	1.027	0.372	0.025	0.054	3.1	0.049 / 0.77	0.297
Burgers	26	1.027	0.331	0.046	0.072	4.5	0.416 / 0.46	0.950
Navier–Stokes (weak cond.)	26	1.018	0.459	0.993	0.027	4.9	0.048 / 0.71	0.0001
Navier–Stokes (nudged)	103	1.002	0.567	0.936	**0.688**	**25.9**	0.004 / 1.00	0.605

E.1.1 Where the deficit lives is set by conditioning strength

The scale distribution of $\mathbb{E}[\mathrm{Cov}(\delta \mid c)]$ varies by more than an order of magnitude across benchmarks, and the ordering is not explained by the refinement factor, the grid, or the coarsening operator—all shared. It tracks how tightly the conditioning field pins the resolved band. The controlled comparison is a pair of Navier–Stokes datasets, with the large one brought to the small one's grid and member count ($N = 128$, $K = 8$) so that they differ only in that one nudges the coarse-band vorticity trajectory toward the fine solution: the share of the deficit above the coarse Nyquist moves from 0.03 to 0.69, and the deficit-power-weighted mean wavenumber from 4.9 to 25.9. In the nudged system the per-band deficit peaks just above the nudging cutoff k_c, which is the signature

of the mechanism rather than a spectral coincidence. Weakly conditioned problems therefore leave large-scale spread for the ensemble to regenerate, while strongly conditioned ones reduce to genuine fine-scale completion; both are instances of Corollary 1, and neither is visible from the refinement geometry alone.

This has a direct consequence for training objectives. The energy score under the unweighted L_2 norm inherits the spectrum, so on red-spectrum benchmarks the overwhelming majority of what it measures is resolved-band spread, and it is correspondingly insensitive to fine-band conditional structure—on precisely the systems where fine bands are the most conditionally random. The energy score's limited discrimination is known (Pinson & Tastu, 2013), but here it is a measured property of our benchmarks rather than a general caution, supported by the per-band deficit localization of Experiment 1 above.

E.2 Experiment 2: innovation fitting and the marginal-source warning

Question. Does fitting the source to the affine innovation remove the variance double-counting predicted by Lemma 2?

Held-out innovation diagnostics. We estimate $\widehat{K}$ on training data and form held-out innovations

$$\hat{r} = \delta - \hat{\mu}_\delta - \widehat{K}(c - \hat{\mu}_c).$$

We report:

1. held-out innovation bias $\|\widehat{\mathbb{E}}\hat{r}\|$;
2. cross-covariance $\|\widehat{\mathrm{Cov}}(\hat{r}, c)\|$;
3. spectral coherence between $\hat{r}$ and c;
4. conditional innovation mean $\widehat{\mathbb{E}}[\hat{r} \mid c \in b]$;
5. conditional innovation variance by amplitude, location, and season; and
6. skewness, kurtosis, and spatial dependence of $\hat{r}$.

The first three diagnose the fitted affine projection. The remaining statistics quantify departures from the fixed Gaussian innovation assumption.

Direct test of Lemma 2. We compare MARG with GCS using the identity transport. For the Gaussian control, the predicted excess covariance is

$$\Delta_{\mathrm{pred}} = \widehat{K}\widehat{\Sigma}_{cc}\widehat{K}^\top.$$

The observed excess is

$$\Delta_{\mathrm{obs}} = \widehat{\mathrm{Cov}}(\hat{\delta}_{\mathrm{MARG}} \mid c) - \widehat{\mathrm{Cov}}(\delta \mid c).$$

We compare them in trace, spectrum, and Frobenius norm. On real benchmarks, where the Gaussian model is approximate, the same comparison is reported as a diagnostic rather than an exact identity.

Table 6: Held-out innovation diagnostics. Columns: explained variance $1 - \mathrm{tr}\widehat{\Sigma}_r/\mathrm{tr}\widehat{\Sigma}_\delta$ (a baseline correction can only lower it, never inflate it); innovation bias and cross-covariance, each normalized by the innovation's own scale; power-weighted multiple coherence over supported shells; conditional-variance range (max/min over train-fitted amplitude quintiles); and pooled excess kurtosis. Absolute magnitudes accompany these ratios in the released diagnostics.

Benchmark	Explained variance	$\|\mathbb{E}\hat{r}\|$	$\|\mathrm{Cov}(\hat{r}, c)\|$	Mean coherence	Conditional variance range	Kurtosis
Gaussian stationary (control)	0.495	0.040	0.002	0.0001	1.03	0.04
Gaussian nonstationary (cosine gain)	0.407	0.055	0.003	0.0003	1.03	0.11
SHE/PAM	0.101	0.076	0.004	0.0002	1.18	4.52
GARD–LENS	0.142	0.035	0.000	0.0000	6.15	138.38
Storm cells	0.091	0.071	0.001	0.0002	2.04	5508
Allen–Cahn	0.265	0.202	0.027	0.0037	1.71	−0.64
Damped wave	0.049	0.183	0.002	0.0014	1.53	−0.00
Burgers	0.385	0.214	0.019	0.0023	1.29	27.75
Navier–Stokes (weak cond.)	0.523	0.072	0.001	0.0002	1.07	0.39
Navier–Stokes (nudged)	0.094	0.061	0.000	0.0001	1.03	0.35

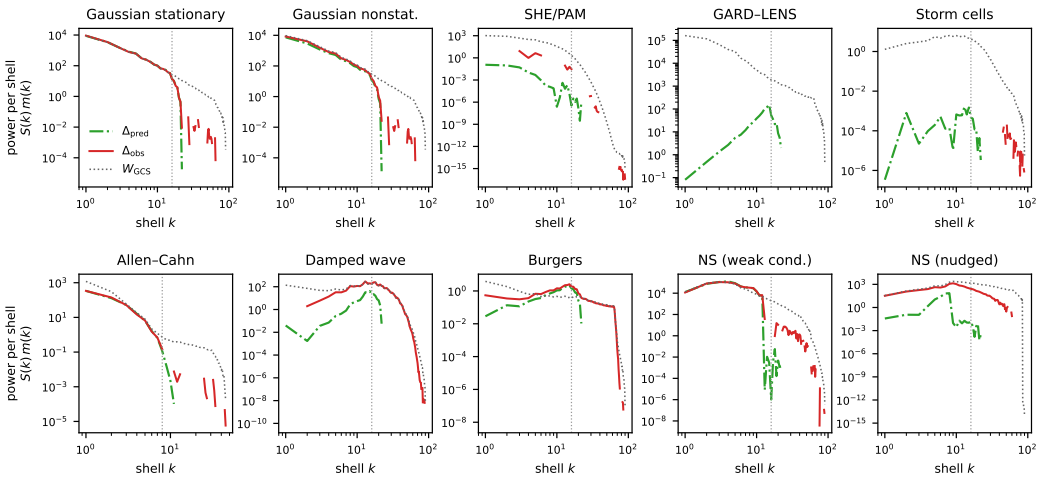


Figure 4: Direct test of Lemma 2, per benchmark (power per radial shell, so the eye integrates the Parseval trace; dotted line: coarse Nyquist k_{nyq}; the ten table benchmarks are shown, the two GARD-LENS split-member controls are reported in the prose and §E.2.2). Each panel plots the predicted excess $\Delta_{\mathrm{pred}} = \widehat{K}\widehat{\Sigma}_{cc}\widehat{K}^{\top}$ against the observed excess Δ_{obs} of a source fitted to the marginal residual (MARG), and the within-spectrum W_{GCS} of the innovation-fitted source (shell 0 excluded on both sides: the sources are DC-free by construction, and the members' spatial-mean share is reported in §E.2.2). On the stationary control Δ_{obs} overlays Δ_{pred} (the identity) and W_{GCS} overlays the data's $\mathbb{E}[\mathrm{Cov}(\delta \mid c)]$ (calibration); on the nonstationary control Δ_{obs} exceeds Δ_{pred} by the planted nonlinear share; on Burgers the MARG excess is roughly double the prediction—the Lemma's double-count plus nonlinear structure. The companion departure detector—the held-out innovation's multiple coherence with c—sits at the stationary-control floor on every benchmark after the affine projection, and is tabulated as the power-weighted multiple-coherence column of Table 6.

E.2.1 The double-count is an identity on the control and measurable on every live row

On the stationary Gaussian control the Lemma's comparison holds as an identity: $\mathrm{tr}\Delta_{\mathrm{obs}}/\mathrm{tr}\Delta_{\mathrm{pred}} = 1.02$ (95% bootstrap interval $[1.00, 1.04]$ over held-out conditions), the innovation-fitted source calibrates ($\mathrm{tr}W_{\mathrm{GCS}}/\mathrm{tr}W_{\mathrm{data}} = 1.01$ $[0.99, 1.04]$), and the marginal-residual source over-disperses by the planted $1 + a^2 = 2.03$. The nonstationary control separates what the columns can each see: for any deterministic modulation $\mathbb{E}[S_{rc}] = 0$ exactly, so the coherence column stays at the stationary control's floor (0.0003 vs 0.0001) while the GCS calibration reads the planted nonlinear excess (1.18 $[1.16, 1.21]$ against the analytic 1.17). Mean structure beyond the stationary filter's reach appears in the variance columns, not the coherence column.

On the live rows the comparison is a diagnostic, made on matched support: the sources zero the DC mode by construction (a per-member spatial mean is not a stationary shell quantity), so shell 0 is excluded from the data's within-condition covariance as well, and its share is reported separately (§E.2.2). The drift-immune form is the train-side ratio $\mathrm{MARG}/\mathrm{GCS} = 1/(1 -$ in-sample affine share): 2.08 on weakly conditioned Navier–Stokes, 1.77 on Burgers, 1.35 on Allen–Cahn, 1.06 on damped wave, and ≤ 1.006 on every row with no affine share. Against the held-out within-condition variance, the innovation-fitted source calibrates where the affine stage is the whole mean structure (Allen–Cahn 1.00 $[0.90, 1.10]$) and over-disperses by exactly the non-affine conditional-mean share elsewhere: Burgers 2.88 $[2.76, 3.03]$ with the marginal source at 5.10, the 2.22 gap being the linearly explained covariance the Lemma removes and the residual 1.88 the conditional-mean structure no affine map reaches—exactly what the transport stage is asked to supply. Damped wave is the extreme case, 7.42 $[7.28, 7.56]$ against 7.84: the affine share is small (gap 0.42) and nearly all of its $7.4\times$ over-dispersion is non-affine mean structure. Nudged Navier–Stokes is the sharpest single exhibit: its $R^2(y \mid c) = 0.60$ is supplied by the nudged baseline, the affine share of the innovation is 0.004, yet its source over-disperses the within by 1.28—a further 28%

of the within variance as non-affine mean structure. Weakly conditioned Navier–Stokes proves the Lemma is a statement about the construction, not about the filter: on the row whose fitted filter is degenerate (gain 0.004, §E.1) the identity still holds ($\mathrm{tr}\Delta_{\mathrm{obs}}/\mathrm{tr}\Delta_{\mathrm{pred}} = 0.96\ [0.94, 0.98]$), the marginal source over-disperses by $2.11\times$, and the innovation-fitted source calibrates at 1.01—the innovation fit removes precisely the double-count that the degenerate filter manufactures. Rows with no affine share have nothing to double-count (MARG and GCS coincide to within 0.006), and their common sub-unity level is block drift. We measure it without any fitted source, as the train/test ratio of the raw within-condition trace, which factors the calibration exactly into a source term and a drift term. The factorization is what makes the reading falsifiable, and it also bounds it: on a row with no affine share the fitted innovation collapses onto the train within, the source term goes to 1, and the calibration becomes numerically *equal* to the drift ratio rather than being corroborated by it. Storm cells is exactly that case—calibration $0.88\ [0.66, 1.13]$, drift $0.880\ [0.66, 1.13]$, source term 1.002—so its agreement is an identity, and all we can say is that its sub-unity level is attributable to drift, not that two independent measurements confirm it. GARD–LENS is the informative one: drift 0.784 $[0.71, 0.88]$ times a source term of 0.892 gives its $0.70\ [0.63, 0.79]$, so roughly three quarters of the shortfall is block drift and the rest is the source under-dispersing against its own training block—consistent with the member-half heterogeneity of the exogenous split controls (§E.2.2), whose drift ratios are 0.634 and 1.044 against calibrations of 0.636 and 1.055. On every row where the affine stage carries structure the drift term is within 0.07 of unity (stationary control 1.014, Burgers 1.010, damped wave 0.997, Allen–Cahn 0.936), so the over-dispersion reported above is not a block artifact.

E.2.2 What the transport stage is asked to close

After the affine projection, no *linear* dependence on c remains anywhere: the power-weighted coherence sits at the control's floor on every row and the cross-covariance norm is ≤ 0.03. What the conditional transport is asked to close is therefore of three kinds, and the diagnostics size each. First, *non-affine conditional-mean power*: the innovation-fitted source over-disperses the within-condition variance by $1.9\times$ on Burgers, $6.4\times$ on damped wave, and 28% on Navier–Stokes (§E.2.1)—second-moment structure the affine stage cannot see, present in the marginal but absent from every condition's own spread. Second, *non-Gaussian shape at matched power*: pooled excess kurtosis runs from 0.04 on the stationary control to 27.8 on Burgers and 5.5×10^3 on storm cells, and the amplitude-conditioned variance ranges from 1.03 (both controls) to 6.15 on GARD–LENS. Third, *per-member spatial-mean structure*: Allen–Cahn members settle at different phase fractions, and each damped-wave member integrates its own velocity-forcing mean into a large-scale offset, so the single DC mode holds for example 50% of the within-condition trace on damped wave and 19% on Allen–Cahn and Burgers. This denotes within-condition structure that a DC-free source under a law-preserving transport can never generate, whatever else it learns. The shell-constant filter's phase content is negligible (phase-loss share below 10^{-3} on every row), so radial shell averaging suffices for the linear stage; per-mode filters would change nothing measurable here.

E.3 Experiment 3: Gaussian exactness and robustness of the classical baseline

Question. Does the implemented Gaussian conditional simulator recover the known conditional law, and how much accuracy is lost through finite-sample covariance estimation and stationarity assumptions?

Oracle Gaussian test. On the stationary Gaussian control, we compare:

1. the analytic conditional law;
2. the innovation-form sampler using the exact Σ_r;
3. the substitution-form sampler using exact joint simulations;
4. the estimated spectral sampler;
5. the regularized spectral sampler; and
6. the marginal-residual additive sampler.

The innovation and substitution forms should agree up to Monte Carlo error. We report:

1. conditional mean error;
2. conditional covariance error;
3. energy distance to the analytic conditional Gaussian;
4. interval coverage;
5. rank uniformity;
6. energy score; and
7. the measured versus predicted MARG overdispersion.

The control and split are those of §E.3.2: $N = 128$, refinement factor 4, $K = 8$, $204/26/26$ conditions, $M = 32$ draws per test condition in double precision.

E.3.1 THE REFERENCE ROW IS THE ACHIEVABLE FLOOR, NOT ZERO

None of these metrics has zero as its attainable value at finite M, so the *analytic* row—an ensemble drawn from the closed-form per-mode law—is scored like every other arm and used as the reference. Two of its readings show why this matters. The conditional-mean error of an exactly correct sampler is $M^{-1/2} = 0.177$, and the analytic row reads 0.192; every other admissible arm lands between 1.03 and 1.09 times that floor, so the correct statement is that they are at the floor, not that they have a 19% error. Likewise the empirical 90% coverage of a 32-member ensemble cannot exceed 0.939, because the 5% quantile interpolates between the first and second order statistic; the analytic row reads 0.857.

Accordingly, each arm is tested for indistinguishability from a reference row rather than against an absolute tolerance, at 4σ of the standard error over test conditions. We also report the smallest deviation the test could have detected, since "indistinguishable" is otherwise a claim about the budget: 0.050 on the covariance trace ratio and 0.023 on coverage.

The two estimated arms are judged against a *binned* analytic reference, and this must be stated plainly because it changes their apparent accuracy by two orders of magnitude. A radially binned source cannot represent a density that varies within a shell, and on this control the true innovation spectrum S_r varies by $2.1\times$ across the eight modes of shell 1 and by $2.7\times$ within shell 2, since the $(1 + |k|)^{-2}$ density and the anisotropic coarse-box support both vary inside a shell. Pushing the *exact* law through the same radial binning already costs a relative Frobenius error of 0.344 at a trace ratio of 1.003—so the shipped estimator's 0.349 is an estimation contribution of $+0.005$ on top of a 0.344 representation cost, not a 35% error. Judged against the per-mode row instead, the estimator would be charged for a discretization no amount of data removes. The per-mode analytic row's own 0.167 is the finite-M floor on the same quantity.

Table 7: **Oracle Gaussian test: the implemented sampler recovers the known conditional law, and MARG fails by the predicted amount.** Stationary Gaussian control, $M = 32$, $n_{\text{test}} = 26$, seed 0; $\pm$ is the standard error over test conditions. *Analytic* (per-mode closed form) and *analytic (binned)* (that same law pushed through the radial binning the estimated arms inherit) are references, not competitors; each metric's target is the applicable reference row's own value, never zero (§E.3.1), which is the floor/cap/target printed under each column heading. The final column is the max $|z|$ over the four metrics with a known target, at a 4σ threshold. MARG's boldface trace ratio (the number Lemma 2 predicts) and italicized coverage (which *exceeds* the finite-ensemble cap) are its two over-dispersion signatures, read out with the energy-distance, energy-score, and fit diagnostics in §E.3.2.

Arm	mean rel. err. (floor 0.177)	tr ratio (target 1)	rel. Frob. err.	cov_{90} (cap 0.939)	rank dev.	max $\lvert z\rvert$ (thr. 4)
analytic (reference)	0.192 ± 0.009	1.011 ± 0.009	0.167 ± 0.013	0.857	0.00044	—
analytic (binned) (ref.)	0.185 ± 0.008	1.003 ± 0.008	0.344 ± 0.013	0.854	0.00044	—
innovation form	0.186 ± 0.007	1.007 ± 0.011	0.173 ± 0.012	0.854	0.00047	0.54
substitution form	0.182 ± 0.006	1.004 ± 0.011	0.156 ± 0.010	0.853	0.00027	0.96
estimated	0.186 ± 0.005	1.013 ± 0.009	0.371 ± 0.014	0.854	0.00032	1.39
regularized (shipped)	0.184 ± 0.006	0.988 ± 0.008	0.349 ± 0.013	0.851	0.00026	1.28
MARG (must fail)	0.250 ± 0.010	$\mathbf{1.982 \pm 0.015}$	1.038 ± 0.019	*0.951*	0.0094	—

E.3.2 RESULT: THE GAUSSIAN ANCHOR IS VALIDATED, WITH MEASURED POWER

Table 7 gives the verdict and Figure 5 shows it against the floors. All four implemented samplers are indistinguishable from their reference on conditional mean, covariance trace, covariance shape and coverage, with a maximum $|z|$ of 1.39 against the 4σ threshold, so the pass is not marginal. The two exact forms also agree with each other, as the theory requires and as no estimation stands between them ($z = 0.18$); this is the check that would catch an implementation bug rather than a statistical one.

The test's power is demonstrated rather than assumed. MARG is a $+100\%$ deviation in conditional spread and is detected as such, and its measured overdispersion of 1.982 ± 0.015 matches the 2.000 predicted by Lemma 2 at $z = 1.20$. That pairing is what converts the lemma from an analytical remark into an empirical claim: "MARG is over-dispersed" would be unfalsifiable on its own, whereas the lemma predicts a number and the number is observed. The $\mathrm{tr}(S_{\delta\delta})/\mathrm{tr}(S_r) = 2.000$ of this control is what makes the prediction exactly 2. The same failure shows up in every other column: MARG's coverage of 0.951 *exceeds* the 0.939 cap a calibrated 32-member ensemble can attain—itself an over-dispersion signature—and its rank deviation is $22\times$ the reference. The energy distance to the analytic law separates the arms the same way (4.48 for MARG against 0.47 for the reference), but because it is an unbiased estimator of a non-negative divergence it scatters through zero on the exact arms (the substitution form reads -0.90), so we do not read its sign as meaningful. Energy score, dominated by the target's irreducible spread, separates the arms only weakly (174.5 reference vs 180.8 for MARG, $+3.6\%$) and is reported for comparability with the later benchmarks rather than as a discriminating test here. The fit itself is clean: 23 supported shells, no negative source eigenvalues, no clipped trace, and a shell-averaged filter whose imaginary part is 3×10^{-8} of its real part—the residual that would be large were the pair not jointly stationary.

We repeated the whole test on sampling seeds 1 and 2: the verdict holds on all three, with every arm passing, the exact forms agreeing ($z = 0.18, 0.28, 1.59$) and MARG matching the lemma (1.982, 1.977, 1.994). One caveat follows from that repetition and we state it rather than reporting the single seed. For the covariance *shape* column the seed-to-seed spread reaches $4.7\times$ the within-seed standard error over conditions (substitution form: 0.156, 0.128, 0.175), so the z values in that column are computed against a standard error that understates the true variability, and the largest $|z|$ anywhere in the three runs is 2.66 rather than the 1.39 visible at seed 0. The verdict is unchanged—2.66 is still well inside 4σ—but a future tightening of this test should widen the Frobenius column's error bar to include the sampling axis. The mean, trace and coverage columns do not have this problem: their seed spread is within $1.7\times$ the condition standard error.

Consequently the NCS Gaussian anchor is validated on this control and results built on this implementation are interpreted. What part 2 then measures is the price of the two approximations this test holds fixed: finite data, and the radial representation whose 0.344 cost is already visible in the binned reference row above.

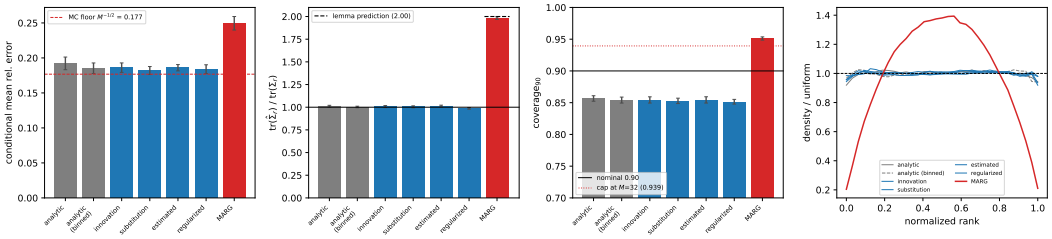


Figure 5: **Oracle Gaussian test.** The same run as Table 7; grey bars are the two *references* and not competitors, and error bars are standard errors over the 26 test conditions. Every panel is read against a line that is not zero. *Left:* an exactly correct sampler still reads the Monte Carlo floor $M^{-1/2} = 0.177$ (dashed) at $M = 32$, and the four implemented arms sit on it. *Centre:* the conditional spread relative to the truth, where 1 is correct; the dashed segment over the last bar is the 2.00 that Lemma 2 predicts for MARG *alone*, and the measured 1.982 ± 0.015 is what converts that lemma from a remark into a tested claim. *Right of centre:* coverage cannot reach the nominal 0.90 with a finite ensemble—the cap at $M = 32$ is 0.939 (dotted)—so the admissible arms' 0.85 is calibrated, whereas MARG *exceeds* the cap, which is itself an over-dispersion signature. *Right:* rank histograms; flat is calibrated, and MARG's inverted-U is the classic over-dispersion shape, with $22\times$ the reference's deviation from uniform. The gap between *analytic* and *analytic (binned)* is the cost of the radial representation itself (0.344 in relative Frobenius error), which is why the estimated arms are scored against the latter (§E.3.1).

Finite-data and regularization study. We vary, one axis at a time from a common reference setting:

1. the number of training conditions;
2. the number of realizations per condition;
3. Tikhonov regularization λ;
4. spectral smoothing bandwidth; and
5. covariance rank truncation.

The source is constructed from the full regularized innovation covariance of Eq. equation 22, not from a damped Schur-complement shortcut. The ladder's purpose is to price that decision: the λ axis is where the shortcut and Eq. equation 22 must visibly disagree, and if they never do, the modelling choice is untested here rather than vindicated.

The control is the stationary Gaussian problem at $N = 128$, refinement factor 4, $K = 8$ members per condition, 256 conditions split contiguously into $204/26/26$ train/val/test, evaluated with $M = 32$ draws per test condition in double precision. The reference rung is all 204 training conditions, all 8 members, $\lambda_{\mathrm{rel}} = 10^{-6}$, no smoothing, no rank truncation.

E.3.3 The error bars are over conditions, not over seeds

A seed-to-seed error bar would have been the wrong instrument here, and measurably so. Every spectral statistic in this study is a deterministic function of the estimated block spectrum: re-running the reference rung under sampling seeds 0, 1 and 2 leaves the relative trace error, the relative Frobenius error, the filter gain, the held-out variance reduction and the Schur-shortcut error *bit-identical*, and moves only the coverage, and only in its third decimal (0.8515–0.8574). A seed interval would therefore have been zero width on five of the six reported columns while looking like a legitimate uncertainty statement.

The uncertainty that exists is over *which conditions were drawn*. We report percentile bootstrap intervals (200 replicates, 95%) that resample the train and test conditions with replacement and **refit the estimator on every replicate**. Refitting is what makes the interval describe the estimator actually used: holding the fit fixed and resampling only the test block would price the evaluation noise and omit the estimation error, which on the shortest rungs is the dominant term. Independent

Table 8: **Finite-data and regularization cost on the stationary Gaussian control.** One axis varied per block from the reference rung (boldface; repeated in each block as that block's own control). Brackets are 95% percentile bootstrap intervals over resampled train and test conditions with the estimator refit per replicate (200 replicates, all usable). Coverage is nominal 90% but is capped at 0.939 by the $M = 32$ ensemble, and the exact analytic sampler of the oracle test itself reads 0.857 on this split, so ≈ 0.85 is the sampler's own level and not an estimation defect. The Schur column is the relative error of the damped-shortcut residual spectrum against Eq. equation 22; it is reported only on the λ block, since every other block sits at the reference $\lambda_{\mathrm{rel}} = 10^{-6}$ where it is 9.9×10^{-7}.

Rung	rel. trace err.	rel. Frobenius err.	filter gain	cov_{90}	Schur err.
Training conditions n (truth: gain $= 1$, trace err. $= 0$)					
8	0.0795 [0.009, 0.170]	0.0599 [0.031, 0.198]	0.9976 [0.931, 1.048]	0.842	—
16	0.0374 [0.002, 0.091]	0.0296 [0.022, 0.104]	1.0042 [0.969, 1.037]	0.844	—
32	0.0260 [0.002, 0.065]	0.0203 [0.013, 0.074]	1.0115 [0.986, 1.040]	0.850	—
64	0.0152 [0.001, 0.039]	0.0142 [0.008, 0.046]	0.9995 [0.980, 1.016]	0.850	—
128	0.0141 [0.002, 0.031]	0.0144 [0.008, 0.041]	1.0013 [0.988, 1.015]	0.852	—
204	**0.0019** [0.0004, 0.017]	**0.0117** [0.004, 0.033]	**0.9982** [0.986, 1.010]	**0.851**	—
Members per condition K ($K = 2$ is the identifiability floor)					
2	0.0149 [0.001, 0.043]	0.0320 [0.005, 0.077]	0.9972 [0.980, 1.012]	0.851	—
3	0.0006 [0.0003, 0.026]	0.0039 [0.003, 0.043]	0.9922 [0.977, 1.006]	0.852	—
4	0.0024 [0.0005, 0.023]	0.0099 [0.004, 0.042]	0.9986 [0.981, 1.013]	0.854	—
6	0.0016 [0.0002, 0.018]	0.0066 [0.003, 0.034]	0.9981 [0.986, 1.010]	0.855	—
8	**0.0019** [0.0004, 0.017]	**0.0117** [0.004, 0.033]	**0.9982** [0.986, 1.010]	**0.856**	—
Tikhonov λ_{rel} (relative to each shell's mean eigenvalue)					
0	0.0019 [0.0004, 0.017]	0.0117 [0.004, 0.033]	0.9982 [0.986, 1.010]	0.851	0 $[0, 2{\cdot}10^{-16}]$
10^{-8}	0.0019 [0.0004, 0.017]	0.0117 [0.004, 0.033]	0.9982 [0.986, 1.010]	0.859	$9.9{\cdot}10^{-9}$
$\mathbf{10^{-6}}$	**0.0019** [0.0004, 0.017]	**0.0117** [0.004, 0.033]	**0.9982** [0.986, 1.010]	**0.853**	$\mathbf{9.9{\cdot}10^{-7}}$
10^{-4}	0.0019 [0.0004, 0.017]	0.0117 [0.004, 0.033]	0.9981 [0.986, 1.010]	0.850	$9.9{\cdot}10^{-5}$
10^{-2}	0.0020 [0.0004, 0.017]	0.0117 [0.004, 0.033]	0.9884 [0.976, 1.000]	0.850	$9.72{\cdot}10^{-3}$ $[9.35, 10.06]{\cdot}10^{-3}$
10^{-1}	0.0101 [0.001, 0.024]	0.0186 [0.003, 0.041]	0.9075 [0.896, 0.918]	0.855	$\mathbf{8.13{\cdot}10^{-2}}$ $[7.82, 8.41]{\cdot}10^{-2}$
1	0.2503 [0.229, 0.266]	0.2587 [0.227, 0.285]	0.4991 [0.493, 0.505]	0.850	$1.99{\cdot}10^{-1}$ $[1.93, 2.04]{\cdot}10^{-1}$
Shell smoothing half-width (of 91 radial shells)					
0	**0.0019** [0.0004, 0.017]	**0.0117** [0.004, 0.033]	**0.9982** [0.986, 1.010]	**0.853**	—
1	0.0227 [0.006, 0.037]	0.6781 [0.6777, 0.6789]	0.9981 [0.985, 1.010]	0.857	—
2	0.0958 [0.076, 0.112]	0.7404 [0.7396, 0.7416]	0.9979 [0.985, 1.010]	0.869	—
4	0.4071 [0.380, 0.430]	0.8463 [0.8457, 0.8469]	0.9977 [0.983, 1.010]	0.908	—
8	1.0340 [0.992, 1.070]	0.9438 [0.9417, 0.9457]	0.9975 [0.982, 1.011]	0.953	—
16	2.4020 [2.329, 2.464]	1.0180 [1.015, 1.021]	0.9973 [0.981, 1.012]	0.985	—
Source rank (retained shells; 23 of 91 are supported by the coarse box)					
2	0.2564 [0.243, 0.273]	0.1168 [0.1163, 0.1209]	0.9982 [0.986, 1.010]	0.796	—
4	0.1215 [0.108, 0.138]	0.0385 [0.037, 0.050]	0.9982 [0.986, 1.010]	0.827	—
8	0.0408 [0.027, 0.057]	0.0138 [0.008, 0.034]	0.9982 [0.986, 1.010]	0.849	—
16	0.0111 [0.001, 0.027]	0.0117 [0.004, 0.033]	0.9982 [0.986, 1.010]	0.849	—
32	0.0011 [0.0002, 0.018]	0.0117 [0.004, 0.033]	0.9982 [0.986, 1.010]	0.850	—
all	**0.0019** [0.0004, 0.017]	**0.0117** [0.004, 0.033]	**0.9982** [0.986, 1.010]	**0.852**	—

resampling is legitimate on this control because its conditions are independent draws from a known law; it would not be on the observational benchmark, whose records are consecutive days, and we do not use it there. All 200 replicates were usable on all 31 rungs.

Two floors bound what any rung can achieve, and neither is a property of the estimator. With $M = 32$ members the Monte Carlo floor on a conditional-mean error is $32^{-1/2} = 0.177$, so mean errors below that are not resolvable at this ensemble size. Separately, the Frobenius column is scored against the *shell-binned* truth, which is itself 0.334 in relative Frobenius distance from the true per-mode spectrum; that column measures agreement with the best a radially binned estimator could do and must not be read as distance to the truth.

One reading caveat applies throughout Table 8. Both error columns are magnitudes of signed quantities whose true value is zero, so their bootstrap distributions are folded at the origin and the point estimate sits near the interval's lower end by construction—the relative trace error at the reference rung is at 9.6% of its interval's width, while the signed trace ratio, which is the same measurement without the fold, sits at 53.9% of its own. Centering should be read off the signed gain column, not off the folded error columns.

E.3.4 RESULT: THE SHORTCUT IS MEASURABLY WRONG, AND SMOOTHING IS THE EXPENSIVE KNOB

Table 8 prices each axis and Figure 6 plots the ladders. Three findings are load-bearing for the rest of the paper.

The Schur shortcut is not a usable stand-in, and this is now measured rather than asserted. The two forms agree exactly at $\lambda = 0$—the shortcut error is 0 with interval $[0, 2 \cdot 10^{-16}]$, which is the correct behaviour, since the shortcut *is* the Schur complement there—and the discrepancy then grows essentially linearly in λ, tracking 0.99λ up to $\lambda_{\mathrm{rel}} = 10^{-4}$ before saturating as the damping starts to dominate (0.81λ at 10^{-1}, 0.20λ at 1). Requiring the whole bootstrap interval to clear a 1% tolerance, the shortcut is established wrong from $\lambda_{\mathrm{rel}} = 10^{-1}$ onward, reaching a relative error of 0.199 $[0.193, 0.204]$ at $\lambda_{\mathrm{rel}} = 1$. The interval does not overturn this verdict—the point estimate places the crossing at the same rung—but it does change the status of one rung: at $\lambda_{\mathrm{rel}} = 10^{-2}$ the shortcut error is $9.72 \cdot 10^{-3}$ with interval $[9.35, 10.06] \cdot 10^{-3}$, straddling the tolerance, so that rung is indeterminate rather than passing. Constructing the source from Eq. equation 22 is therefore a real requirement at any regularization strong enough to matter, and free at the regularization we use.

Regularization is safe up to $\lambda_{\mathrm{rel}} = 10^{-2}$ and not beyond, on the mean rather than the spread. Requiring both a trace error within 5% and a gain within 5% of unity, the largest admissible value is $\lambda_{\mathrm{rel}} = 10^{-2}$ under the point estimates and the same value under the CI-strict reading. What disqualifies 10^{-1} is not the spectrum—its trace error is still 0.0101—but the gain, which has fallen to 0.9075 $[0.897, 0.918]$: Tikhonov damping shrinks the conditional *mean* before it distorts the innovation spectrum. At $\lambda_{\mathrm{rel}} = 1$ the gain is 0.4991 $[0.493, 0.505]$, i.e. the filter has halved the mean prediction, and the held-out variance reduction falls from 0.493 to 0.366 against an analytic 0.5. This is the failure mode to watch on real data, where the gain is observable but the truth is not.

Shell smoothing is far more expensive than it looks, and coverage hides it. A single-shell half-width already inflates the relative Frobenius error from 0.0117 to 0.6781 $[0.6777, 0.6789]$—twice the 0.334 that radial binning itself costs, so the damage exceeds the discretization it is applied on top of—while the trace error rises only to 0.0227. The intervals here are extremely narrow, which is the signature of a *bias*: bootstrap replicates disagree about the sampling noise, not about the systematic tilt smoothing imposes on a red spectrum. Coverage moves the wrong way over the same block, rising monotonically to 0.985 at half-width 16, because a flattened source is over-dispersed at low k and over-wide intervals cover more. Any pipeline tuned on coverage alone would have selected the worst rung on the ladder. We therefore use no smoothing, and we report the spectral error alongside coverage everywhere rather than coverage alone.

The remaining two axes are cheap. Members per condition matter little above the $K = 2$ identifiability floor: $K = 2$ already reaches a trace error of 0.0149 $[0.001, 0.043]$ and a gain of 0.9972, and $K = 3$ is statistically indistinguishable from the full $K = 8$. Rank truncation is harmless once the retained count exceeds the 23 shells the coarse box actually supports; below that its cost appears in coverage (0.796 at rank 2) rather than in the gain, for the reason given in the table footnote. Condition count behaves as expected—the interval width on the trace error contracts monotonically from 0.161 at $n = 8$ to 0.016 at the full $n = 204$, a factor of 10 for a factor of 25 in n—and we note that it is *still* contracting at the largest count available, so this ladder bounds the finite-data cost from above and does not show that the shipped 256 conditions saturate it.

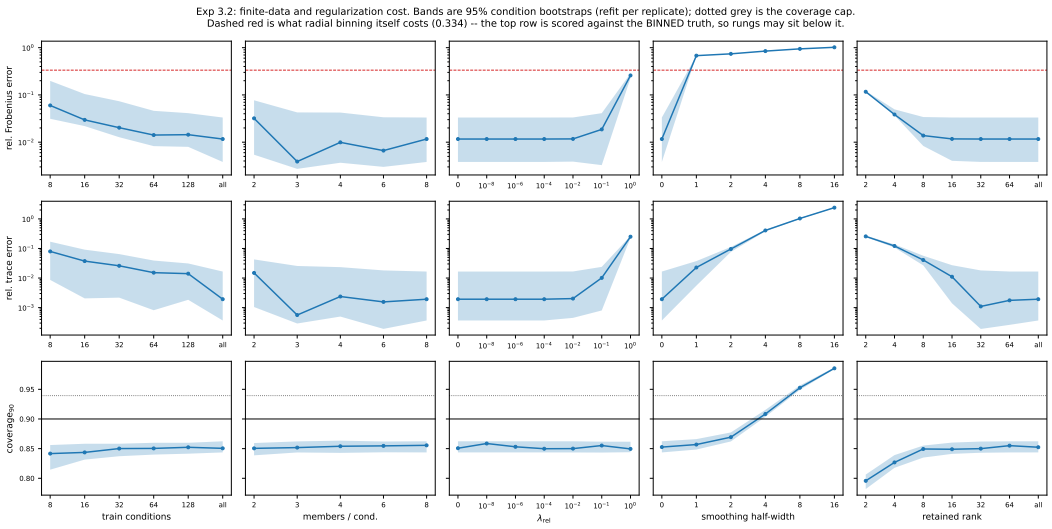


Figure 6: **Finite-data and regularization ladders** (the rungs of Table 8). Bands are 95% condition bootstraps that refit per replicate, so they include estimation error and not merely evaluation noise (§E.3.3); rungs are equally spaced because the ladders are discrete and λ_{rel} includes 0. Dashed red on the top row is the 0.334 that radial binning itself costs against the per-mode spectrum; that column is scored against the *binned* truth, so rungs legitimately fall below the line, and it marks the scale at which a degradation is large rather than a bound. The *smooth* column carries the finding that generalizes: one shell of smoothing pushes the covariance *shape* error past that line and eventually to ~ 1, while $\mathrm{coverage}_{90}$ climbs monotonically *through* the nominal 0.90 and past its own finite-ensemble cap (dotted). A pipeline tuned on coverage alone would select the worst rung on the ladder; the bands there are also narrow, which marks the damage as bias rather than variance. Note also that the λ axis degrades the conditional *mean* well before this spectrum (gain 0.91 at $\lambda_{\mathrm{rel}} = 10^{-1}$, where the trace error is still 0.010), so neither of these two rows would have flagged it—the reason Table 8 reports gain as its own column.

Stationarity stress test. On the controlled nonstationary Gaussian problem and GARD–LENS, we compare:

1. a global stationary spectral filter;
2. season-specific filters;
3. region-specific or locally stationary filters;
4. a non-spectral low-rank covariance estimator where feasible; and
5. the learned NCS correction.

We measure held-out innovation coherence and local conditional variance after each fit. This quantifies the burden placed on the neural transport by the stationarity approximation. All five arms estimate the *same* quantity—the per-condition mean of δ—so the comparison is between estimators of one target rather than between objectives.

E.3.5 A BURDEN IS ONLY READABLE AGAINST A CEILING

A statement of the form "the learned correction recovers $Y\%$ of the error the stationary filter leaves" is uninterpretable on its own, because it does not say how much of that error is recoverable *at all*. We therefore add a sixth, non-competing arm: the controlled problem is built so that its conditional mean $\mathbb{E}[\delta \mid c]$ is available in closed form, and evaluating it supplies the exact ceiling on the same split and the same metric. Every burden below is reported as a fraction of that ceiling.

This is not a refinement of the presentation; it changes the conclusion. Our first nonstationary control modulated the signal by a smooth per-pixel gain $a(x)$, which makes the true conditional

mean $a(x)\,\mathcal{P}c$—a per-pixel affine function of the stationary filter's own output. A per-pixel affine correction is then the *matched* estimator rather than a stand-in for a transport, and it duly reaches $1.01\times$ the ceiling (Table 9, second block). Its "54% recoverable" is arithmetically correct and argues *against* needing a learned transport, since a pointwise rescaling closes the entire achievable gap.

The discriminating control instead makes the position-dependence *spectral*:

$$m(x) = 2\big[(1 - u(x))\,P_{<k_s}s + u(x)\,P_{\geq k_s}s\big], \tag{38}$$

a position-dependent mixture of two disjoint frequency bands of the same field ($k_s = 4$, splitting the signal power $82/18$). Because the two projections have exactly disjoint support they are exactly uncorrelated, so no per-pixel gain applied to any fixed mixture can reproduce a spatially varying one; yet Eq. equation 38 is still linear in s, so the conditional mean remains in closed form and the ceiling still exists. Both projections are radially symmetric, which matters for the reason given next.

The confound: isotropy is not stationarity. The global arm is an isotropic, radially binned Wiener filter, so it can be defeated for two unrelated reasons—genuine nonstationarity, or mere angular variation within a radial shell while the field is perfectly stationary. We therefore re-run every control with the modulation *frozen* at its mean, which yields a stationary field built by the identical code path; the ceiling must then collapse to zero. Eq. equation 38 reads -0.4% frozen, so its gap is attributable to nonstationarity. A position-dependent *shift* was rejected on this test: it produced a large 60–75% ceiling that survived freezing, because the phase $e^{ik\cdot d}$ of a diagonal shift varies within every radial shell. Its tell was that the locally stationary arm got monotonically *worse* with more tiles ($-0.07\times$ at 4, $-1.30\times$ at 8), tiling adding spatial rather than angular resolution. Any claim that a stationary filter "cannot represent" an operator must be checked this way or it measures the estimator's isotropy instead.

E.3.6 The learned arm must be able to represent nonstationarity

The learned arm predicts a residual on top of the global filter's output with a spectral operator, and this creates the same hazard as the $M = 12$ truncation of §E.1: a Fourier neural operator is built from spectral convolutions and 1×1 convolutions, both of which commute with translation, so the composed map is translation equivariant—measured at 5×10^{-8} relative at this arm's configuration. A translation-equivariant map *is* a stationary operator. Fed only the filter output it could not represent position dependence of any kind, would score near zero on any nonstationary control for purely architectural reasons, and would manufacture support for the stationarity burden it is meant to measure. We therefore append two normalized coordinate ramps, which break the equivariance by four orders of magnitude (to 1.5×10^{-3} here; the absolute value is a property of the untrained initialization and moves with seed and width, so we assert the *separation* rather than a magnitude), and set the mode truncation to the coarse Nyquist rather than the conventional 12. As in §E.1, a negative result is only evidence if the estimator could have produced a positive one.

The learned arm also *abstains*: if its validation score does not improve on the global filter it returns that filter unchanged. Validation is a contiguous tail of the training block, never the test block, so this is a fit-time decision. On the stationary null it abstains and ties the filter exactly (Table 9, first block), which is the honest outcome there—the filter is already the exact conditional mean, so there is nothing above it to learn and an arm that "improved" on it would be fitting noise.

Table 9: Stationarity burden on held-out blocks. The metric is the spread of the per-tile ratio of achieved to true local conditional standard deviation (8×8 tiles), reported as a relative RMSE; a spatially uniform arm can match the domain-average variance exactly while being wrong tile by tile, so the spread and not the mean is the quantity a stationary source cannot fix. *Oracle* is the exact $\mathbb{E}[\delta \mid c]$ from the control's own closed form: it is the denominator, never a competitor. $\times$ceil is each arm's share of the oracle's own gap over the global filter. The first block is the stationary NULL, where every arm should tie: its ceiling is -0.1%, so fractions of it are undefined rather than large, and we report none. The second block is degenerate *by construction* for a pointwise correction and is retained only for the stratified arms; the third is the discriminating control of Eq. equation 38. Burdens are given as the range over 6 slid test blocks, since one block is not an error bar.

Control	Arm	rel. RMSE	$\times$ceil	coh./null q_{95}	burden (6 blocks)	best arm
Gaussian stationary (NULL; $n = 256$, $n_{\text{test}} = 26$; ceiling -0.1%, no gap)						
	global	0.0379	—	0.85		
	seasonal	0.0382	—	1.09		
	locally stat.	0.0366	—	14.9		
	low-rank	*0.0954*	—	13.7		
	pointwise ref.	0.0378	—	0.87		
	NCS (learned)	**0.0379**	—	0.85	0.0–7.4%	varies
	oracle	0.0380	—	3.16		
Gaussian nonstationary, per-pixel gain (degenerate; $n = 128$, $n_{\text{test}} = 13$; ceiling $\mathbf{53.1}\%$)						
	global	0.0905	—	0.86		
	seasonal	0.0790	0.24	0.87		
	locally stat.	0.0759	0.31	5.81		
	low-rank	*0.1294*	—	5.33		
	pointwise ref.	0.0420	*1.01*	0.78		
	NCS (learned)	0.0416	*1.02*	0.93	39.9–55.2%	varies
	oracle	0.0425	1.00	0.77		
Gaussian nonstationary, band mixture (Eq. equation 38; $n = 256$, $n_{\text{test}} = 26$; ceiling $\mathbf{78.1}\%$)						
	global	0.1740	—	0.78		
	seasonal	0.1944	-0.15	0.92		
	locally stat.	0.1012	0.54	11.0		
	low-rank	*0.1055*	—	11.5		
	pointwise ref.	0.1046	0.51	0.77		
	NCS (learned)	**0.0385**	**1.00**	1.24	**75.1–80.9%**	stable
	oracle	0.0381	1.00	0.83		
GARD–LENS (observational; $n = 1024$, $n_{\text{test}} = 103$; **no closed form, hence no ceiling**)						
	global	3.1027	n/a	6.05		
	seasonal	3.0950	n/a	6.04		
	locally stat.	3.4348	n/a	6.10		
	low-rank	*3.0005*	n/a	2.94		
	pointwise ref.	1.4787	n/a	26.5		
	NCS (learned)	**1.2041**	n/a	2.48	60.9–67.6%	stable

E.3.7 Result: the burden is real, and it is the learned arm that carries it

Table 9 gives the four blocks and Figure 7 summarizes them. On the stationary null every arm ties the global filter to within the Monte Carlo margin except low-rank, which is disqualified. On the discriminating control the global stationary filter leaves a local-variance error of 0.174 against an achievable 0.038: a 78.1% ceiling. The classical repairs recover about half of it—locally stationary $0.54\times$, season-specific $-0.15\times$, i.e. actively harmful, since a spectral band mixture has no seasonal structure for it to exploit—while the learned NCS correction reaches 0.0385, or $1.00\times$ the ceiling, with a burden of 75.1–80.9% across six slid blocks and the same winning arm in all six.

We read that $1.00\times$ as closing the gap to the exact conditional mean, and deliberately not as a bare percentage. The distinction from the degenerate case is the pointwise reference arm on the same split: it cannot represent Eq. equation 38 and reaches only $0.51\times$, whereas on the per-pixel-gain control it reaches $1.01\times$ and the learned arm's $1.02\times$ carries no information. A ratio near unity is evidence about the *arm* only when an arm that provably cannot solve the problem scores far below it.

On GARD–LENS the ordering survives contact with real data: the learned arm reads 1.204 against the global filter's 3.103 and the pointwise reference's 1.479, a 60.9–67.6% improvement that is stable across blocks. There is no closed-form conditional mean here, so this is not a fraction of anything achievable and we do not present it as one—it bounds the burden from below only.

The burden is therefore not an artifact of the stationary filter being a weak baseline: on a control where the optimum is known, the stationary filter forgoes 78% of the available local-variance accuracy, no classical stratification of it recovers more than half, and a learned correction recovers essentially all. That is the quantitative form of the claim that the stationarity approximation places a real burden on the neural transport.

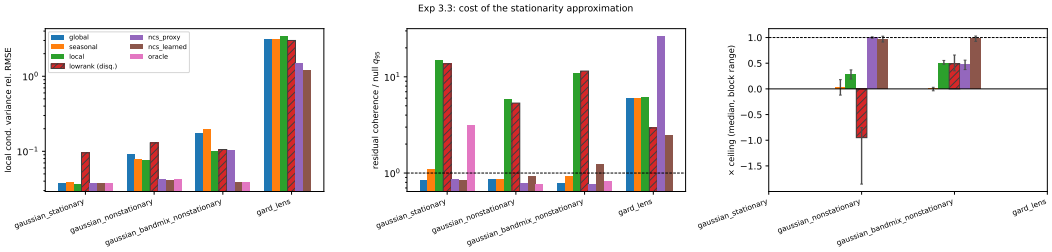


Figure 7: **Cost of the stationarity approximation** (Table 9). Hatched bars are DISQUALIFIED: the low-rank arm fails the stationary-null gate, and it is shown rather than dropped so the gate is visible. *Left:* the ranked quantity, per-tile local conditional variance error. *Centre:* residual coherence against a permutation null, a diagnostic only—it is a domain-averaged and therefore stationary statistic, and the locally stationary arm's large values are its tile seams, not conditional structure. *Right:* each arm's share of the *oracle's own* gap over the global filter, as the median and range over six slid test blocks. Two systems are deliberately absent from that panel: the stationary null, whose ceiling is -0.1% so that fractions of it are undefined rather than large, and GARD–LENS, which has no closed-form conditional mean and therefore no ceiling to take a fraction of. On the discriminating band-mixture control the learned arm reaches the ceiling while the pointwise reference reaches only about half of it—that separation, and not the height of the learned bar, is what shows the control is not degenerate (§E.3.5).

Falsification criterion. If the oracle or sufficiently sampled stationary implementation fails to recover the analytic Gaussian conditional law, the NCS Gaussian anchor is not validated and results using that implementation are not interpreted. This is adjudicated at 4σ against a reference row of the same ensemble size rather than against an absolute tolerance, and the test is required to report the smallest deviation it could have detected, so that a pass cannot be earned by a small Monte Carlo budget; the MARG arm is the standing check that this power is real (§E.3.2). For the finite-data study the criterion is that a knob is reported as safe only where the whole bootstrap interval, not merely the point estimate, stays inside the stated tolerance; and if the two innovation forms had agreed across the entire λ ladder we would have recorded the distinction as untested rather than as support for Eq. equation 22. For the stationarity stress test the criterion is separate and has two parts, both of which a burden must survive to be reported: the control's ceiling must not be reachable by the pointwise reference arm (or the control is degenerate and measures nothing about transport), and the ceiling must vanish when the modulation is frozen at its mean (or the gap is the global filter's isotropy rather than nonstationarity). Our first nonstationary control fails the first part and a position-dependent shift fails the second; both are reported above rather than replaced silently.

F FULL RESULTS AND REMAINING STUDIES

F.1 FULL BASELINE LADDER

Table 10 is the complete sixteen-arm ladder behind Table 1, with the variogram score, raw spread/skill, and P99 columns. Arm names: NCS-Drift+ES is the NCS of the main text; NCS-ES is NCS (no pretrain); NCS-twES is NCS+tw.

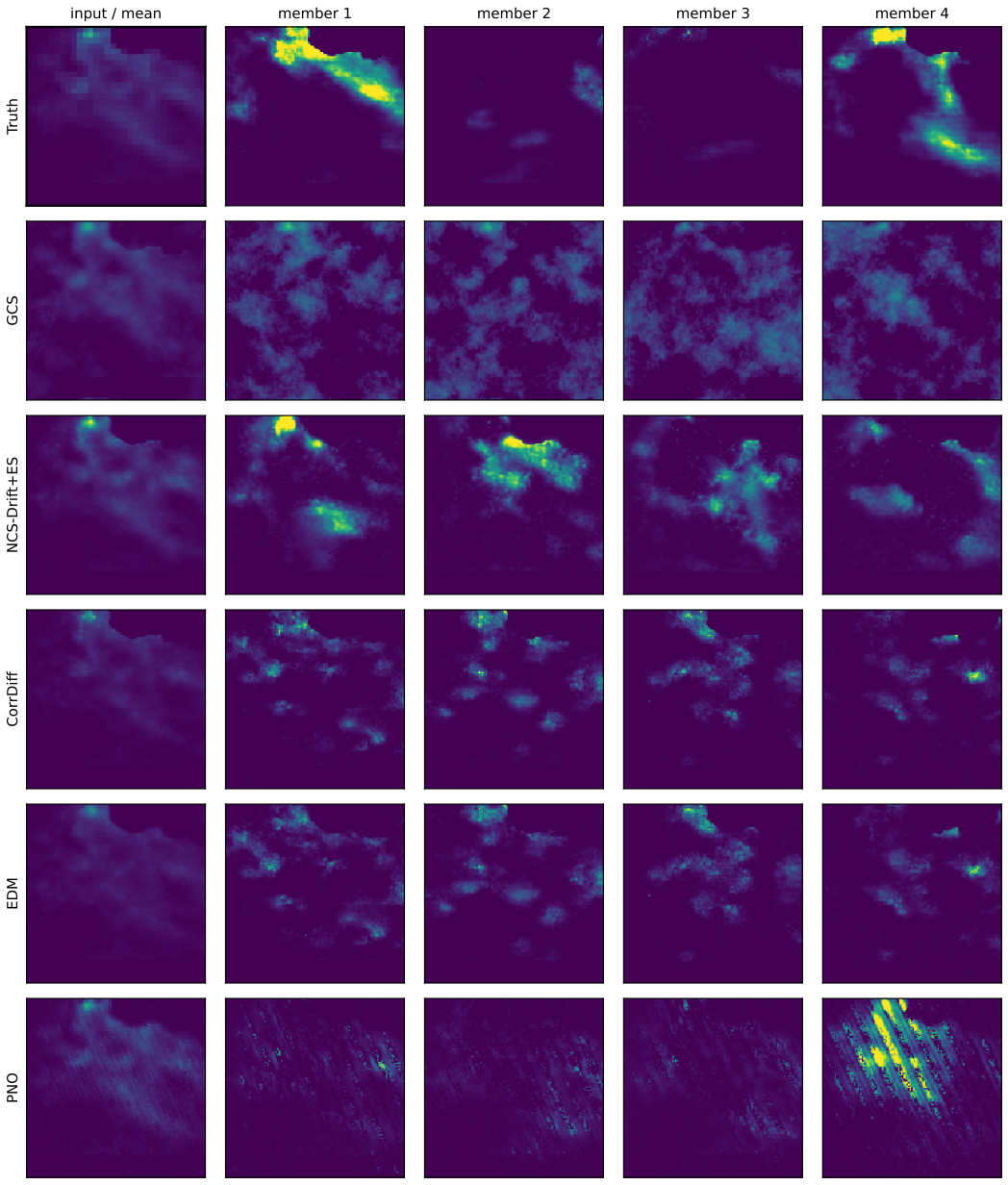


Figure 8: GARD-LENS precipitation on a held-out day. The 32^2 coarse field (top left) is refined to 128^2. Top row: data members. Following rows: members from the classical floor (GCS), NCS, CorrDiff, EDM, and PNO, each preceded by its ensemble mean; one shared color scale from zero to the data's 99.5th percentile. Coarse field and ensemble means carry only the large-scale amount; members restore the localized wet cells, and PNO's show the streak artifacts of a one-pass operator on a heavy-tailed target.

Table 10: Full held-out results, sixteen arms on three benchmarks. Raw and Fortin-corrected spread/skill are both shown. The NCS-twES tail weight is set per system by a gradient-norm match at initialization with $\alpha = 0.3$ (0.1 on Allen–Cahn, where the stability gate triggers). P99 is the 99th percentile of the standardized |increment| pool; the data's own P99 is 3.00 (Allen–Cahn), 2.90 (Navier–Stokes), and 3.87 (GARD-LENS), against 2.58 for a Gaussian field. "–" is $0/0$ for the spread-collapsed deterministic arms. Navier–Stokes ± are condition bootstraps (frozen split).

Benchmark	Method	RMSE	ES	CRPS	Variogram	Raw σ/s	Corrected σ/s	logPSD	$\kappa/\kappa_{\text{data}}$	P99	NFE
Allen–Cahn											
	LIN	0.288	46.4	0.623	0.038	0.000	0.000	2.670	–	–	0
	NN-L_2	0.326 ± 0.004	47.4 ± 0.1	0.625 ± 0.001	0.036 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.312 ± 0.034	–	–	1
	MARG	0.311 ± 0.001	32.8 ± 0.0	0.420 ± 0.001	0.033 ± 0.001	0.971 ± 0.009	0.986 ± 0.009	0.063 ± 0.001	0.70 ± 0.00	2.58 ± 0.00	0
	GCS	0.305 ± 0.001	33.0 ± 0.1	0.425 ± 0.001	0.021 ± 0.000	0.837 ± 0.007	0.850 ± 0.008	0.026 ± 0.001	0.70 ± 0.00	2.58 ± 0.00	0
	HET-GCS	0.304 ± 0.001	33.1 ± 0.1	0.426 ± 0.001	0.020 ± 0.000	0.819 ± 0.008	0.832 ± 0.008	0.035 ± 0.001	0.70 ± 0.00	2.58 ± 0.00	0
	TGCS	0.310 ± 0.001	32.8 ± 0.0	0.418 ± 0.001	0.028 ± 0.000	0.940 ± 0.007	0.954 ± 0.007	0.319 ± 0.001	0.97 ± 0.00	3.01 ± 0.00	0
	Bilinear	0.506 ± 0.000	52.6 ± 0.0	0.576 ± 0.000	0.075 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.193 ± 0.000	–	–	0
	FNO	0.358 ± 0.005	48.2 ± 0.1	0.628 ± 0.001	0.038 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	0.399 ± 0.009	–	–	1
	PNO	0.400 ± 0.000	37.9 ± 0.0	0.491 ± 0.004	0.075 ± 0.002	0.488 ± 0.003	0.495 ± 0.003	1.076 ± 0.090	0.98 ± 0.10	2.92 ± 0.10	1
	Consistency	0.310 ± 0.001	32.6 ± 0.0	0.401 ± 0.001	0.139 ± 0.003	0.997 ± 0.006	1.013 ± 0.006	1.634 ± 0.017	0.75 ± 0.01	2.64 ± 0.01	2
	PDE-Refiner	0.280 ± 0.003	44.7 ± 0.1	0.591 ± 0.001	0.032 ± 0.001	0.047 ± 0.001	0.048 ± 0.001	0.552 ± 0.033	0.94 ± 0.02	2.77 ± 0.01	4
	Diffusion (EDM)	0.316 ± 0.022	33.1 ± 0.6	0.388 ± 0.008	0.038 ± 0.001	0.853 ± 0.048	0.866 ± 0.049	0.775 ± 0.024	5.78 ± 0.11	3.98 ± 0.21	35
	Diffusion (CorrDiff)	0.306 ± 0.008	32.8 ± 0.1	0.384 ± 0.003	0.039 ± 0.001	0.884 ± 0.016	0.897 ± 0.016	0.793 ± 0.012	5.35 ± 0.06	4.00 ± 0.07	35
	NCS-ES	0.306 ± 0.003	32.8 ± 0.1	0.417 ± 0.002	0.168 ± 0.005	0.908 ± 0.003	0.922 ± 0.003	1.500 ± 0.008	0.87 ± 0.01	2.70 ± 0.01	2
	NCS-Drift+ES	0.304 ± 0.011	32.8 ± 0.2	0.409 ± 0.002	0.025 ± 0.001	0.894 ± 0.002	0.908 ± 0.002	0.134 ± 0.001	0.874 ± 0.00	2.64 ± 0.01	2
	NCS-twES	0.304 ± 0.011	32.8 ± 0.2	0.409 ± 0.003	0.025 ± 0.001	0.900 ± 0.004	0.914 ± 0.004	0.124 ± 0.006	0.75 ± 0.00	2.65 ± 0.01	2
Navier–Stokes											
	LIN	0.826 ± 0.010	445.1 ± 1.6	1.374 ± 0.005	2.005 ± 0.015	0.000 ± 0.000	0.000 ± 0.000	3.590 ± 0.002	–	–	0
	NN-L_2	0.929 ± 0.004	458.2 ± 0.8	1.417 ± 0.002	1.601 ± 0.007	0.000 ± 0.000	0.000 ± 0.000	1.226 ± 0.006	–	–	1
	MARG	0.882 ± 0.005	314.8 ± 0.7	0.976 ± 0.002	0.369 ± 0.005	0.986 ± 0.002	1.001 ± 0.002	0.033 ± 0.001	0.75 ± 0.00	2.67 ± 0.00	0
	GCS	0.882 ± 0.005	314.8 ± 0.7	0.976 ± 0.002	0.369 ± 0.005	0.983 ± 0.002	0.998 ± 0.002	0.033 ± 0.001	0.75 ± 0.00	2.67 ± 0.00	0
	HET-GCS	0.881 ± 0.005	314.8 ± 0.7	0.976 ± 0.002	0.369 ± 0.005	0.982 ± 0.002	0.997 ± 0.002	0.033 ± 0.002	0.75 ± 0.00	2.66 ± 0.00	0
	TGCS	0.881 ± 0.005	314.8 ± 0.7	0.976 ± 0.002	0.369 ± 0.005	0.982 ± 0.002	0.997 ± 0.002	0.037 ± 0.001	0.76 ± 0.00	2.69 ± 0.00	0
	Bilinear	0.862 ± 0.006	452.0 ± 1.0	1.395 ± 0.003	1.954 ± 0.009	0.000 ± 0.000	0.000 ± 0.000	3.258 ± 0.015	–	–	0
	FNO	1.150 ± 0.005	489.9 ± 1.0	1.515 ± 0.003	1.506 ± 0.007	0.000 ± 0.000	0.000 ± 0.000	1.835 ± 0.008	–	–	1
	PNO	1.113 ± 0.007	341.7 ± 0.9	1.203 ± 0.033	1.010 ± 0.049	0.819 ± 0.002	0.832 ± 0.003	0.326 ± 0.002	2.39 ± 0.90	4.85 ± 0.51	1
	Consistency	0.850 ± 0.005	313.0 ± 0.7	0.970 ± 0.002	0.342 ± 0.005	0.922 ± 0.002	0.937 ± 0.002	0.087 ± 0.001	0.97 ± 0.00	2.72 ± 0.01	2
	PDE-Refiner	0.787 ± 0.005	411.4 ± 1.0	1.269 ± 0.003	1.896 ± 0.010	0.100 ± 0.002	0.102 ± 0.002	1.535 ± 0.012	0.82 ± 0.00	0.41 ± 0.01	4
	Diffusion (EDM)	0.859 ± 0.005	312.8 ± 0.7	0.969 ± 0.002	0.335 ± 0.005	0.979 ± 0.003	0.995 ± 0.003	0.045 ± 0.001	1.01 ± 0.00	2.92 ± 0.01	35
	Diffusion (CorrDiff)	0.836 ± 0.005	310.9 ± 0.6	0.963 ± 0.002	0.324 ± 0.004	0.981 ± 0.002	0.996 ± 0.002	0.041 ± 0.001	1.01 ± 0.00	2.94 ± 0.01	35
	NCS-ES	0.836 ± 0.005	311.0 ± 0.6	0.965 ± 0.002	0.401 ± 0.005	0.970 ± 0.002	0.985 ± 0.002	0.453 ± 0.002	0.97 ± 0.01	1.85 ± 0.01	2
	NCS-Drift+ES	0.852 ± 0.005	312.6 ± 0.6	0.969 ± 0.002	0.385 ± 0.005	0.958 ± 0.002	0.973 ± 0.002	0.364 ± 0.012	0.90 ± 0.01	2.00 ± 0.02	2
	NCS-twES	0.855 ± 0.005	312.8 ± 0.7	0.970 ± 0.002	0.372 ± 0.005	0.958 ± 0.002	0.973 ± 0.002	0.208 ± 0.004	1.01 ± 0.01	3.04 ± 0.01	2
GARD–LENS											
	LIN	1.206	661.9	2.483	1.496	0.000	0.000	1.306	–	–	0
	NN-L_2	1.248 ± 0.056	645.9 ± 0.2	2.316 ± 0.004	1.640 ± 0.031	0.000 ± 0.000	0.000 ± 0.000	1.206 ± 0.004	–	–	1
	MARG	1.286 ± 0.010	505.9 ± 0.4	2.499 ± 0.004	7.738 ± 0.011	0.910 ± 0.000	0.924 ± 0.000	0.019 ± 0.000	0.01 ± 0.00	2.58 ± 0.00	0
	GCS	1.285 ± 0.010	505.9 ± 0.4	2.499 ± 0.004	7.713 ± 0.011	0.909 ± 0.000	0.923 ± 0.000	0.018 ± 0.000	0.01 ± 0.00	2.58 ± 0.00	0
	HET-GCS	1.289 ± 0.008	498.3 ± 0.3	2.458 ± 0.004	7.700 ± 0.013	0.914 ± 0.001	0.928 ± 0.001	0.020 ± 0.000	0.01 ± 0.00	2.82 ± 0.00	0
	TGCS	1.474 ± 0.004	507.3 ± 0.2	1.895 ± 0.002	2.524 ± 0.016	0.819 ± 0.001	0.831 ± 0.001	0.230 ± 0.001	0.13 ± 0.00	4.64 ± 0.00	0
	Bilinear	1.213 ± 0.000	662.5 ± 0.0	2.429 ± 0.000	1.457 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	1.529 ± 0.000	–	–	0
	FNO	0.500 ± 0.001	664.7 ± 0.2	2.422 ± 0.001	1.072 ± 0.003	0.000 ± 0.000	0.000 ± 0.000	0.969 ± 0.003	–	–	1
	PNO	1.290 ± 0.030	476.2 ± 0.6	1.612 ± 0.007	5.005 ± 0.793	1.003 ± 0.026	1.018 ± 0.027	0.550 ± 0.080	0.09 ± 0.00	4.58 ± 0.13	1
	Consistency	1.241 ± 0.045	514.7 ± 0.2	1.581 ± 0.004	1.156 ± 0.047	0.407 ± 0.004	0.414 ± 0.004	0.464 ± 0.017	0.21 ± 0.01	4.21 ± 0.02	2
	PDE-Refiner	0.757 ± 0.043	626.7 ± 1.6	2.226 ± 0.012	1.341 ± 0.036	0.051 ± 0.001	0.052 ± 0.001	1.194 ± 0.011	0.02 ± 0.00	3.61 ± 0.06	4
	Diffusion (EDM)	1.445 ± 0.165	513.0 ± 5.9	1.535 ± 0.009	1.233 ± 0.116	0.420 ± 0.038	0.427 ± 0.039	0.329 ± 0.071	0.16 ± 0.03	4.36 ± 0.04	35
	Diffusion (CorrDiff)	1.150 ± 0.117	508.0 ± 3.1	1.550 ± 0.017	1.297 ± 0.095	0.506 ± 0.033	0.513 ± 0.033	0.122 ± 0.040	0.12 ± 0.01	4.50 ± 0.03	35
	NCS-ES	1.184 ± 0.013	482.6 ± 0.3	1.921 ± 0.004	12.101 ± 0.224	0.802 ± 0.006	0.815 ± 0.006	0.466 ± 0.006	0.02 ± 0.00	3.45 ± 0.02	1
	NCS-Drift+ES	1.183 ± 0.035	473.7 ± 0.1	1.588 ± 0.010	1.896 ± 0.117	0.948 ± 0.012	0.962 ± 0.012	0.148 ± 0.013	0.17 ± 0.03	4.06 ± 0.04	1
	NCS-twES	1.184 ± 0.036	473.7 ± 0.1	1.583 ± 0.009	1.829 ± 0.145	0.950 ± 0.011	0.965 ± 0.011	0.159 ± 0.010	0.19 ± 0.02	4.06 ± 0.04	1

F.2 Experiment 4: when learned conditional transport is necessary

Question. Which departures from joint Gaussianity require a learned transport rather than a stronger classical innovation model?

Controlled non-Gaussian ladder. For each controlled departure, we compare GCS, HET-GCS, TGCS, and NCS while holding the conditional mean and source-estimation sample size fixed, at four severity rungs (Table 11).

The expected separation is:

1. a nonlinear conditional mean should separate NN-L_2 and NCS from LIN;
2. heteroscedasticity should separate HET-GCS from GCS;
3. fixed non-Gaussian marginals should favor TGCS;
4. condition-dependent skewness, tails, or dependence should require a conditional transport; and
5. cross-level dependence should expose the limitations of independent per-level innovations.

These are empirical hypotheses, not assumptions used to define the controls.

Real-system baseline ladder. On each PDE and GARD–LENS, Table 10 reports the complete baseline progression. We emphasize paired improvements:

$$\Delta_{\text{mean}} = \text{Metric}(\text{LIN}) - \text{Metric}(\text{NN-}L_2), \tag{39}$$
$$\Delta_{\text{innovation}} = \text{Metric}(\text{MARG}) - \text{Metric}(\text{GCS}), \tag{40}$$
$$\Delta_{\text{hetero}} = \text{Metric}(\text{GCS}) - \text{Metric}(\text{HET-GCS}), \tag{41}$$
$$\Delta_{\text{transform}} = \text{Metric}(\text{GCS}) - \text{Metric}(\text{TGCS}), \tag{42}$$
$$\Delta_{\text{learned}} = \text{Metric}(\text{TGCS}) - \text{Metric}(\text{NCS}). \tag{43}$$

We fix the classical benchmark in Δ_{learned} to TGCS: it is the top of the classical ladder (a transformed-Gaussian innovation source is the most expressive source the classical family admits), and on the tail metrics it is the classical ceiling—no other classical arm reaches a higher $\kappa/\kappa_{\text{data}}$. How much headroom that ceiling leaves is benchmark-dependent, and the transform step $\Delta_{\text{transform}} = \text{Metric}(\text{GCS}) - \text{Metric}(\text{TGCS})$ already exposes it: on Allen–Cahn the transform lifts $\kappa/\kappa_{\text{data}}$ from 0.70 to 0.97, so TGCS is a genuinely strong classical baseline there; on Navier–Stokes it lifts it only from 0.75 to 0.76, so the classical ceiling itself is nearly degenerate and Δ_{learned} on the tail there is measured against a baseline that a fixed marginal transform cannot move. Because metrics have different scales and orientations, these differences are reported metric by metric rather than combined into a single "non-Gaussianity index."

Each difference is reported *signed*, not as a magnitude: the sign carries the finding. Δ_{mean} makes this concrete. With RMSE and CRPS oriented so that lower is better, a positive Δ_{mean} means the learned conditional mean (NN-L_2) improves on linear interpolation (LIN); a negative value means LIN is ahead. On the accuracy metrics Δ_{mean} is a signed null on Allen–Cahn ($\text{RMSE}\ -0.04$, $\text{CRPS}\ -0.00$: LIN is, if anything, marginally ahead), so a nonlinear mean buys nothing over interpolation there; it is negative on Navier–Stokes ($\text{RMSE}\ -0.10$, $\text{CRPS}\ -0.04$), where LIN is ahead on both; and mixed on GARD–LENS ($\text{RMSE}\ -0.54$ but $\text{CRPS}\ +0.17$). Folding these to $|\Delta_{\text{mean}}|$ would report a uniform "mean effect" and hide that on one benchmark it is absent, on another it points the wrong way, and on the third the two metrics disagree in sign.

Source-versus-transport factorial ablation. We cross:

1. raw-residual Gaussian source;
2. Gaussian innovation source;
3. heteroscedastic Gaussian innovation source;
4. transformed-Gaussian innovation source;

with:

1. identity transport;
2. affine learned transport;
3. full neural transport.

This factorial design determines whether gains arise from source calibration, conditional transformation, or their interaction.

Factorial findings: three separable levers. Crossing the four sources with the three transports dissociates three effects that a single "non-Gaussianity" axis would conflate; each metric is moved by a different axis, and the three are close to orthogonal. *(i) The source spectrum sets logPSD.* Replacing the fixed-Gaussian source with a spectrally-fitted one drops logPSD from 1.3–1.7 to 0.12–0.52 across every Gaussian-family source, whereas at fixed source the affine-to-full transport step changes logPSD by less than 0.2 and never by an order of magnitude. Spectral realism is the source's job, not the transport's. *(ii) The transport sets the proper score.* At fixed source, full neural transport improves the energy score over the affine transport by $\approx 12\%$ on Allen–Cahn ($37.3 \to 33.0$) and $\approx 11\%$ on Navier–Stokes ($573 \to 513$), while leaving logPSD and the tail ratio essentially unchanged. The learned transport earns its keep on the headline proper score, not on the structural or tail metrics. *(iii) The source marginal sets the tail.* Only the transformed-Gaussian marginal (TGCS) reaches the data tail: on Allen–Cahn it attains $\kappa/\kappa_{\text{data}} \approx 1.1$ against 0.71–0.81 for the raw,

Gaussian-innovation, and heteroscedastic sources, and this is invariant to both spectrum and transport. On Navier–Stokes no cell exceeds 0.65, so the tail there is unreached by any source–transport combination—consistent with the objective, not the source, being the operative tail lever on that system (Experiment 5).

These three readings are only separable because the Gaussian-family sources were re-run on the spectrally-fitted envelope: the original factorial placed them on a fixed-Gaussian blur that could not match the residual spectrum, which confounded "transport does not fix logPSD" with "these sources cannot match it in the first place." Holding the spectrum correct and observing that full transport *still* leaves logPSD an order of magnitude below the source effect is what isolates logPSD as a property of the source spectrum alone.

Ladder table. Table 11 reports every arm across all four severity rungs, scored on the metric its hypothesis targets; the main-text reading is in §5.3 and Fig. 2a. For the condition-dependent spectrum, pooled metrics are blind by construction, so we report a condition-resolved *tracking* ratio: the innovation slope is set to $p_{\text{eff}} = p + \mathrm{d}p \cdot z(c)$ for a monotone function $z(c)$ of the coarse amplitude; test conditions are binned by $z(c)$ into low and high groups and $\mathrm{track} = \langle |\log_{10} \mathrm{PSD}_{\text{high}} - \log_{10} \mathrm{PSD}_{\text{low}}| \rangle_k$ for the arm is divided by the same quantity for the data, so 0 is one global spectrum and 1 is tracking c as the data does. Classical arms read ≤ 0.01 (95% CI $[0.003, 0.013]$); NCS reads 0.78 $[0.74, 0.81]$, 0.57 $[0.54, 0.61]$, and 0.24 $[0.22, 0.32]$ at severities 1–3 (condition bootstrap, three seeds pooled). The decline with severity is the limit of a scalar power-law tilt on a separation that is superlinear in $\log(1+k)$; an oracle tilt set to the generator's own $\mathrm{d}p$ tops out at the same level. NCS's pooled logPSD (0.15–0.44) stays inside the band every spectrally fitted source occupies, so per-condition reshaping costs nothing in pooled spectral fidelity.

Table 11: Controlled non-Gaussian ladder across four severity rungs. Each departure is scored on the metric its hypothesis targets (CRPS↓ for the nonlinear mean; $\kappa/\kappa_{\mathrm{data}}$, closest to 1 best, for marginal and scale departures; the tracking ratio for the spectrum). Best arm per rung in bold; the severity-0 anchor is an identity tie and is left unbolded. The spectrum anchor is undefined at $\mathrm{d}p = 0$.

Departure	Metric	Sev	GCS	HET-GCS	TGCS	NCS
nonlinear mean	CRPS↓	0	0.56	0.56	0.56	0.57
		1	0.66	0.66	0.66	**0.57**
		2	0.89	0.88	0.87	**0.58**
		3	1.16	1.15	1.13	**0.60**
hetero. scale	$\kappa/\kappa_{\mathrm{d}}$	0	1.00	1.00	1.00	1.10
		1	0.91	**0.99**	0.92	1.04
		2	0.72	0.95	0.81	**0.98**
		3	0.41	**0.88**	0.73	1.88
skew	$\kappa/\kappa_{\mathrm{d}}$	0	1.00	1.00	1.00	1.10
		1	0.88	0.88	0.96	**0.99**
		2	0.77	0.77	**0.97**	0.92
		3	0.68	0.68	**0.98**	0.88
heavy tail	$\kappa/\kappa_{\mathrm{d}}$	0	1.00	1.00	1.00	1.10
		1	0.80	0.80	0.84	**0.94**
		2	0.49	0.49	0.69	**0.87**
		3	0.16	0.16	0.50	**0.73**
point mass	$\kappa/\kappa_{\mathrm{d}}$	0	1.00	1.00	1.00	1.10
		1	0.53	0.53	0.56	**0.60**
		2	0.64	0.64	0.77	**0.81**
		3	0.61	0.61	0.93	**0.96**
cond. spectrum	track	0	–	–	–	–
		1	0.01	0.01	0.01	**0.78**
		2	0.01	0.01	0.01	**0.57**
		3	0.00	0.00	0.00	**0.24**
cross scale	$\kappa/\kappa_{\mathrm{d}}$	0	1.00	1.00	1.00	1.10
		1	0.92	0.92	0.92	**1.01**
		2	0.83	0.83	0.83	**0.91**
		3	0.60	0.60	0.61	**0.67**

Factorial cells. Table 12 gives every cell of the source×transport factorial on both source-spectrum variants. Variant A places the Gaussian-family sources on a fixed-Gaussian blur that cannot match the residual spectrum; variant B re-runs them on the spectrally fitted envelope. The three levers of §5.3 are separable only under variant B, because variant A confounds "transport does not fix logPSD" with "these sources cannot match it in the first place."

Table 12: Source×transport factorial (Experiment 4), mean over three seeds, on the small Navier–Stokes dataset (Table 3). Columns A: fixed-Gaussian source spectrum; columns B: spectrally fitted source. TGCS has no variant B because its source is already fitted. The full transport improves ES by 11–12% over the affine transport at every source; the fitted spectrum moves logPSD by an order of magnitude; only the transformed marginal moves the tail.

		fixed-Gaussian source spectrum (A)				fitted source spectrum (B)			
Source	Transport	ES	logPSD	σ/s	$\kappa/\kappa_{\mathrm{data}}$	ES	logPSD	σ/s	$\kappa/\kappa_{\mathrm{data}}$
Allen–Cahn									
MARG	affine	37.3	1.73	0.95	0.73	37.3	0.44	0.94	0.72
	full	33.0	1.65	0.92	0.74	33.0	0.36	0.91	0.75
GCS	affine	37.5	1.64	0.85	0.73	37.5	0.30	0.84	0.71
	full	33.0	1.61	0.91	0.73	33.0	0.20	0.90	0.73
HET-GCS	affine	38.0	1.53	0.72	0.81	38.0	0.23	0.72	0.78
	full	33.0	1.57	0.90	0.81	32.9	0.12	0.90	0.78
TGCS	affine	37.1	0.45	0.94	1.10	–	–	–	–
	full	32.7	0.36	0.94	1.12	–	–	–	–
Navier–Stokes									
MARG	affine	576.7	1.58	1.17	0.61	576.9	0.46	1.16	0.59
	full	512.8	1.43	1.00	0.62	513.2	0.52	1.00	0.63
GCS	affine	573.5	1.39	0.91	0.60	573.8	0.24	0.90	0.59
	full	512.8	1.31	0.99	0.61	514.0	0.37	0.99	0.63
HET-GCS	affine	578.2	1.32	0.82	0.60	580.1	0.21	0.82	0.59
	full	513.4	1.29	0.99	0.61	514.7	0.35	0.99	0.63
TGCS	affine	573.8	0.26	0.91	0.63	–	–	–	–
	full	513.8	0.41	0.99	0.65	–	–	–	–

F.3 Experiment 5: what proper-score training calibrates

Question. Which aspects of the conditional law are learned by energy-score training, and which require additional tail- or dependence-sensitive supervision?

Objective ablation. Using the same architecture and source, we compare:

1. L_2 training;
2. energy-score training;
3. energy score plus a dependence-sensitive score (variogram score of order $p = 1$);
4. a *band-weighted* energy score, computed on spectrally whitened fields so that each radial band contributes in proportion to its conditional variance rather than its raw power;
5. energy score plus threshold-weighted score; and
6. drifting pretraining without proper-score fine-tuning.

The arm numbering above is the one used throughout this section and in Table 2; arm 4 was cut (see below), so the table reports the five surviving arms.

Arm 3's score and its order are both fixed by measurement. A multiscale score is the other conventional choice, but the average-pool band bank we implement reweights the spectrum by a factor of just 1.31 from the largest to the smallest scale (mean implied weight 0.980 below the coarse-grid Nyquist versus 0.808 above), which makes it nearly plain energy score—arm 2—and in any case the same kind of object as arm 4, whose reweighting is orders of magnitude stronger; we therefore drop it. A differentiable variogram is not the obstacle it appears: on CPU one forward–backward pass costs under $2\times$ plain ES at $M = K = 8$ on a 128^2 grid, so cost does not decide the arm. The order does, and it decides in favour of $p = 1$.

The gradient of $|\delta|^p$ carries $p|\delta|^{p-1}$, so $p < 1$ and $p > 1$ fail at opposite ends and $p = 1$ at neither. For $p < 1$ the weight diverges as $\delta \to 0$, and an atom of exactly-zero increments captures the objective in a proportion set by the smoothing floor rather than by the data. Three of our seven

systems have such an atom, and its size is not what matters: at $p = 0.5$ the zero pairs take $1.57\times$ the share of gradient weight they already hold by distance weight on GARD–LENS (where 62.8% of sampled pairs are dry at both ends), $4.93\times$ on storm cells, and $224\times$ on stochastic Allen–Cahn, whose atom is a mere 4×10^{-6} of pairs. Changing the floor from 10^{-6} to 10^{-1} moves that share by 19.5%, 77%, and 99.3% respectively, which is the defect stated plainly: a hyperparameter, not the field, sets the emphasis. For $p > 1$ the weight instead grows without bound in δ, recovering the raw high-order variogram whose sensitivity to rare large increments motivated flatness matching in the first place: the top 1% of increments by distance-weight mass collect 2.2–$13.5\times$ their uniform share at $p = 1.5$ and 4.0–$36.2\times$ at $p = 2$, across all seven systems. At $p = 1$ exactly, $\mathrm{d}|\delta|/\mathrm{d}\delta = \operatorname{sign}\delta$ is constant, and all three diagnostics are clean by construction rather than by compromise: zero pairs receive exactly their base-rate share ($1.0000\times$, on each of the three systems that define the statistic), the objective is exactly floor-independent, and the top 1% of increments take exactly their uniform share ($1.0000\times$ on all seven). We therefore train arm 3 with $p = 1$ and, to keep the comparison from being scored on its own training loss, *judge* it on the $p = 0.5$ variogram score, reporting $p = 1$ only as a training trace.

Arm 4 is motivated by a measurement rather than by the literature. Because the plain energy score uses an unweighted L_2 norm, it inherits the red spectrum of the target: on our benchmarks the resolved band accounts for the large majority of $\mathbb{E}[\mathrm{Cov}(\delta \mid c)]$ in absolute terms (§E.1.1), so ES gradients are dominated by large-scale spread and are nearly indifferent to the fine bands where the conditional law is most diffuse. Whitening by the measured per-band deficit equalizes that contribution. Since the weights are estimated from training conditions only, the score remains proper. We nonetheless cut this arm: under both admissible readings its reweighting collapses onto an existing objective on three of seven systems, `gard_lens` among them, so it adds a hyperparameter without a distinct measurement. It is described here because it motivates the band-deficit diagnostic that §E.1.1 reports; the ablation below is the five surviving arms.

The comparison separates four distinct outcomes:

1. conditional-mean accuracy;
2. global ensemble spread;
3. condition-dependent spread;
4. tails and spatial dependence.

Table 2 reports these on Navier–Stokes, and the four outcomes separate onto different objectives rather than tracking one another. The four proabilistic arms tie on conditional-mean accuracy once the ensemble exists (ES, CRPS within their intervals for arms 2/3/5/6); the L_2 arm alone collapses spread ($\sigma/s = 0.375$ versus ≈ 0.97), which is outcome (ii) failing exactly where the energy score is meant to act. The variogram term (arm 3) buys spatial/spectral structure—logPSD 0.236 versus the energy score's 0.364, a 35% reduction—at no accuracy cost, but widens the increment-kurtosis gap (-0.95 versus -0.39). The threshold-weighted term (arm 5) is the only objective whose flatness (κ_{h2}) and $\Delta\zeta_4$ gaps close to the data, and it also has the lowest logPSD (0.208).

Table 13: Objective ablation on Navier–Stokes (§5.4, Table 2). One neural generator and source; arms differ only in the training objective. Values are condition-bootstrap point estimates with 95% intervals (2000 replicates, resampling seed→condition); the frozen split makes seed std a training-RNG quantity, not an error bar. Tail columns are gaps to the data increment law, so 0 is the target and the sign is the direction of the miss.

Objective	ES	CRPS	σ/s	logPSD	κ_{h2} gap	$\Delta\zeta_4$ gap	P99 gap
L_2 (arm 1)	367.0 [364.0, 369.6]	1.132 [1.123, 1.139]	0.375 [0.372, 0.378]	0.706 [0.700, 0.714]	-0.924 [-0.949, -0.900]	+0.308 [0.300, 0.315]	-1.713 [-1.722, -1.703]
ES (arm 2)	312.6 [311.3, 313.8]	0.969 [0.965, 0.973]	0.973 [0.969, 0.977]	0.364 [0.341, 0.392]	-0.386 [-0.440, -0.331]	+0.222 [0.211, 0.232]	-0.900 [-0.942, -0.857]
ES + variogram (arm 3)	312.7 [311.4, 313.9]	0.973 [0.969, 0.977]	0.964 [0.960, 0.968]	0.236 [0.231, 0.244]	-0.953 [-0.989, -0.913]	+0.270 [0.260, 0.281]	-0.740 [-0.757, -0.722]
ES + twCRPS (arm 5)	312.8 [311.5, 314.1]	0.970 [0.966, 0.974]	0.973 [0.969, 0.977]	0.208 [0.201, 0.217]	+0.027 [-0.101, 0.119]	+0.007 [-0.026, 0.036]	+0.143 [0.113, 0.167]
pretrain-only (arm 6)	312.1 [310.8, 313.4]	0.968 [0.964, 0.971]	0.951 [0.947, 0.954]	0.493 [0.487, 0.501]	-0.402 [-0.443, -0.363]	+0.213 [0.205, 0.222]	-0.917 [-0.929, -0.904]

Reach-limited systems: restoring the tail the energy score cancels. The Navier–Stokes ablation locates the tail deficit at the objective, and the supply-versus-reach taxonomy predicts *where* a fix must act: on a reach-limited system the increment tail is not supplied by any Gaussian-family source

(Experiment 4), and the energy score is weakly sensitive to the increment fourth moment, so a bare energy score does not recover it either. The remedy follows from the diagnosis rather than from a new architecture. A scale-free term that matches that fourth-order structure directly—the increment flatness $F(h) = S_4(h)/S_2(h)^2$, whose ratio form leaves the energy-score–calibrated second moment untouched—supplies exactly that structure back. We add this term to the energy-score objective on the controlled reach exemplar SHE/PAM ($\beta = 1$; increment kurtosis $\kappa_{h2} = 38.3$ against a Gaussian-source ceiling of 3.99, a reach ratio of 0.10) and sweep its dose across four decades of the initialization-normalized weight ρ_{init}, gating each arm on the guards the taxonomy identifies as binding: conditional-mean accuracy (nRMSE, one-sided $+2\%$), calibration (σ/s, two-sided $\pm 10\%$), and—the guard that actually binds here—the log-power spectrum (one-sided $+10\%$). Table 14 reports the sweep.

The reading is twofold. First, the fix works and is nearly free: at the smallest dose ($\rho_{\text{init}} = 0.03$) the term closes 93% of the kurtosis deficit (3.99 $\rightarrow$ 35.9, i.e. $\kappa/\kappa_{\text{data}}$ from 0.10 to 0.94) while nRMSE, σ/s, and logPSD all stay inside their guards—indeed nRMSE *improves* slightly and σ/s is unchanged. Second, more weight buys nothing and eventually costs: κ is flat at $\approx$36 across $\rho_{\text{init}} \in \{0.03, 0.3, 3\}$ while logPSD degrades monotonically and, by $\rho_{\text{init}} = 30$, σ/s collapses and the tail itself regresses. The correct dose is the smallest one that moves the tail; logPSD, not accuracy or calibration, is the binding guard. This is the taxonomy's payoff, not a proposed method—the diagnosis of *where* to intervene (the objective, on a reach-limited system) is what makes a one-line, scale-free term sufficient. The companion heavy-tailed system, storm cells, is evaluated under the same sweep as a paired control: there the vanilla objective's admissible kurtosis range already brackets the data law, so it is not a deficit case and no dose is expected to close a gap.

Table 14: Flatness-dose sweep on the reach-limited exemplar SHE/PAM ($\beta = 1$), Gaussian-source (ν_∞) family, 3 seeds per arm (§5.4, Table 2). A scale-free fourth-order matching term is added to the energy-score objective at increasing dose ρ_{init} (the gradient-norm–normalized weight at initialization; λ_{flat} is the raw coefficient). Data increment kurtosis $\kappa_{h2} = 38.30$; the $\rho_{\text{init}} = 0$ control (energy score alone) reaches only 3.99. "Deficit closed" is the fraction of the control-to-data kurtosis gap recovered. Guards are relative to the control (nRMSE $+2\%$ one-sided, σ/s $\pm 10\%$ two-sided, logPSD $+10\%$ one-sided); *Free* = all guards held, *Costly* = at least one broken. The smallest dose is the operating point; logPSD is the binding guard.

ρ_{init}	λ_{flat}	κ_{h2}	deficit closed	nRMSE	σ/s	logPSD	verdict
control	0	3.99 ± 0.61	–	0.415	0.948	3.896	Control
0.03	0.00042	35.91 ± 0.13	93.0%	0.413	0.949	3.923	Free
0.3	0.0042	36.02 ± 1.64	93.4%	0.411	0.948	4.386	Costly
3	0.042	35.90 ± 1.20	93.0%	0.413	0.874	4.556	Costly
30	0.42	29.23 ± 6.29	73.6%	0.413	0.415	4.349	Costly

Where the scale-free fix stops: GARD-LENS precipitation. The same one-line term does *not* transfer to the real-world reach system. GARD-LENS daily precipitation has the largest fourth-moment deficit in the study: the energy-score control of this sweep reaches $\kappa_{h2} = 35.6$ against a data law of 431.3 (ratio 0.08; single-stage recipe, whereas the two-stage NCS of Table 1 reads 0.17). Sweeping the identical flatness term across the same four decades of ρ_{init} recovers at most 54.7% of the gap ($\kappa \rightarrow 252$ at $\rho_{\text{init}} = 0.3$) even when *no* guard is enforced, and κ saturates and then decays with further dose—there is no dose that closes the tail. The recovery is also costly: the single dose that keeps every guard inside its band ($\rho_{\text{init}} = 0.03$) recovers only $\approx$47%, binding on nRMSE, and at the nominal, unnormalized weight the term closes 40% ($\kappa = 195$) while tripling logPSD ($0.10 \rightarrow 0.30$). Moreover, tail weighting is not helpful here. Adding the threshold-weighted score moves κ/κ_{data} only from 0.17 to 0.18, inside the seed spread, although the same term closes the homogeneous NS tail gap (13). A single threshold reweights marginal exceedances and it does not act on an amplitude-dependent spread ratio, which is what the GARD-LENS deficit is.

The reason is the same ratio form that makes the term free on SHE/PAM. The GARD-LENS deficit is not a homogeneous kurtosis gap but an *amplitude-conditional* one: binning wet cells by local intensity, the ensemble is over-dispersed by $3.8\times$ at low amplitude and under-dispersed to $0.87\times$

in the heavy tail, a within-condition variance ratio whose slope in log-amplitude crosses one. The increment flatness $F(h) = S_4(h)/S_2(h)^2$ divides out S_2 by construction and so matches only the pooled, scale-free fourth moment; it cannot bend an amplitude-dependent spread ratio, and pushing it hard enough to force the pooled statistic instead corrupts the spectrum. This is the boundary of the taxonomy's free fix: the diagnosis still says intervene at the objective, but on GARD-LENS the required statistic is amplitude-conditional, not scale-free.

Making the objective amplitude-conditional. The diagnosis is prescriptive: if the deficit is an amplitude-dependent spread ratio, an objective that is itself conditioned on local amplitude should bend it where a pooled term cannot. We build a four-rung ladder of objectives, each changing only the training score while the source and transport stay fixed: an energy score whitened pixelwise by the fitted per-amplitude scale $\sigma_{\mathrm{data}}(a)$ (L1); an energy score stratified by amplitude band (L2); a per-band Dawid–Sebastiani score, which matches each band's mean and variance (L3); and K62 slope matching, which matches the slope and intercept of $\log \mathrm{Var}(\delta_h \mid a)$ against $\log a$ at lag 1 (L4). L2–L4 are dosed by gradient norm and gated to the same guards as the global term; L1 has no dose, its lever being the frozen $\sigma_{\mathrm{data}}(a)$ curve (Table 15). L2–L4 close the precipitation tail without a guard violation, and L4 lifts the lag-1 kurtosis ratio from 0.14 to 0.81—78% of the deficit—while ES and σ/s are unchanged and logPSD *improves* from 0.148 to 0.071. L4's target is a second moment, not a fourth: increments whose variance scales with amplitude pool into a heavy-tailed scale mixture, so matching the variance–amplitude relation restores most of the pooled kurtosis that the scale-free flatness term could not reach at any admissible dose. L1, which only reweights the energy score across amplitudes, is the negative control: it under-disperses the ensemble ($\sigma/s = 0.64$), triples the spectral error, and lowers the kurtosis ratio to 0.07. Equalizing the score's emphasis across amplitudes is therefore not enough; the within-band second moment has to be constrained. The tail does not close fully ($\kappa/\kappa_{\mathrm{data}}$ plateaus below unity, as on every NCS arm). On the lag-2 scale of Table 1, L4 reads 0.83 ± 0.02, above every baseline in Table 10 (the strongest, Consistency, reads 0.21); tuned multi-step consistency sampling reaches higher values only while under-dispersed ($\sigma/s \leq 0.57$, Table 19), so L4 is the only arm that reaches this tail while calibrated. This confirms the taxonomy's prediction that the GARD-LENS deficit is reachable once the statistic is conditioned on amplitude.

Table 15: Amplitude-conditional objective ladder on GARD-LENS. Each rung adds one amplitude-aware term to the spectral-envelope NCS anchor. $\kappa/\kappa_{\mathrm{data}}$ is the increment-kurtosis ratio (1 is the data tail) at lag 1 ($\kappa_{\mathrm{data}} = 762$) and at lag 2 ($\kappa_{\mathrm{data}} = 431$, the scale of Table 1). Guards are relative to the anchor (nRMSE $+2\%$, σ/s $\pm 10\%$, logPSD $+10\%$ one-sided, ceiling 0.163). Mean $\pm$ s.d. over three seeds. L1 (amplitude-whitened ES) is the negative control: it under-disperses ($\sigma/s = 0.64$) and triples logPSD, missing every guard. Levers L2–L4 hold all guards; L4 lifts the tail from 0.14 to 0.81 (deficit closed 78%) while *improving* the spectrum.

				$\kappa/\kappa_{\mathrm{data}}$	
Arm	ES	σ/s	logPSD	lag 1	lag 2
NCS (spectral envelope)	473.7 ± 0.1	0.96 ± 0.01	0.148 ± 0.016	0.14	0.174 ± 0.030
+ amp.-whitened ES (L1)	489.8 ± 6.6	0.64 ± 0.07	0.433 ± 0.111	0.07	0.088 ± 0.005
+ amp.-stratified ES (L2)	473.7 ± 0.1	0.96 ± 0.02	0.160 ± 0.015	0.68	0.693 ± 0.009
+ per-band DSS (L3)	473.9 ± 0.1	0.96 ± 0.01	0.109 ± 0.022	0.77	0.781 ± 0.019
+ K62 slope matching (L4)	473.7 ± 0.1	0.96 ± 0.02	**0.071** ± 0.001	**0.81**	**0.834** ± 0.016

Training ensemble size. We measure the finite-M noise in the energy-score gradient directly—the relative standard deviation across independent batches of $\partial \mathrm{ES}/\partial\theta$ for a spread parameter, with surrogate ensembles carrying each dataset's own residual spectrum and scored against real residual targets. On three benchmarks (`stochastic_burgers`, `gard_lens`, `she_pam`) the noise follows the population $M^{-1/2}$ law over the whole range we measured ($M = 2$–128; fitted log–log slopes -0.71, -0.70, -0.69, and -0.56, -0.53, -0.53 over the top three rungs), so there is no elbow to find: the per-doubling gains are ratios of two noisy standard deviations and are individually consistent with $\sqrt{2}$. What bounds M is cost. Since the per-update cost is linear in M, a fixed

budget buys $1/M$ as many steps, and the budget-normalized noise RSD $\cdot \sqrt{M}$ falls by only 2.46–2.64× across $M = 2$–128 while varying by under 10% among the rungs at $M \geq 16$, so the training ensemble size is a variance-versus-cost tradeoff rather than a knob with an optimum.

F.4 Optimization study: is drifting a useful warm-start?

Question. Does drift pretraining improve the trained model once the training budget is controlled?

Comparisons. We compare drift pretraining followed by energy-score fine-tuning (drift+ES; 100 drift and 50 ES epochs) with energy-score training from the identity initialization (ES only), with budgets matched in epochs: *(i)* equal ES epochs (50 each), with the cold arm's plateaued ES curve at an 150-epoch budget as an equal-total-budget bound (Table 16); and *(ii)* equal total epochs, with a cold arm trained for 150 ES epochs and nothing else changed (Table 17). Drift pretraining without the ES stage is the pretrain-only arm of the objective ablation (Table 2). We report final held-out scores with their spread over three seeds.

Decision criterion. Drifting is described as a useful warm-start only if it provides a statistically stable improvement in final quality at a matched training budget. Otherwise it is removed from the principal method and retained only as a negative ablation.

Result: the gain survives matching the training budget. *(i)* At **equal ES steps** the warm arm already ends lower—ES 501.6 vs. 552.0 on Navier–Stokes, 30.7 vs. 31.6 on Allen–Cahn—despite both curves having plateaued (cold last-10-epoch slope -0.27 and $+0.007$ respectively). *(ii)* For an **equal-total-budget** the warm arm wins on both systems: Navier–Stokes 501.6 vs. cold 525.0; Allen–Cahn 30.7 vs. 32.1 . *(iii)* Final held-out quality recomputed from the stored ensembles confirms the gap is not an artifact of the ES history: energy score, ensemble nRMSE (0.826 vs. 1.181 on Navier–Stokes), and the log-PSD spectral ratio all favor the warm arm.

Drifting is therefore not merely an accelerator. At a matched training budget it delivers a modest but reproducible quality gain on both synthetic systems (Navier–Stokes and Allen–Cahn). A second, initialization-only ablation at equal epochs (Table 17) shows where the gain lies: in spatial structure (variogram, logPSD). This is why drift pretraining is part of the NCS recipe (§3).

Table 16: Matched-budget drifting ablation. "ES final" is the mean of the last five ES epochs; "ES held-out", nRMSE, and logPSD are recomputed from the stored ensembles. "ES @ matched" is the equal-total-budget comparison. Three seeds per cell, on the small Navier–Stokes dataset (Table 3).

System	Arm	ES final	ES held-out	nRMSE	logPSD	ES @ matched
Navier–Stokes	drift+ES	501.6	525.8	0.826	1.762	501.6
	ES only	552.0	574.7	1.181	1.843	525.0
Allen–Cahn	drift+ES	30.7	33.8	0.614	1.104	30.7
	ES only	31.6	34.1	0.648	1.232	32.1

Table 17: Initialization-only ablation: drift pretraining (100 drift + 50 ES epochs) against identity initialization (150 ES epochs), otherwise identical. Mean ± s.d. over three seeds, scored as Table 10 (variogram of order $p = 1$).

Initialization	ES	CRPS	logPSD	σ/s	Variogram
Allen–Cahn					
drift pretraining	**34.45** ± 0.05	**0.404** ± 0.001	**1.218** ± 0.009	**0.95** ± 0.00	**0.048** ± 0.001
identity	36.84 ± 0.06	0.417 ± 0.002	1.500 ± 0.008	0.92 ± 0.00	0.168 ± 0.005
GARD-LENS					
drift pretraining	**485.9** ± 0.4	**1.663** ± 0.012	**0.425** ± 0.015	**0.91** ± 0.00	**5.73** ± 0.92
identity	522.7 ± 0.3	1.921 ± 0.004	0.466 ± 0.006	0.82 ± 0.01	12.10 ± 0.22

F.5 MULTILEVEL COMPOSITION AND CROSS-SCALE CONSISTENCY

Question. When does the multilevel cascade improve over a single-level transport, and does independent per-level noise reproduce the required cross-scale law?

Capacity and training study. We compare a single level, a capacity-matched single level (*single-wide*, $\approx 97.7\%$ of the two-level parameter count), and the two-level cascade, the latter both with levelwise training only and with levelwise initialization followed by end-to-end proper-score training.

Cross-level latent ablation. We compare:

1. fresh independent innovation at each level;
2. a shared base latent projected across levels;
3. correlated latents fitted to empirical cross-level innovation covariance; and
4. an autoregressive latent in which level-$\ell+1$ noise is conditioned on previous-level innovations.

Consistency diagnostics. We measure the relative restriction error $\|R_{\ell+1\to\ell}\hat{u}_{\ell+1}-\hat{u}_\ell\|$, the cross-level residual coherence against its value on the data, and the finest-level proper scores and spectra.

Decision criterion. The cascade is presented as a substantive contribution only if it improves finest-level quality or cross-scale consistency over a capacity-matched single level. Where one level remains competitive, the hierarchy is not described as the source of that gain.

Result: the hierarchy earns its keep on cross-scale fidelity, not aggregate score. Table 18 reports both blocks, recomputed from each run's stored ensemble (three seeds per cell on Allen–Cahn and on the small Navier–Stokes dataset of Table 3). Against the *capacity-matched* single-level control (*single-wide*, $\approx 97.7\%$ of the two-level parameter count) the full hierarchy leaves the aggregate proper scores essentially unchanged—energy 509.6 vs. 511.6 and nRMSE 0.823 vs. 0.840 on Navier–Stokes—but transforms the *cross-scale* statistics: finest-level logPSD halves ($1.458 \to 0.697$), cross-level residual coherence rises from 0.462 to 0.944 (against a data target of 0.905), and the relative restriction inconsistency drops from 0.432 to 0.096. A single level is competitive on aggregate accuracy, and the hierarchy's contribution is spectral and cross-scale consistency. Levelwise training *alone* is not enough: *full, levelwise-only* loses accuracy on Allen–Cahn (nRMSE 0.791 vs. 0.610) and spectrum on Navier–Stokes (logPSD 0.866 vs. 0.697), so the end-to-end proper-score stage after levelwise initialization is doing real work.

Latent ablation. Across the three correlated-latent schemes (shared, empirical-covariance, and autoregressive) neither the finest-level proper scores nor the cross-level coherence separate from the independent-latent baseline beyond seed variance (Table 18), so we retain independent per-level noise as the default. We do not read the small empirical-covariance restriction/coarse-spectrum differences as a reliable effect at three seeds.

Table 18: Multilevel study, full table (§5.5). Finest-level quality (left) and cross-level consistency (right). *Single-wide* is the capacity-matched control: *full* must beat it for the hierarchy to earn its keep. $\bar{\rho}_b$ is the power-weighted sample cross-level residual coherence; $\bar{\rho}_b^{data}$ is the same statistic on the held-out data (the target the two new latent schemes aim to match). r_{restr} is the relative restriction inconsistency, logPSD$_c$ the coarse-band spectrum error. Mean±std over 3 seeds.

Dataset	Arm	σ/s	Energy	nRMSE	logPSD	$\bar{\rho}_b$	$\bar{\rho}_b^{data}$	r_{restr}	logPSD$_c$
Allen-Cahn	Single (1 level)	0.920 ± 0.003	33.0 ± 0.1	0.623 ± 0.007	1.655 ± 0.001	0.546 ± 0.006	0.891 ± 0.000	0.353 ± 0.002	0.735 ± 0.004
	Single-wide (capacity-matched)	0.924 ± 0.003	33.0 ± 0.1	0.621 ± 0.008	1.655 ± 0.001	0.545 ± 0.003	0.891 ± 0.000	0.353 ± 0.001	0.729 ± 0.004
	Full (indep. latent)	0.907 ± 0.003	32.8 ± 0.1	0.610 ± 0.008	1.114 ± 0.010	0.956 ± 0.000	0.891 ± 0.000	0.093 ± 0.001	0.827 ± 0.001
	Full, levelwise-only	0.907 ± 0.003	34.5 ± 0.1	0.791 ± 0.009	1.187 ± 0.007	0.956 ± 0.000	0.891 ± 0.000	0.094 ± 0.001	0.895 ± 0.003
	Full + shared latent	0.910 ± 0.005	32.8 ± 0.0	0.609 ± 0.003	1.110 ± 0.006	0.962 ± 0.001	0.891 ± 0.000	0.078 ± 0.002	0.818 ± 0.008
	Full + empcov latent	0.911 ± 0.005	32.8 ± 0.0	0.609 ± 0.003	1.114 ± 0.010	0.971 ± 0.001	0.891 ± 0.000	0.068 ± 0.001	0.843 ± 0.004
	Full + autoreg latent	0.909 ± 0.003	32.8 ± 0.1	0.610 ± 0.008	1.122 ± 0.012	0.956 ± 0.000	0.891 ± 0.000	0.091 ± 0.001	0.823 ± 0.001
Navier-Stokes	Single (1 level)	0.997 ± 0.000	512.8 ± 0.3	0.852 ± 0.003	1.435 ± 0.015	0.505 ± 0.009	0.905 ± 0.000	0.406 ± 0.004	0.842 ± 0.006
	Single-wide (capacity-matched)	0.990 ± 0.001	511.6 ± 0.5	0.840 ± 0.004	1.458 ± 0.007	0.462 ± 0.006	0.905 ± 0.000	0.432 ± 0.001	0.885 ± 0.007
	Full (indep. latent)	0.986 ± 0.001	509.6 ± 0.6	0.823 ± 0.005	0.697 ± 0.024	0.944 ± 0.004	0.905 ± 0.000	0.096 ± 0.004	0.330 ± 0.025
	Full, levelwise-only	1.003 ± 0.002	509.1 ± 0.5	0.820 ± 0.005	0.866 ± 0.029	0.918 ± 0.007	0.905 ± 0.000	0.151 ± 0.006	0.557 ± 0.022
	Full + shared latent	0.984 ± 0.002	509.4 ± 0.3	0.820 ± 0.003	0.673 ± 0.022	0.955 ± 0.005	0.905 ± 0.000	0.082 ± 0.004	0.278 ± 0.027
	Full + empcov latent	0.984 ± 0.002	509.1 ± 0.5	0.818 ± 0.004	0.664 ± 0.018	0.966 ± 0.004	0.905 ± 0.000	0.074 ± 0.003	0.234 ± 0.024
	Full + autoreg latent	0.986 ± 0.000	509.6 ± 0.6	0.822 ± 0.006	0.700 ± 0.032	0.950 ± 0.005	0.905 ± 0.000	0.093 ± 0.006	0.330 ± 0.033

F.6 Comparison with iterative generative baselines

Sampling-budget sweep. Multi-step baselines are evaluated across their sampling budget, not only at their default setting: EDM over Heun steps and FM-stoch over passes per level, three seeds each (Fig. 9).

Noise-schedule sweep on GARD-LENS. Zero-inflated, heavy-tailed fields can require a different diffusion noise range than the defaults of Table 1 ($\sigma_{\max} = 80$, 18 Heun steps). We therefore re-sample the stored GARD-LENS checkpoints of EDM, CorrDiff, and Consistency across the sampling-time noise ceiling $\sigma_{\max} \in \{2.5, 5, \ldots, 5120\}$ (twelve doubling values) and the number of sampler steps, 132 settings in total (Table 19). Spread rises monotonically with $\sigma_{\max}$ for every arm. Only CorrDiff reaches calibration, $\sigma/s = 1.00$ at $\sigma_{\max} = 2560$, and all 11 settings at which any arm reaches $\sigma/s \geq 0.90$ are CorrDiff's, at a CRPS of at least 1.64 and a logPSD of at least 0.37. EDM reaches 0.79 at the largest value tested and is still rising slowly; Consistency stays below 0.63. At their lowest-CRPS settings EDM and Consistency beat NCS on CRPS (1.51 and 1.49 vs. 1.59) and on $\kappa/\kappa_{\mathrm{data}}$, while remaining under-dispersed (σ/s 0.53 and 0.59). Multi-step consistency sampling (8–12 steps) reaches the data tail, up to $\kappa/\kappa_{\mathrm{data}} = 1.10$, but only at $\sigma/s \leq 0.57$ and logPSD ≥ 0.36. No swept setting is calibrated and matches NCS on CRPS and spectrum at once.

Table 19: Noise-schedule sweep of the diffusion-class baselines on GARD-LENS: for each arm, the default setting of Table 1, the lowest-CRPS setting, the setting closest to $\sigma/s = 1$, and (Consistency) the highest-κ setting, over 132 settings of the sampling-time $\sigma_{\max}$ (2.5–5120) and sampler steps. Mean ± s.d. over three seeds on the 64 test conditions of Table 1, $M = 32$; σ/s Fortin-corrected, $\kappa/\kappa_{\mathrm{data}}$ on the scale of Table 1. Only $\sigma_{\max}$ and the step count are varied at sampling time; the training noise distribution is unchanged, and settings are selected on the validation conditions. EDM's closest-to-one point is the largest $\sigma_{\max}$ tested.

Arm	Setting	$\sigma_{\max}$ / steps	NFE	σ/s	CRPS	logPSD	$\kappa/\kappa_{\mathrm{data}}$
EDM	default	80 / 18	35	0.43 ± 0.04	1.535 ± 0.009	0.329 ± 0.071	0.16 ± 0.03
	lowest CRPS	160 / 9	17	0.53 ± 0.08	1.512 ± 0.018	0.171 ± 0.075	0.22 ± 0.09
	closest to $\sigma/s = 1$	5120 / 9	17	0.79 ± 0.15	1.584 ± 0.043	0.291 ± 0.120	0.28 ± 0.15
CorrDiff	default	80 / 18	35	0.51 ± 0.03	1.550 ± 0.017	0.122 ± 0.040	0.12 ± 0.01
	lowest CRPS	160 / 9	17	0.72 ± 0.07	1.541 ± 0.019	0.188 ± 0.066	0.12 ± 0.01
	closest to $\sigma/s = 1$	2560 / 9	17	1.00 ± 0.10	1.692 ± 0.103	0.531 ± 0.126	0.10 ± 0.02
Consistency	default	80 / 2	2	0.41 ± 0.00	1.581 ± 0.004	0.464 ± 0.017	0.21 ± 0.01
	lowest CRPS	320 / 4	4	0.59 ± 0.01	1.491 ± 0.002	0.346 ± 0.014	0.49 ± 0.01
	closest to $\sigma/s = 1$	640 / 1	1	0.63 ± 0.01	1.543 ± 0.004	0.200 ± 0.031	0.29 ± 0.02
	highest $\kappa/\kappa_{\mathrm{data}}$	640 / 12	12	0.56 ± 0.01	1.516 ± 0.007	0.406 ± 0.015	1.10 ± 0.22
NCS	(Table 1)	–	1	0.96 ± 0.01	1.588 ± 0.010	0.148 ± 0.013	0.17 ± 0.03

Matched information and architecture. All baselines receive the same coarse conditioning variables and are trained on the same condition-level split. Where architectural matching is impossible,

we report parameter count and inference cost (Table 20) and avoid attributing differences uniquely to the generative objective.

Result: the one-step operating point sits on the multi-step front. Figure 9 sweeps the two multi-step baselines across their sampling budgets—EDM over Heun steps $\{1, \ldots, 18\}$ (NFE $= 2$ steps $- 1$) and FM-stoch (the NCS backbone trained with plain conditional flow matching and sampled stochastically) over per-level NFE $\{1, 2, 4, 8\}$—by re-sampling the same trained checkpoints on the frozen evaluation subset, and overlays NCS as a single fixed one-step operating point (NFE=2). Three findings. First, the few-step regime of the Heun sampler is *unreliable*: at three steps EDM's energy score blows up (to ≈770 on Allen–Cahn and ≈2500 on Navier–Stokes) before recovering, so a naive low-NFE diffusion point is not a calibrated operating point. Second, both multi-step methods only reach their quality floor by NFE ≈8–11 (Allen–Cahn energy ≈33, Navier–Stokes ≈550). Third, NCS's single NFE=2 point already lies on or below that floor at every metric: it matches the best multi-step energy on Allen–Cahn (32.6 vs. 32.6–33.0) and *beats* it on Navier–Stokes (512 vs. 548–552), while its log-PSD is far lower on both systems (0.29/0.34 vs. 0.78–1.09) and its spread/skill sits at the calibration line. The sweep is shown on the two systems spanning the diffusive (Allen–Cahn) and turbulent (Navier–Stokes) regimes; the remaining systems carried no swept checkpoints in this inference-only run.

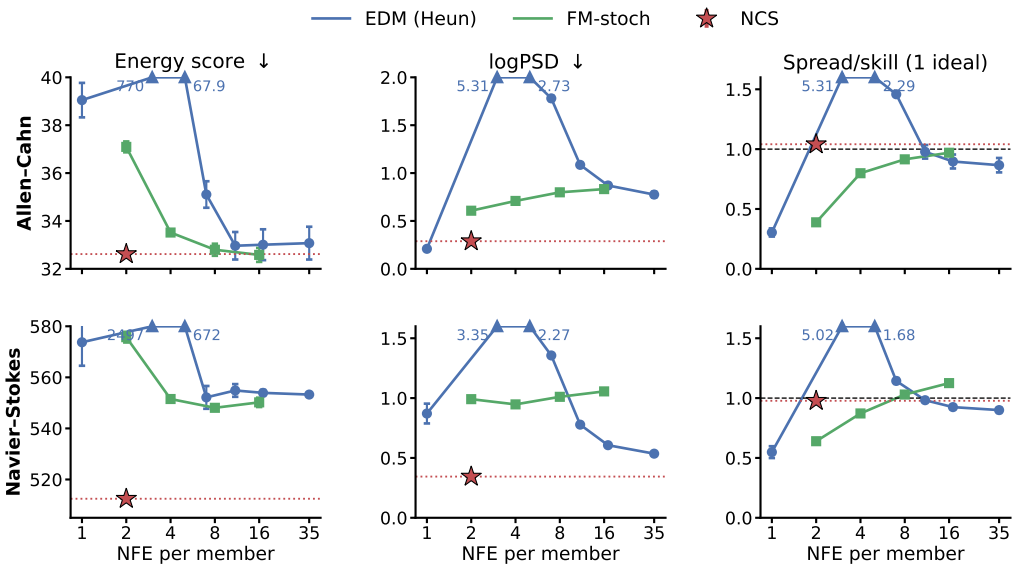


Figure 9: Quality against sampling budget. EDM (Heun) and a stochastic flow-matching cascade (FM-stoch) are swept by re-sampling one checkpoint per seed; NCS ($\star$, dotted line at its value) is one point at NFE=2. Axes are clipped to the frontier; triangles mark off-scale points with their value. Mean $\pm$ s.d. over three seeds; dashed line $\sigma/s = 1$. This is a separate re-sampling evaluation on the small Navier–Stokes dataset (Table 3), so absolute values are not comparable with Table 1.

F.7 Sampling cost

Neural-function evaluations are reported but are not treated as a sufficient measure of cost. For each arm we report measured per-member wall-clock latency, peak accelerator memory, and parameter count (Table 20).

For the cascade, the per-member cost is

$$C_{\mathrm{NCS}} = \sum_{\ell=0}^{L-1} C_\ell, \tag{44}$$

not simply L equal-cost evaluations, because levels have different spatial resolutions. For iterative baselines,

$$C_{\text{iter}} = \sum_{t=1}^{T} C_t \tag{45}$$

at the resolutions actually used by the sampler.

Table 20 reports the measured sampling cost of every arm on one NVIDIA GH200 GPU, drawing a 32-member ensemble for 8 conditions with 7 timed repeats after warmup. Sampling cost is a property of the architecture, the sampler-step count, and the batch shape, and is independent of the learned weight values, so the arms are built on synthetic data at each benchmark's real resolution and timed with device synchronization around each sampling call. The comparison that matters is at matched network capacity. NCS and EDM share the same 759K-parameter velocity-UNet backbone; NCS uses one instance per cascade level (1.52M on the two-level Allen–Cahn and Navier–Stokes cascades, 759K on GARD-LENS) and CorrDiff two (1.52M). NCS draws each member in one network pass per level against EDM's 35, and the measured per-member wall-clock tracks that ratio almost exactly: NCS is 17× faster on Allen–Cahn, 28× on Navier–Stokes, and 36× on GARD-LENS. CorrDiff runs the same 35-step wall as EDM, and on the two-level systems NCS matches its parameter count exactly while still being 17–28× faster, confirming the gap is the sampler and not the model size. Peak GPU memory is comparable across the UNet-based arms because a multi-step sampler holds only one state at a time, so the speedup is not bought with memory.

Table 20: Measured sampling cost on one NVIDIA GH200 GPU (32 members, 8 conditions, median of 7 timed repeats after 3 warmup passes). NFE and parameter count are for the two-level cascades (Allen–Cahn, Navier–Stokes); GARD-LENS is a single refinement level, where the cascade arms' NFE halve and every UNet arm has 759K parameters except CorrDiff (1.52M). Navier–Stokes is the small dataset ($32 \to 128$). Wall-clock is per ensemble member; peak GPU memory is for the Navier–Stokes (128^2) run. NCS uses one 759K velocity-UNet instance per level, the same backbone as EDM, isolating the sampler as the source of the wall-clock gap; CorrDiff runs two instances (1.52M). FM-det is the deterministic flow-matching cascade (an over-smoothing, ensemble-size-1 anchor); FM-stoch is the same cascade trained with plain conditional flow matching and sampled stochastically, i.e. the drift-versus-plain-flow ablation of the NCS backbone.

			Wall-clock (ms/member)			Peak GPU
Arm	NFE	Params	AC	NS	GARD	(MiB, NS)
Classical (zero-learning)						
Bilinear	0	0	0.0003	0.0002	0.0002	81
Linear	0	23	0.006	0.017	0.017	81
Source-only	0	0	0.025	0.024	0.024	97
MARG	0	113	0.108	0.418	0.411	80
GCS	0	113	0.108	0.420	0.411	80
HET-GCS	0	118	0.112	0.422	0.415	80
TGCS	0	1,137	0.389	1.508	1.505	80
One-pass learned						
FNO	1	1.19M	0.010	0.010	0.010	230
NN-L_2	1	8.40M	0.010	0.010	0.010	337
PNO	1	1.19M	0.025	0.080	0.080	2478
FM-det	2	1.01M	0.028	0.030	0.025	250
MeanFlow	1	759K	0.429	0.728	0.729	440
NCS (this work)						
NCS	2	1.52M	0.886	0.950	0.742	446
NCS (no pretrain)	2	1.52M	0.898	0.954	0.744	446
Multi-step generative						
FM-stoch	8	1.01M	3.261	3.513	2.915	266
Consistency	2	759K	0.873	1.463	1.465	444
PDE-Refiner	4	759K	1.644	2.894	2.905	440
EDM	35	759K	15.03	26.43	26.35	441
CorrDiff	35	1.52M	15.12	26.40	26.42	449

G ARCHITECTURE AND TRAINING DETAILS

Backbone. All learned arms share one backbone class, a velocity UNet. NCS uses one 759K-parameter instance per cascade level (1.52M for the two-level cascades); EDM, Consistency, PDE-Refiner, and MeanFlow use one instance of the same size; CorrDiff uses two instances (regression and residual-diffusion networks), 1.52M in total. PNO and FNO use an FNO backbone with 1.19M parameters; NN-L_2 is an 8.40M-parameter FNO. The conditioning field is concatenated as an input channel; the classical innovation ε enters NCS as a second input channel, and the network output is added to ε so that a zero-initialized final layer gives the identity map at initialization.

NCS training schedule. Drift pretraining: 100 epochs, $M = 16$ generated members per condition, Riesz kernel with $\beta = 1$, drift scale 1.0, batch size 2, Adam with base learning rate 2×10^{-4}. Energy-score fine-tuning: 50 epochs, $M = 8$ members, end-to-end through the cascade, batch size 1. The NCS+tw arm adds a threshold-weighted score at threshold $u = 1.5$ (in standardized units, two-sided absolute) with weight set by a gradient-norm match at initialization ($\alpha = 0.3$; 0.1 on Allen–Cahn where the stability gate triggers). Three training seeds per arm; hyperparameters selected on the validation split and frozen before test evaluation. Evaluation uses $M = 32$ members per condition throughout.

Classical stage. Wiener filter at $\lambda_{\mathrm{rel}} = 10^{-6}$ relative to each shell's mean eigenvalue, no shell smoothing, all supported shells retained; the innovation spectrum is Eq. 22. Statistics are estimated on the training split only.

Compute. Training and timing on NVIDIA GH200 GPUs; sampling-cost measurements in Table 20.

H CLAIM CRITERIA AND THEORY-TO-EXPERIMENT MAP

We adopted the following criteria before making the corresponding claims; Table 21 then maps each theoretical step to the experiment that tests it and the result that would falsify it.

1. **Value of stochastic simulation:** probabilistic methods must improve held-out proper scores and conditional coverage relative to deterministic point prediction.
2. **Value of innovation fitting:** GCS must improve aggregate or conditional calibration over MARG, and the Gaussian control must reproduce the covariance difference predicted by Lemma 2.
3. **Value of nonlinear learning:** NCS must improve over GCS, HET-GCS, TGCS, and source-only controls on at least one preregistered distributional property without unacceptable degradation elsewhere.
4. **Calibration claim:** "well calibrated" is used only if aggregate spread, conditional spread, coverage, and tail diagnostics jointly support it. Otherwise claims are qualified by the property measured, such as "globally variance calibrated."
5. **Value of drifting:** drifting must improve final performance at a matched training budget.
6. **Value of the cascade:** the hierarchy must improve finest-level quality or cross-scale consistency relative to a capacity-matched single-level model.
7. **Sampling-efficiency claim:** NCS must lie on or improve the empirical quality–cost frontier in measured latency or FLOPs, not only in neural-function evaluations.
8. **Any "recovers $Y\%$" claim:** a fraction of an error gap is reported only alongside the gap's ceiling, computed on the same split and metric, and only where a reference estimator that provably cannot solve the problem scores far below the arm being credited (§E.3.5). Where no ceiling exists—as on the observational benchmark, whose conditional mean has no closed form—the number is stated as an unnormalized improvement over the named baseline and not as a share of what is achievable.

H.1 THEORY-TO-EXPERIMENT MAP

Table 21 summarizes the empirical implication associated with each theoretical step and the result that would falsify the proposed interpretation.

Table 21: Theory-to-experiment map. The final column specifies the empirical outcome that would invalidate or materially weaken the corresponding interpretation.

Theory step	Prediction	Primary experiment	Falsification or weakening criterion
Projection	Deterministic point prediction omits mean conditional covariance	Held-out variance and spectral decomposition	Estimated deficit is inconsistent with within-condition covariance beyond finite-sample uncertainty
Innovation	Raw-residual additive noise double-counts linearly explained covariance	MARG versus GCS on Gaussian control and real data	Observed Gaussian-control overdispersion does not match the predicted covariance difference
Gaussian simulation	Schur-complement innovation sampling is exact under joint Gaussianity	Oracle and estimated Gaussian controls	Oracle implementation fails conditional covariance, coverage, or proper-score checks
Regularization	The damped Schur complement is not the residual spectrum of the regularized filter, so the source must be built from Eq. equation 22	λ ladder of the finite-data study (§E.3.4)	The two forms remain within tolerance across the whole ladder, in which case the distinction is untested here rather than confirmed
Stationarity	A stationary spectral filter forgoes conditional-variance accuracy that a learned correction can recover	Stationarity stress test against an exact ceiling (§E.3.7)	The ceiling is reached by a pointwise reference arm (degenerate control), or survives freezing the modulation (the gap is filter isotropy, not nonstationarity)
Conditional transport	Learning is needed only for departures not captured by classical innovation models	Controlled non-Gaussian ladder and real-system baseline progression	NCS does not improve over the strongest classical baseline or gains disappear under matched capacity and seeds
Proper scoring	Population-distribution training improves conditional spread relative to point prediction	Objective and ensemble-size ablations	ES training does not improve held-out proper scores or calibration relative to deterministic/source-only controls
Drifting warm-start	Drifting improves optimization but is not the source of calibration	Matched-budget pretraining study	No final-quality benefit at a matched training budget
Cascade	Multilevel composition improves cross-scale quality	Capacity-matched single-level and latent-sharing study	A capacity-matched single-level model matches the cascade's cross-scale consistency