CONDITIONING WITHOUT GUIDANCE: WHEN THE OBSERVATION IS ALREADY A DIFFUSION MARGINAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditioning a diffusion model usually approximates an unavailable likelihood score at inference. We identify a restricted case where this is unnecessary: when an observation has the law of a reachable forward marginal, conditioning reduces to restarting the reverse process there. With an exact score and exact marginal match, restart samples the Bayes posterior.

We instantiate this idea with *scale-time diffusion*, which destroys Fourier modes in scale order. A two-segment schedule removes the unobserved band and then matches sensor noise. Noiseless observations remain fixed by the reverse dynamics; noisy ones are denoised before they freeze.

Across two chaotic dynamical systems involving turbulent flows, the method substantially reduces reconstruction error relative to learned baselines in the first three missing octaves while retaining fine-scale power through repeated assimilation. A progressive clock uses a finer unseen observation with an exactness–utility tradeoff. The theorem covers matched projection sensors. The production endpoint has a climatology-averaged joint-law total-variation bound of 9.5×10^{-4}; sparse sensing and useful off-design entry remain outside exact matching.

1 INTRODUCTION

Conditioning a diffusion model (Ho et al., 2020) on a measurement usually requires an inference-time correction. Classifier guidance (Dhariwal & Nichol, 2021) and inverse-problem methods such as diffusion posterior sampling (DPS) (Chung et al., 2023), reconstruction guidance (Ho et al., 2022b), and ΠGDM (Song et al., 2023) add or approximate a likelihood score because the observation lies outside the model's forward process. For inverse problems, $\nabla_x \log p(y \mid x_s)$ is generally unavailable at intermediate noise levels, so conditional accuracy depends on both the learned prior and this second approximation.

We study the case in which that mismatch can be removed by design. If a measurement has the law of a reachable forward marginal, conditioning can be performed by starting the reverse process at that marginal. With an exact score and exact marginal match, this restart samples the Bayes posterior without a guidance gradient or tuning weight. We realize this idea with *scale-time diffusion*: a cutoff descends through wavenumber and destroys modes in scale order, while a separate noise ramp matches the sensor variance. A band-limited observation can then be an entry state of the process. The condition is restrictive and testable: the sensor must be a matched projection onto a reachable subspace and must satisfy the entry noise budget. Spectral truncation qualifies; sparse point sensing does not.

The comparison with guidance clarifies the mechanism. In a noiseless diagnostic, restart maintains its conditioning value to 2.8×10^{-8} at every reverse state. DPS matches its missing-band reconstruction at the unique unit effective step, $w/\sigma_\varepsilon^2 = 1$; because spectral H is an orthogonal projection, that clean-state correction is itself a projection. Its full sampler-state residual remains 8.0×10^{-3}. Thus the result is not an untunable performance gap: DPS can reproduce the update after selecting a special likelihood weight, whereas the forward construction supplies it without that extra tuning parameter (Section 4).

We evaluate super-resolved data assimilation for two turbulent systems. A 256^2 field is observed on a 32^2 spectral grid and each analysis is fed back into the next forecast. The missing band contains

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.

98.3% of live modes but only 17.8% of variance, so aggregate error can conceal a method that has stopped predicting fine scales. We therefore report phase and power per octave, as well as cycled stability. The continuous clock also supports entry from a finer observation than seen during training, although finite transition width makes that useful off-design entry approximate.

Contributions.

- We define a scale-ordered, mode-dependent forward process within the blurring-diffusion formalism (Hoogeboom & Salimans, 2023) (Section 3).
- We prove that exact reverse integration from a matched entry state samples $p(\omega_t \mid y_t, c)$, and delimit the result to matched projections. We quantify the production schedule's finite endpoint and show that DPS reproduces the reconstruction only when its clean-state correction becomes a projection (Section 4).
- We show zero-shot entry from a finer 96^2 observation, improving the first missing octave $9.4\times$, and measure its exactness–utility tradeoff (Section 7.2).
- Against a retrained refinement cascade (Dhingra et al., 2026), deterministic super-resolution, and EnKF variants, we evaluate two systems over repeated assimilation cycles; a paired test resolves no total-error difference from a perfect-dynamics EnKF at cycle 25 (Section 7.3).

2 Related Work

Conditioning a pretrained diffusion model. DPS (Chung et al., 2023), reconstruction guidance (Ho et al., 2022b), and ΠGDM (Song et al., 2023) approximate the measurement score in different ways. DDNM (Wang et al., 2023) instead projects onto the measurement null space, while RePaint (Lugmayr et al., 2022) and ILVR (Choi et al., 2021) repeatedly restore known pixels or low frequencies. Thus exact linear data consistency is not our novelty. Our claim concerns distributional matching: below s_{dead} the measured coordinates have zero diffusion coefficient, so retaining them follows the model's own dynamics (Section 4). DDRM (Kawar et al., 2022) is closest technically: it matches noise per singular value at inference time for an isotropic pretrained model. We incorporate that matching into the trained forward process. SDEdit (Meng et al., 2022) also restarts a reverse process, but its guide need not be a forward marginal.

Forward processes other than isotropic noise. Blurring diffusion (Hoogeboom & Salimans, 2023) establishes score matching and reverse sampling for fixed-basis Gaussian processes with per-coordinate schedules. Inverse heat dissipation (Rissanen et al., 2023) and cold diffusion (Bansal et al., 2023) study deterministic coarse-to-fine degradations, and soft diffusion (Daras et al., 2023) treats general corruptions. Cascaded diffusion (Ho et al., 2022a), SR3 (Saharia et al., 2023), and the assimilation cascade of Dhingra et al. (2026) refine a discrete resolution ladder; our continuous clock instead exposes intermediate cutoffs, including a 96^2 observation between dyadic rungs (Section 7.2). CorrDiff (Mardani et al., 2025) uses generative downscaling for physical fields without a cycled assimilation loop.

Diffusion models for data assimilation. Score-based data assimilation (DA) (Rozet & Louppe, 2023) conditions trajectory samples, DiffDA (Huang et al., 2024) combines a diffusion analysis with a forecast, and learned forecasters (Lam et al., 2023) now support variational DA (Adrian et al., 2025); our context model is a Fourier neural operator (FNO) (Li et al., 2021). MASF (Yoon et al., 2026) is the closest recent comparison: it designs a measurement-aware forward process and likelihood score for nonlinear and sparse sensors, whereas we prove an unguided restart for a narrower matched projection. Ayapilla et al. (2026) study cycled super-resolution DA with uncertainty and an EnKF comparison, using conditional generation rather than restart. Following Hodyss & Morzfeld (2026), we name the posterior through its prior, context, and likelihood in Proposition 1. Sequence-level conditioning (Chen et al., 2024) and function-space diffusion (Lim et al., 2025; Pidstrigach et al., 2024) address adjacent temporal and discretization questions; we make no discretization-independence claim.

3 SCALE-TIME DIFFUSION

3.1 FORWARD PROCESS

A standard diffusion corrupts every coordinate at once. Ours corrupts them in scale order. At scale-time s, a cutoff $k_c(s)$ descends through wavenumber and each Fourier mode follows the Gaussian marginal

$$\widehat{x}_{\boldsymbol{k}}(s) = a_{\boldsymbol{k}}(s)\,\widehat{x}_{\boldsymbol{k}}(0) + \sigma_{\boldsymbol{k}}(s)\,\widehat{\xi}_{\boldsymbol{k}}, \qquad \widehat{\xi}_{\boldsymbol{k}} \sim \mathcal{CN}(0,1) \text{ i.i.d.}, \tag{1}$$

with signal coefficient $a_{\boldsymbol{k}}$ and noise scale $\sigma_{\boldsymbol{k}}$. Each mode has a clock synchronized to the cutoff,

$$\tau_{\boldsymbol{k}}(s) = \Phi\left(\frac{\log k - \log k_c(s)}{w}\right), \qquad a_{\boldsymbol{k}} = \sqrt{\bar{\alpha}(\tau_{\boldsymbol{k}})}, \quad \sigma_{\boldsymbol{k}}^2 = E(k)\left(1 - \bar{\alpha}(\tau_{\boldsymbol{k}})\right), \tag{2}$$

where Φ is a smooth switch of width w octaves and $\bar{\alpha}$ is the cosine schedule of Nichol & Dhariwal (2021). High wavenumbers run through the clock first. The observed square $\|\boldsymbol{k}\|_\infty \le k_0$ is exempt, which enables the restart in Section 4.

In whitened coordinates $z_{\boldsymbol{k}} = \widehat{x}_{\boldsymbol{k}}/\sqrt{E(k)}$, every mode follows the same variance-preserving diffusion on a shifted clock. This is a fixed-basis non-isotropic diffusion (Hoogeboom & Salimans, 2023), so standard reverse sampling, including DDIM-type solvers (Song et al., 2021a) and diagonal updates (Zhang et al., 2023), applies. We use the measured per-mode climatological variance $E(k)$, rather than a power-law fit, so a destroyed mode ends with its data variance. Section C gives the measured spectrum, geometry, flat-noise control, and relation to deterministic degradations (Rissanen et al., 2023; Bansal et al., 2023); Figure 1 summarizes the matched restart.

3.2 THE TWO SEGMENTS, AND WHY THERE MUST BE TWO

Scale-time splits at a *sweep-completion time* s_{dead} into two segments. During the **sweep** $[0, s_{\mathrm{dead}}]$, k_c descends and destroys only the unobserved band; measured modes remain at $a_{\boldsymbol{k}} = 1, \sigma_{\boldsymbol{k}} = 0$. During the **ramp** $[s_{\mathrm{dead}}, 1]$, the cutoff is fixed, the missing band remains climatological noise, and the measured band accumulates sensor noise. Thus the two segments corrupt disjoint coordinates.

This separation is required for matched entry. If sweeping and ramping overlap, modes between k_0 and $k_c(s)$ still contain signal although an entry state built from the observation must fill them with noise, an order-one violation of Proposition 1. A finite-width clock also requires the sweep to pass below k_0; otherwise the first missing modes retain $a_{\boldsymbol{k}} \approx 0.7$. We therefore overshoot to $\kappa = k_0 2^{-mw}$ and start above the square storage corner at $k_c(0) = 336$. This leaves measured coordinates untouched but creates the off-design information tradeoff in Section 7.2. Full endpoint and corner-mode calculations appear in Sections C and D; the exact-versus-finite endpoint distinction is in Section B.

3.3 PROGRESSIVE VERSUS SHARED CLOCKS

Equation (2) gives every mode its own clock. The obvious alternative is a *shared* clock, $\tau_{\boldsymbol{k}}(s) = s$ for all unobserved modes. A progressive clock produces an approximately step-shaped marginal: modes below $k_c(s)$ are intact and modes well above it are destroyed. A shared clock partially corrupts every missing mode at each interior time and has only one band-limited endpoint.

The step shape lets a progressive model use a finer observation by entering at $k_c(s^*) = k_0'$. Finite transition width makes this entry approximate, and Section 7.2 measures the tradeoff between marginal matching and retained data. Shared clocks and discrete ladders (Ho et al., 2022a; Saharia et al., 2023) cannot expose an intermediate 96^2 cutoff. This flexibility has a cost: at the trained resolution, the shared clock is better by $25.4/12.4/5.4/2.3\%$ per octave because the progressive model generates the first missing band last and then propagates it through the remainder of the reverse path (Section 7.1). The same band receives the largest benefit from later off-design entry. Figure 7 in the appendix visualizes this geometry.

3.4 TRAINING

The denoiser is trained by conditional denoising score matching under elucidated diffusion model (EDM) preconditioning (Karras et al., 2022). The network receives the physical field under one

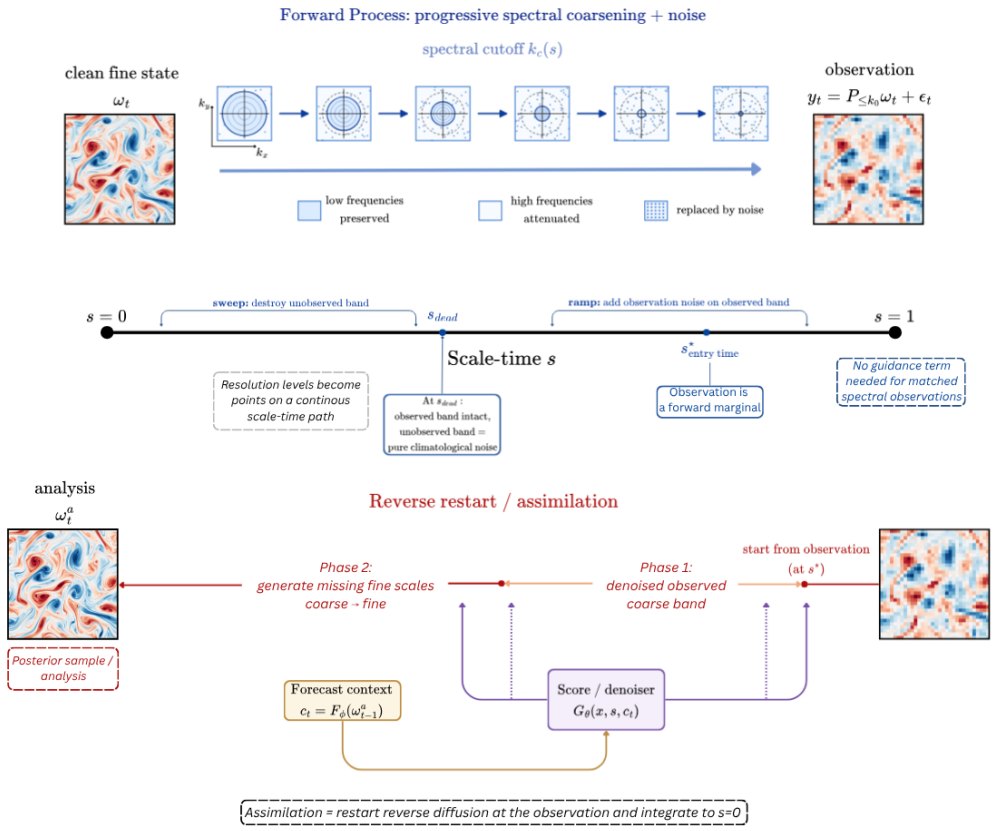

Figure 1: **The observation is a frame of the forward process.** The process destroys modes as $k_c(s)$ descends. At the matched endpoint, a 32^2 spectral observation supplies the retained band and climatological noise supplies the rest. Assimilation reverses from that entry state. The implementation approaches the exact endpoint with finite overshoot, as detailed in Section B.

global normalization; whitening is confined to the target and loss weights, following Hoogeboom & Salimans (2023). We supervise a mode exactly where its diffusion coefficient satisfies $g_{\boldsymbol{k}}(s) > 0$ and compose the coefficients algebraically so they remain finite as $a_{\boldsymbol{k}} \to 0$. We assign 25% of training draws to the sensor-noise ramp and sample the sweep uniformly in $\log k_c$. The context c_t is an FNO forecast (Li et al., 2021), teacher-forced in training and autoregressive at inference. The objective, architecture, sampling rationale, exposure gap, and forecaster diagnostics are in Section E.

4 Conditioning Is a Restart

Definition 1 (Spectral observation)**.** *The sensor returns the field restricted to a square of low wavenumbers,*

$$y_t \;=\; H\omega_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, R), \tag{3}$$

where H projects onto $\{\boldsymbol{k} : \|\boldsymbol{k}\|_\infty \le k_0\}$ and R is Fourier-diagonal with variance $r_{\boldsymbol{k}}$. White physical noise has $r_{\boldsymbol{k}} = \sigma_\varepsilon^2$. An N_0^2 real grid gives $k_0 = N_0/2-1$; for $N_0 = 32$ the observed square contains $31^2 = 961$ modes. Section D details the Nyquist convention and its baseline consequences.

The entry state. Fix an entry time s^* and write $a_{\boldsymbol{k}}, \sigma_{\boldsymbol{k}}$ for the forward process's per-mode signal and noise coefficients there. Given an observation resolving the mode set $\mathcal{O}$, build a fine-grid state X^* by

$$\widehat{X}^*_{\boldsymbol{k}} = \begin{cases} a_{\boldsymbol{k}}(s^*)\,\widehat{y}_{t,\boldsymbol{k}} + \sqrt{\sigma_{\boldsymbol{k}}(s^*)^2 - a_{\boldsymbol{k}}(s^*)^2 r_{\boldsymbol{k}}}\;\widehat{\xi}_{\boldsymbol{k}}, & \boldsymbol{k} \in \mathcal{O}, \\ \sigma_{\boldsymbol{k}}(s^*)\,\widehat{\xi}_{\boldsymbol{k}}, & \boldsymbol{k} \notin \mathcal{O}, \end{cases} \tag{4}$$

with fresh unit noise $\widehat{\xi}$. It is valid when $\sigma_{\boldsymbol{k}}^2 \geq a_{\boldsymbol{k}}^2 r_{\boldsymbol{k}}$ on $\mathcal{O}$: sensor and added noise then sum to variance $\sigma_{\boldsymbol{k}}^2$. At the trained entry, $a_{\boldsymbol{k}} = 1$ and $\sigma_{\boldsymbol{k}}^2 = r_{\boldsymbol{k}}$ on the observed band, so the added term vanishes; off-band, $a_{\boldsymbol{k}} = 0$. Thus one pastes y_t and fills the missing band with climatological noise. This is the per-mode matching principle used by DDRM (Kawar et al., 2022), incorporated here into the trained process.

Proposition 1 (Reverse restart is exact conditional sampling)**.** *Suppose the forward process satisfies (C1)–(C4), the observation follows equation 3, and the observation and entry noises are independent of the conditioning context c given ω_t. Suppose the entry time is matched to the sensor: $a_{\boldsymbol{k}}(s^*) = 1$ and $\sigma_{\boldsymbol{k}}(s^*)^2 = r_{\boldsymbol{k}}$ on the observed band, with $a_{\boldsymbol{k}}(s^*) = 0$ off it. Let X^* be the entry state equation 4. Then*

(a) (ω_t, c, X^*) has the same joint law as $(\omega_t, c, x(s^*))$ under the forward process;

(b) $p(\omega_t \mid X^*, c) = p(\omega_t \mid y_t, c)$, the Bayes posterior given the observation;

(c) exact reverse-time integration from (s^*, X^*) to $s = 0$ produces samples from $p(\omega_t \mid y_t, c)$.

Proof idea. Given (ω_t, c), equation 4 matches the Gaussian conditional law of $x(s^*)$. Its off-band coordinates are independent noise, while its observed coordinates equal y_t; hence conditioning on X^* is equivalent to conditioning on y_t. Anderson time reversal (Anderson, 1982; Song et al., 2021b) then gives the stated posterior. The full proof and assumptions (C1)–(C4) are in Section B.

The theorem and implementation have separate approximations. Proposition 1 assumes an exact score and exact destruction of the unobserved signal at entry. Under those premises, restart adds no conditioning approximation. Our experiments use a learned score and finite overshoot. On the production table the largest off-band coefficient is 2.48×10^{-4} ($\bar{\alpha} = 6.16 \times 10^{-8}$); summing the induced Gaussian mean shifts over all $56{,}160$ live unobserved coordinates gives expected joint Kullback–Leibler (KL) divergence 1.80×10^{-6} and, by Pinsker, total variation (TV) at most 9.5×10^{-4}. This is a climatology-averaged entry-law bound, not a uniform posterior-error bound (Section B). For noiseless data, entry occurs at s_{dead} and $g_{\boldsymbol{k}} = 0$ freezes the observed band. For noisy data, reverse integration first denoises that band on the ramp; repeatedly imposing raw y_t there would target a different distribution. Hard replacement is therefore a noiseless special case of matched restart.

One rule, both sides. We supervise and update mode $\boldsymbol{k}$ exactly when $g_{\boldsymbol{k}}(s) > 0$. Consequently, measured modes are frozen during the sweep and missing modes are frozen during the ramp. Masking on $\sigma_{\boldsymbol{k}} > 0$ would wrongly supervise the latter at zero signal-to-noise ratio (SNR). The complete region table and dead-band case appear in Section B.

Which operators qualify. The identification requires H to be a projection diagonal in the basis the forward process is diagonal in. The proof covers matched nested orthogonal projections with a feasible noise budget. Box averaging mixes weighting and aliasing and has no matched entry time. At each arm's best refined weight, the box observation costs $12.2/6.9/2.7/1.4\times$ per octave; unlike spectral guidance, it gains no benefit at $w = 4$ (Section G.3). General linear operators additionally require their singular values and transformed noise covariance, as in DDRM (Kawar et al., 2022); Proposition 1 does not prove that extension.

Measured against guidance. In a noiseless diagnostic, Figure 9 compares the residual $\|Hx - y^c\|/\|y^c\|$ along the reverse sweep against a DPS-style likelihood-gradient baseline (Chung et al., 2023); here the conditioning value y^c equals raw y because this arm is noiseless. Ours holds 2.8×10^{-8} at all 401 states. Every guidance trace takes one update and plateaus. At $w = 4$ its missing-band errors become $0.026/0.042/0.103/0.369$, equal to restart to three decimals, while its sampler-state residual is 8.0×10^{-3}.

The equality is algebraic. DPS updates its clean prediction by $x_0 \leftarrow x_0 + \eta H^\top (y - Hx_0)$ with $\eta = w/\sigma_\varepsilon^2$. Since spectral H is an orthogonal projection, $w = \sigma_\varepsilon^2 = 4$ makes this clean-state correction a projection. It does not make the full DPS trajectory a hard-projected one, as its nonzero residual shows. The experiment therefore supports a mechanism claim: DPS reproduces restart after selecting the unit effective step; scale-time restart obtains the frozen-coordinate update from its

dynamics and uses the known sensor variance only to select the entry time, not to tune an additional likelihood weight. The full sweep and the mismatched box-operator control are in Figure 9 and Section G.3.

5 Experimental Setup

5.1 Systems and protocol

Systems. We use 256^2 forced 2D turbulence (Kraichnan, 1967; Boffetta & Ecke, 2012) and surface quasi-geostrophic (SQG) flow (Held et al., 1995) on periodic domains. The first has nearly scale-independent turnover; the second has a measured scale-dependent turnover and tests interaction with the scale-indexed schedule. Both simulations are dealiased and stored through $k_R^{\mathrm{eff}} = 119$. Solver and spectrum details are in Sections C and G.5.

Observations and the cycled loop. The sensor returns the 32^2 spectral square $\|\boldsymbol{k}\|_\infty \le k_0 = 15$ every $\Delta t_{\mathrm{obs}} = 0.100 = 0.235\,\tau$. Each cycle forecasts the previous analysis with the FNO, assimilates one observation, and feeds the analysis back. All methods share this loop and differ only in the analysis. The interval correlations and square-versus-radial bookkeeping are in Sections D and G.2.

Metric. We report relative Fourier error per octave, $\|\widehat{x} - \widehat{\omega}\|/\|\widehat{\omega}\|$ restricted to each band, over $(15, 30]$, $(30, 60]$, $(60, 120]$ and $(120, 168]$. Normalized this way, 1 is the score of predicting zero. We use the whole-field norm, rather than the mean of octave ratios, for total error. Octave resolution is necessary because $E(k)$ spans four decades; Section G.2 gives the metric rationale and conventions.

5.2 Baselines

The principal learned baseline is the iterative-refinement (IR) cascade of Dhingra et al. (2026), trained on this dataset at parameter parity (4.17M versus our 4.11M parameters). It uses a discrete resolution ladder with learned coarse conditioning. Our three-rung implementation appears only as the controlled discretization ablation in Section G.1, not as an IR baseline. An *ensemble Kalman filter* (EnKF) (Evensen, 2003) with 20 members either shares our FNO or uses the data-generating solver (Section 7.3). Both retain documented, untuned covariance defaults. We also include observation-only EDSR (Lim et al., 2017), direct insertion, FNO free-run (Li et al., 2021), and spectral zero-padding. Geometry asymmetries and EnKF settings are documented in Section F.

5.3 Statistical protocol

Every method comparison is *paired*: arms share the start frame and the entry-noise seed, so a difference is measured within a start rather than across a population of them. We report paired t throughout and treat $|t| > 3$ as resolved.

The 25-cycle studies use $n = 40$ starts, strided eight frames apart within five test trajectories. They are paired samples but not 40 independent flows, and the tests do not model this clustering. The 198-cycle study has one start per trajectory ($n = 3$), so we report no paired statistics there. Calibration uses the octave-wise continuous ranked probability score (CRPS) (Gneiting & Raftery, 2007); protocol and calibration details are in Sections G.2 and G.10.

6 Results

6.1 Single-analysis fidelity

Table 1 reports per-octave error at a frozen snapshot, and Figure 2 shows the corresponding fields. On 2D we reach $0.262/0.447/0.631/0.855$ across the four octaves. Every learned baseline is at or above the zero-prediction score from the second missing octave: deterministic SR loses power, whereas IR retains uncorrelated power (Figure 8). On SQG our errors rise to

Table 1: Single-analysis relative error per octave, at a frozen snapshot of the cycled run. Lower is better; 1.000 is the score of predicting zero, so a method at or above it has stopped contributing in that band.

system	method	$(15, 30]$	$(30, 60]$	$(60, 120]$	$(120, 168]$
2D Kraichnan	**scale-time (ours)**	0.262	0.447	0.631	0.855
	IR cascade	0.762	0.985	1.099	1.118
	deterministic SR	0.866	0.988	1.001	1.002
	observation (zero-pad)	1.000	1.000	1.000	1.000
	FNO free-run			*diverged*	
	direct insertion			*diverged*	
SQG	**scale-time (ours)**	0.445	0.784	1.031	1.227
	IR cascade	0.685	0.968	1.081	1.080
	deterministic SR	0.847	0.968	0.998	1.001
	observation (zero-pad)	1.000	1.000	1.000	1.000
	FNO free-run			*diverged*	
	direct insertion			*diverged*	

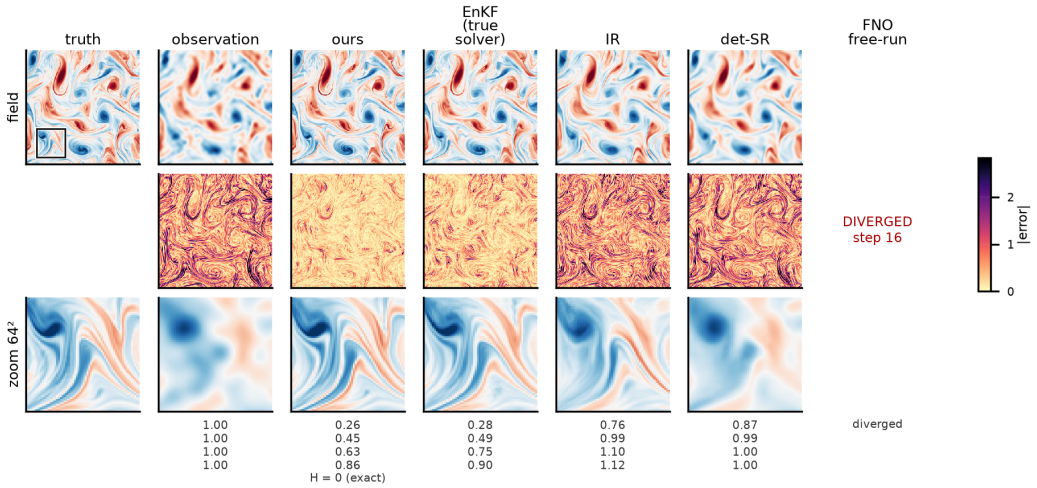

Figure 2: **Single 2D analysis.** Fields, absolute errors, and zooms at a frozen cycled snapshot. IR retains structured but misplaced fine scales; deterministic SR smooths them away. Per-octave errors are shown in each column.

$0.445/0.784/1.031/1.227$, crossing the zero-prediction score in the upper two octaves; full results are in Section G.5.

Why the baselines fail, quantitatively. Relative error in a band decomposes exactly, in the recovered power ratio P and the correlation C with truth, as

$$\text{err} = \sqrt{1 + P - 2\sqrt{P}\,C}, \tag{5}$$

which reproduces all reported errors to 10^{-6}. Deterministic SR has $P = 0.445/0.029/0.003/0.011$; its mean-squared-error (MSE) objective drives the conditional mean toward zero at unpredictable scales. IR retains power ($P = 0.83/0.54/0.28/0.28$) but its correlation reaches 0.073 and then -0.063; for $C \approx 0$, nonzero power makes error exceed one. Ours keeps $P \approx 1$ and $C = 0.98/0.92/0.83/0.69$. Thus spectral shrinkage and stable but uncorrelated generation are distinct failures, neither captured by a divergence count.

Probabilistic calibration. At $k_0 = 15$, spread/error is 0.70–1.00 across octaves against a computed calibrated-sampler reference of 0.956. Octave-wise CRPS is $0.010/0.017/0.041/0.122$, versus $0.114/0.191/0.443/0.966$ for the EnKF sharing our forecast. Calibration deteriorates sharply

for the sparse $k_0 = 7$ sensor; effective-sample-size estimation and the full diagnostics are in Section G.10.

6.2 CYCLED STABILITY

Ours completes 50 cycles on both systems with saturated rather than compounding error. Free-run and direct insertion diverge at cycles 16 and 23 on 2D and at 12 and 19 on SQG. Deterministic SR remains stable by approaching the zero-prediction line, showing why stability alone is insufficient. Figure 13 traces the 2D error by octave and cycle.

At cycle 197 of the descriptive $n = 3$ long run, total error is 0.310 for ours, 0.428 for IR, and 0.453 for zero-padding (Figure 12). Ours remains below one per octave except in the corner (1.06; Figure 16). The perfect-dynamics EnKF crosses total climatology at cycle 154 under untuned inflation and additive noise; this documents that configuration rather than the EnKF ceiling (Houtekamer & Mitchell, 2005).

7 ABLATIONS

7.1 THE CLOCK

The progressive clock of equation 2 is what makes off-design entry possible, but at the design resolution a shared clock improves the first two octaves by 17.4% and 16.0% ($t = 3.63, 4.26$), with no resolved difference above $k = 60$ (Table 4). Matching per-mode steps and sampler cost does not close the gap (Section G.1). A partial refinement sweep then improves progressive-only by $22.6/19.2/11.2/7.9\%$ (all $|t| > 3$), matching or exceeding the shared clock while retaining approximate off-design entry. Thus flexibility has a measured design-resolution cost, and refinement recovers it at additional sampling cost.

7.2 OFF-DESIGN ENTRY

A model trained with 32^2 observations receives a noiseless 96^2 observation ($k_0' = 47$) without retraining. Entering the progressive process at $k_c(s^*) = 47$ improves its first missing octave $9.4\times$ over design-resolution entry and $7.0\times$ over the shared clock (Figure 3). Shared clocks have no intermediate band-limited frame, while a dyadic IR ladder restricts 96^2 back to its 32^2 rung; both reproduce their design-resolution results to four decimals.

This gain is approximate. At zero overshoot the largest discarded signal coefficient is 0.6355; increasing overshoot improves marginal matching but removes more of the added observation, eventually recovering the 32^2 result. Noisy off-design entry violates the noise budget because the mid-sweep measured band has $\sigma_{\boldsymbol{k}} = 0$, and observations coarser than training are outside the fixed exempt subspace. Section G.6 gives the full curve and limiting cases.

7.3 WHAT THE LEARNED PRIOR COSTS: A PERFECT-DYNAMICS COMPARATOR

All learned arms share the FNO, confounding the analysis method with the forecast prior. We partially separate them using an EnKF whose forecast uses the data-generating solver. At cycle 25 over 40 paired starts, total error is 0.183 for ours and 0.185 for EnKF ($t = 0.49, p = 0.63$); EnKF is better in $(15, 30]$ ($t = -4.32$), and no octave resolves in our favor. This finite, untuned ensemble is not an upper bound. Its $55\,\mathrm{s}$ serial-CPU timing is not comparable to accelerator inference, and estimating the EnKF ceiling requires validation-tuned inflation and localization (Houtekamer & Mitchell, 2005). Further configuration and timing details are in Sections F and G.7 and Fig. 22.

7.4 SAMPLER COST

At converged operating points, IR costs $483\,\mathrm{ms}$ versus $793\,\mathrm{ms}$ for ours; recovered fine-scale power costs $1.64\times$ more. NFE counts are not comparable because IR evaluates lower-resolution rungs. Section G.7 gives convergence curves, per-call timings, and Figure 22.

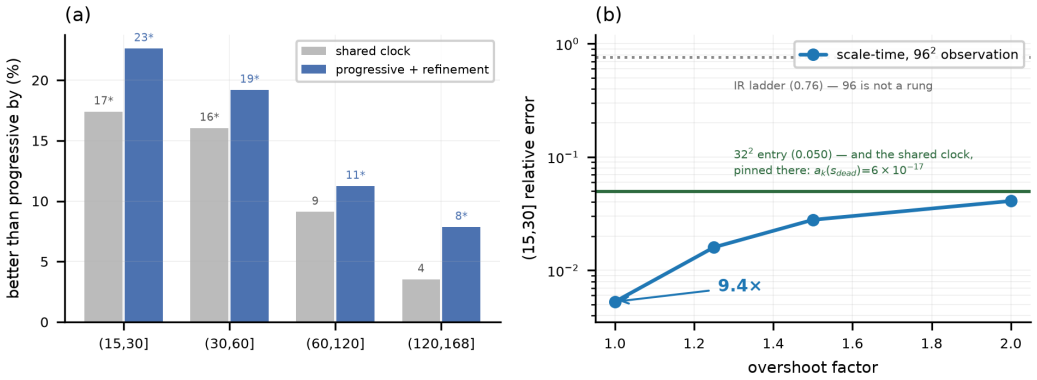

Figure 3: **Clock cost and off-design benefit.** (a) A shared clock wins at 32^2; refinement recovers the gap. (b) At 96^2, progressive entry improves the first missing octave $9.4\times$. Increasing overshoot improves marginal matching while discarding more observed information.

7.5 Observation noise

Sensor noise selects a matched ramp entry without a tuned likelihood weight. Across $\sigma_\varepsilon = 0.5$–4.511, s^* moves from 0.868 to 1 and observed-band error from 0.021 to 0.044. On $k \in (10, 15]$, assuming half the true noise costs $3.55\times$, whereas assuming $1.5\times$ too much does not hurt. Section G.8 gives the full sweep and band dependence.

8 Limitations

Exact restart requires a matched projection that is a reachable forward marginal and satisfies the noise budget. Spectral truncation qualifies; aliased box averages and point sensors do not. General linear operators need singular-value and noise-covariance treatments. We test two flows but one sensor family. On SQG, our error exceeds the zero-prediction score in the upper two octaves (Section G.5), leaving other sensors and higher-dimensional systems open. With the sparse $k_0 = 7$ observation, IR wins in the corner, 1.048 versus 1.179 ($t = -3.96$), because root-mean-square error rewards low power when neither method knows the phase (Section G.4). The cascade receives extra Nyquist modes and its SQG augmentation is untuned; the finite, untuned perfect-dynamics EnKF is neither an upper bound nor a speed comparator, and its long-run divergence is configuration-specific (Sections F, 7.3 and G.7). Forty paired starts span five trajectories and the long run has $n = 3$, precluding population-level uncertainty without trajectory-aware inference (Sections 5.3 and G.2). The clock ablation is noiseless, the $1.49\times$ speedup is analytic, the 2.8×10^{-8} residual tests frozen-coordinate replacement rather than score accuracy, and the finite endpoint only approximates exact killing despite its small joint-law bound. Sections G.3 and G.9 give full scope.

9 Conclusion

When an observation has the law of a reachable forward marginal, reverse restart can replace an inference-time likelihood correction. With an exact score and exact marginal match, it samples the Bayes posterior. Scale-time diffusion realizes this construction for spectral projections by destroying modes in scale order and matching sensor noise on a separate ramp; its finite endpoint approximates the exact killing premise with a quantified joint-law discrepancy. Across two chaotic dynamical systems, restart recovers information in the first three missing octaves and remains stable under cycling. At $w = \sigma_\varepsilon^2$, DPS matches its missing-band error because the clean-state correction becomes a projection; scale-time instead obtains frozen sampler coordinates from its dynamics. The guarantee remains limited to matched projections with a feasible noise budget. Extending it to general linear operators requires singular-value and transformed-noise treatments. These results support aligning a generative forward process with the sensor so that conditioning follows from the reverse dynamics.

AI USE STATEMENT

ChatGPT (GPT-5.6) was used solely for language editing, including correcting grammatical errors, improving wording, and identifying minor writing issues in the manuscript. It was not used to generate scientific results, perform experiments, develop code, derive mathematical results, or determine the scientific conclusions of this work. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This study uses synthetic fields generated by numerical simulations and does not involve human subjects, personal data, or sensitive attributes. The method generates unresolved fine scales conditioned on partial observations; such samples should not be presented as direct measurements or used in downstream scientific decisions without preserving their ensemble uncertainty and validating the sensor and operator assumptions. We report calibration, sparse-sensing failures, operator mismatch, and the limits of off-design entry to make these risks explicit. The authors have read and adhered to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The forward process, entry construction, and assumptions are specified in Sections 3 and 4, with a complete proof in Section B. Measured schedule constants, training details, baseline configurations, evaluation protocols, and calibration procedures appear in Sections C, E, F, G.2 and G.10. The supplementary repository contains the implementation, experiment scripts, configuration files, manifest-backed numerical records, figure and table generators, and consistency checks used for this paper. Paired comparisons use shared starts and random seeds as described in Section 5.3; limitations of the available trajectories and uncertainty analysis are stated in Section 8.

REFERENCES

Melissa Adrian, Daniel Sanz-Alonso, and Rebecca Willett. Data assimilation with machine learning surrogate models: A case study with FourCastNet. *Artificial Intelligence for the Earth Systems*, 4 (3):e240050, 2025. doi: 10.1175/AIES-D-24-0050.1.

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. doi: 10.1016/0304-4149(82)90051-5.

Aditya Sai Pranith Ayapilla, Kazuya Miyashita, Yuki Yasuda, and Ryo Onishi. Uncertainty-aware spatiotemporal super-resolution data assimilation with diffusion models, 2026. URL https://arxiv.org/abs/2604.21180. arXiv:2604.21180.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2208.09392.

Guido Boffetta and Robert E. Ecke. Two-dimensional turbulence. *Annual Review of Fluid Mechanics*, 44:427–451, 2012. doi: 10.1146/annurev-fluid-120710-101240.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.01392.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14367–14376, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Choi_ILVR_Conditioning_Method_for_Denoising_Diffusion_Probabilistic_Models_ICCV_2021_paper.html.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14687.

Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G. Dimakis, and Peyman Milanfar. Soft diffusion: Score matching with general corruptions. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=W98rebBxlQ.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Mrigank Dhingra, Ramchandran Muthukumar, Rebecca Willett, and Omer San. Iterative refinement diffusion for super-resolved data assimilation of multiscale physical systems, 2026. URL https://arxiv.org/abs/2608.14744. arXiv:2608.14744.

Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003. doi: 10.1007/s10236-003-0036-9.

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.

Isaac M. Held, Raymond T. Pierrehumbert, Stephen T. Garner, and Kyle L. Swanson. Surface quasi-geostrophic dynamics. *Journal of Fluid Mechanics*, 282:1–20, 1995. doi: 10.1017/S0022112095000012.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

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, 2022a. URL https://www.jmlr.org/papers/v23/21-0635.html.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022b. URL https://arxiv.org/abs/2204.03458.

Daniel Hodyss and Matthias Morzfeld. Using diffusion models to do data assimilation. *Monthly Weather Review*, 154(2):165–182, 2026. doi: 10.1175/MWR-D-25-0125.1.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.05557.

P. L. Houtekamer and Herschel L. Mitchell. Ensemble kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3269–3289, 2005. doi: 10.1256/qj.05.135.

Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter Dominik Dueben, and Torsten Hoefler. DiffDA: a diffusion model for weather-scale data assimilation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19798–19815, 2024. URL https://proceedings.mlr.press/v235/huang24h.html.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.00364.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2201.11793.

Robert H. Kraichnan. Inertial ranges in two-dimensional turbulence. *Physics of Fluids*, 10(7): 1417–1423, 1967. doi: 10.1063/1.1762301.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

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*, 2021. URL https://arxiv.org/abs/2010.08895.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. URL https://arxiv.org/abs/1707.02921.

Jae Hyun Lim, Nikola B. Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, Christopher Pal, Arash Vahdat, and Anima Anandkumar. Score-based diffusion models in function space. *Journal of Machine Learning Research*, 26(158):1–62, 2025. URL https://www.jmlr.org/papers/v26/23-1472.html.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lugmayr_RePaint_Inpainting_Using_Denoising_Diffusion_Probabilistic_Models_CVPR_2022_paper.html.

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*, 6:124, 2025. doi: 10.1038/s43247-025-02042-5.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2108.01073.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2102.09672.

Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models. *Journal of Machine Learning Research*, 25(414):1–52, 2024. URL https://www.jmlr.org/papers/v25/23-1271.html.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2206.13397.

François Rozet and Gilles Louppe. Score-based data assimilation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.10574.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. URL https://arxiv.org/abs/2104.07636.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://arxiv.org/abs/2010.02502.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://arxiv.org/abs/2011.13456.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.00490.

Eunbi Yoon, Won Chang, Donghan Kim, and Dae Wook Kim. Rethinking forward processes for score-based nonlinear data assimilation in high dimensions, 2026. URL https://arxiv.org/abs/2604.02889. arXiv:2604.02889.

Qinsheng Zhang, Molei Tao, and Yongxin Chen. gDDIM: Generalized denoising diffusion implicit models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2206.05564.

A Notation

Table 2 collects every symbol used in the main text, with the value it takes for the 2D system where that value is fixed by the dataset rather than chosen. Three conventions are worth restating because they are the ones most likely to be misread. First, ρ is deliberately overloaded: it denotes the lag-1 correlation of the data in Section 5.1 and the ramp's noise profile in (C2). The two never appear in the same expression, and we keep both because each is standard in its own literature. Second, $E(k)$ is the *per-mode* climatological variance $\langle|\widehat{\omega}_{\boldsymbol{k}}|^2\rangle$, never a shell-integrated spectrum; the two differ by the shell count $\sim 2\pi k$, i.e. by exactly one power of k, and every exponent we quote is per-mode. Third, sets defined by $\|\boldsymbol{k}\|_\infty$ are squares while $E(k)$ and $\tau_{\boldsymbol{k}}$ are radial, so a square-defined set summarized by a radial statistic mixes two populations (Section D).

B Proof of Proposition 1

The forward process and its constraints. The marginals equation 1 are subject to four constraints, which we state in full here because the proof uses each of them by name. **(C1) Clean start:** $a_{\boldsymbol{k}}(0) = 1$ and $\sigma_{\boldsymbol{k}}(0) = 0$ for every live mode. **(C2) Observed band preserved, then noised:** for $\|\boldsymbol{k}\|_\infty \le k_0$, $a_{\boldsymbol{k}} \equiv 1$ on all of $[0,1]$, with $\sigma^2_{\boldsymbol{k}} = 0$ for $s \le s_{\mathrm{dead}}$ and a nondecreasing white ramp $\sigma_{\boldsymbol{k}}(s)^2 = \rho\Big(\frac{s-s_{\mathrm{dead}}}{1-s_{\mathrm{dead}}}\Big)\,\sigma^2_{\max}$ thereafter. **(C3) Unobserved band destroyed by s_{dead}:** for $\|\boldsymbol{k}\|_\infty > k_0$, $a_{\boldsymbol{k}}(s) = 0$ and $\sigma_{\boldsymbol{k}}(s)^2 = E(k)$ for all $s \ge s_{\mathrm{dead}}$, giving pure noise at climatological variance, independent of $x(0)$, and *staying* that way, since the ramp does not touch it. **(C4) Monotone corruption:** $\mathrm{SNR}_{\boldsymbol{k}}(s) = a^2_{\boldsymbol{k}}E(k)/\sigma^2_{\boldsymbol{k}}$ is nonincreasing in s for every $\boldsymbol{k}$, understood weakly (infinite on the observed band below s_{dead}, identically zero on the unobserved band above it).

These marginals are realized by the linear Itô SDE

$$\mathrm{d}\widehat{x}_{\boldsymbol{k}} = \frac{\dot{a}_{\boldsymbol{k}}}{a_{\boldsymbol{k}}}\,\widehat{x}_{\boldsymbol{k}}\,\mathrm{d}s + g_{\boldsymbol{k}}(s)\,\mathrm{d}\widehat{W}_{\boldsymbol{k}}, \qquad g^2_{\boldsymbol{k}} = \frac{\mathrm{d}}{\mathrm{d}s}\sigma^2_{\boldsymbol{k}} - 2\frac{\dot{a}_{\boldsymbol{k}}}{a_{\boldsymbol{k}}}\sigma^2_{\boldsymbol{k}}. \tag{6}$$

On differentiable intervals where $a_{\boldsymbol{k}} > 0$, (C4) implies $g^2_{\boldsymbol{k}} \ge 0$ and hence a valid linear diffusion. If a schedule reaches $a_{\boldsymbol{k}} = 0$ at a finite endpoint, equation 6 is interpreted through its limiting Markov transition and the resulting data-independent Gaussian coordinate is then frozen; the quotient $\dot{a}/a$ need not remain finite at that endpoint. The proposition needs the Markov transition and the marginals in (C1)–(C4), rather than a nonsingular coefficient representation at the killing time. Markovianity licenses restarting from an intermediate state rather than only from the terminal marginal.

Table 2: Notation. Values are for the 2D Kraichnan system at 256^2 with 32^2 observations.

symbol	meaning	value	defined
Scale-time and the schedule			
s	scale-time	$[0, 1]$	Section 3.1
s_{dead}	sweep-completion time	0.8658	Section 3.2
s^*	entry time, matched to σ_ε	0.892	Definition 1 ff.
$k_c(s)$	cutoff schedule, strictly decreasing	$336 \to 7.5$	Eq. equation 9
κ	overshoot terminus, $k_0 2^{-mw}$	7.5	Section C
w, m	switch bandwidth (octaves), overshoot margin	0.25, 4	Section C
$\tau_{\boldsymbol{k}}(s)$	mode clock, 0 untouched $\to$ 1 destroyed	—	Eq. equation 2
$\bar{\alpha}$	scalar VP noise schedule on the mode clock	cosine	Section C
Φ	smooth switch, nondecreasing, $\to 0$ / $\to 1$	$\frac{1}{2}(1+\tanh)$	Eq. equation 2
Wavenumber geometry (∞-norm unless radial)			
k_0	observation Nyquist, $N_0/2 - 1$	15	Definition 1
k_R^{eff}	storage truncation	119	Section D
k_{top}	top live radius, $\sqrt{2}\, k_R^{\text{eff}}$	168	Section D
$\mathcal{O}$	resolved mode set of an observation	961 modes	Eq. equation 4
Forward process, per mode			
$a_{\boldsymbol{k}}(s)$	signal coefficient	$[0, 1]$	Eq. equation 1
$\sigma_{\boldsymbol{k}}(s)$	noise scale	≥ 0	Eq. equation 1
$V_{\boldsymbol{k}}$	total variance, $a^2 E(k) + \sigma^2$	—	Section C
$g_{\boldsymbol{k}}(s)$	diffusion coefficient; $g > 0$ iff diffusing	—	Eq. equation 6
$\text{SNR}_{\boldsymbol{k}}$	$a^2 E(k)/\sigma^2$, nonincreasing in s	—	(C4)
$z_{\boldsymbol{k}}$	whitened coordinate, $\widehat{x}_{\boldsymbol{k}}/\sqrt{E(k)}$	—	Section 3.1
$\tilde{\sigma}_{\boldsymbol{k}}$	effective noise level, $\sigma/(a\sqrt{E(k)})$	—	Section 3.4
X^*	entry state built from y_t	—	Eq. equation 4
Data, sensor and evaluation			
ω_t	fine vorticity field at time t	256^2	Section 5.1
ω_t^a	analysis state	—	Section 5.1
y_t	observation	32^2	Eq. equation 3
c_t	FNO forecast used as conditioning	—	Section 5.1
H	spectral truncation operator	—	Eq. equation 3
R, $r_{\boldsymbol{k}}$	observation-noise covariance, per-mode variance	—	Eq. equation 3
σ_ε	sensor noise	2.0	Section 7.5
$\sigma_{\max}$	design ceiling on supported sensor noise	$\sqrt{20.35}$	Section C
$E(k)$	*per-mode* climatological variance	$\alpha = 2.47$	Section C
$E_{\min}$	dead-mode floor, $\epsilon \max_{k>0} E(k)$	$\epsilon = 10^{-12}$	Section C
M	reverse steps per analysis	100	Section 7.4
P, C	recovered power ratio, correlation with truth	—	Eq. equation 5
ρ	lag-1 correlation per octave (data), or	$0.743/0.344/0.048$	Section 5.1
	the ramp's noise profile (schedule)	—	(C2)

(a) The entry state is a forward marginal. Condition on $(x(0), c) = (\omega_t, c)$. Under the forward process at s^*, the observed-band coordinates are $\widehat{x}_{\boldsymbol{k}}(0) + \sigma_{\boldsymbol{k}}(s^*)\widehat{\xi}_{\boldsymbol{k}}$ with $\sigma_{\boldsymbol{k}}(s^*)^2 = r_{\boldsymbol{k}}$ by the matching of the entry time, and the unobserved-band coordinates are $\sigma_{\boldsymbol{k}}(s^*)\widehat{\xi}_{\boldsymbol{k}}$—pure noise independent of $x(0)$, by (C3). The entry state equation 4 has, conditionally on $(x(0), c)$, identical law: its observed band is $\widehat{y}_{t,\boldsymbol{k}} = \widehat{x}_{\boldsymbol{k}}(0) + \widehat{\varepsilon}_{\boldsymbol{k}}$ with $\widehat{\varepsilon}_{\boldsymbol{k}} \sim \mathcal{CN}(0, r_{\boldsymbol{k}})$, and its unobserved band is fresh noise at the matching variance. Equality of conditional laws given $(x(0), c)$, together with their shared marginal, gives equality of the joints. □

(b) Conditioning on the state equals conditioning on the observation. Write $X^* = (X^*_{\text{obs}}, X^*_{\text{un}})$. By construction X^*_{un} is drawn independently of $(x(0), c, X^*_{\text{obs}})$, so it is uninformative and

$$p\big(x(0) \mid X^*, c\big) = p\big(x(0) \mid X^*_{\text{obs}}, c\big) \;\propto\; p_{\text{data}}\big(x(0) \mid c\big)\, \mathcal{N}\big(y_t;\, Hx(0),\, R\big),$$

using $X^*_{\text{obs}} = y_t$ as a value and the Gaussian conditional from (a). The right-hand side is the Bayes posterior for the observation model equation 3. This is the step the construction exists to

make available: the unobserved coordinates of the entry state carry no information about $x(0)$, so conditioning on the full fine-grid state is the same as conditioning on the observation alone. □

(c) Reverse integration samples it. By (a), $(x(0), c, X^*) \overset{d}{=} (x(0), c, x(s^*))$. For a Markov forward process, Anderson's time-reversal (Anderson, 1982; Song et al., 2021b) states that the reverse SDE driven by the exact score $\nabla_x \log p_s(x \mid c)$ generates the time-reversed process; run from the fixed initial condition $x(s^*) = X^*$, its time-zero law is $q(x(0) \mid x(s^*) = X^*, c)$. Combining with (b) gives samples from $p(x(0) \mid y_t, c)$. The same statement holds for the discretized ancestral chain with the exact posterior transition kernels, and DDIM-type samplers inherit it in the appropriate limit. ■

Where the overshoot enters. Hypothesis $a_{\boldsymbol{k}}(s^*) = 0$ off the observed band is not free, and it is where the schedule's design shows up in the proof. Under the implemented two-segment schedule, the production table instead gives $\max_{\boldsymbol{k}} a_{\boldsymbol{k}}(s_{\mathrm{dead}}) = 2.48 \times 10^{-4}$, or $\max \bar{\alpha} = 6.16 \times 10^{-8}$, over the $56{,}160$ live unobserved coordinates. The implementation therefore does not literally satisfy (C3) or the proposition's premise.

The resulting joint-law discrepancy can be bounded rather than inferred from the largest coefficient alone. Conditional on $x(0)$, the ideal and implemented entry laws are Fourier-diagonal Gaussians with the same covariance; their off-band means differ by $a_{\boldsymbol{k}}(s_{\mathrm{dead}})\widehat{x}_{\boldsymbol{k}}(0)$. Averaging the conditional KL over the climatology gives, in the real-field Fourier convention,

$$\mathbb{E}\,\mathrm{KL}_{\boldsymbol{k}} = \frac{a_{\boldsymbol{k}}(s_{\mathrm{dead}})^2 E(k)}{2\sigma_{\boldsymbol{k}}(s_{\mathrm{dead}})^2} = \frac{a_{\boldsymbol{k}}(s_{\mathrm{dead}})^2}{2\{1 - a_{\boldsymbol{k}}(s_{\mathrm{dead}})^2\}}, \tag{7}$$

where the second equality uses the implemented $\sigma_{\boldsymbol{k}}^2 = E(k)(1 - a_{\boldsymbol{k}}^2)$. Summing the production table yields

$$\sum_{\boldsymbol{k}} \mathbb{E}\,\mathrm{KL}_{\boldsymbol{k}} = 1.80 \times 10^{-6}, \qquad \mathrm{TV} \le \sqrt{\tfrac{1}{2} \sum_{\boldsymbol{k}} \mathbb{E}\,\mathrm{KL}_{\boldsymbol{k}}} = 9.5 \times 10^{-4}. \tag{8}$$

Because the clean-field marginal is shared, this is a bound between the two joint entry laws after averaging over the climatology. It is not uniform in $x(0)$ and does not bound posterior error from the learned score. The KL is concentrated near the boundary: radial bin 16 contributes 61.5%, bins 16–17 contribute 86.8%, and bins 16–20 contribute 99.3%.

Terminating the sweep at $k_c(s_{\mathrm{dead}}) = k_0$ would make it *unsatisfiable*: modes just above k_0 would sit at clock argument ≈ 0, hence $\tau \approx \frac{1}{2}$ and $a_{\boldsymbol{k}} \approx 0.7$ for every choice of w. This is an order-one violation concentrated on the first unobserved octave, which is precisely the band super-resolution is judged on. Under a single-segment schedule that sweeps and ramps simultaneously, the hypothesis instead fails at order one on the still-untouched band between k_0 and $k_c(s^*)$, and no margin repairs it. This is the formal content of the claim in Section 3.2 that there must be two segments.

The observed band is conditioning, not state. A consequence worth stating plainly: where $a_{\boldsymbol{k}} \equiv 1$ and $\sigma_{\boldsymbol{k}} \equiv 0$, the marginal p_s is supported on the affine slice fixed by the coarse field. It has no density with respect to Lebesgue measure on the full space, and a score with respect to the frozen coordinates is not merely degenerate but undefined. The honest description of the model is therefore a *conditional* diffusion over the unobserved band, with the observed band acting as a conditioning input that happens to be stored inside the state tensor. Stating it this way is what keeps the standard machinery available: the blurring-diffusion equivalence of Hoogeboom & Salimans (2023) and the variational bound that comes with it apply verbatim to the unobserved-band subsystem, whose terminal marginal at s_{dead} is the fixed, data-independent prior $\bigotimes_{\boldsymbol{k}} \mathcal{CN}(0, E(k))$. The trained objective is conditional denoising score matching for $p(x_{\mathrm{un}} \mid x_{\mathrm{obs}}, c)$, and the bound is a bound on that conditional likelihood.

The mask rule, and the band-limited case. The single rule *supervise and update mode $\boldsymbol{k}$ at s iff $g_{\boldsymbol{k}}(s) > 0$* follows from equation 6: modes need score supervision exactly where they diffuse. On a band-limited dataset it acquires a third frozen region: the dead band $E(k) \le E_{\min}$, where $\sigma_{\boldsymbol{k}} \equiv 0$ forces $g_{\boldsymbol{k}} \equiv 0$ on both segments. Thus, the check becomes six cells rather than four. The rule itself needs no amendment, but the implementation does need an explicit $E(k) > E_{\min}$ guard

applied *before* the preconditioning coefficients are formed: those coefficients are $0/0$ rather than merely zero on the dead band, and a mask evaluated afterwards inherits their NaNs.

C SCHEDULE CONSTRUCTION AND MEASURED CONSTANTS

The cutoff and the mode clock. The sweep is geometric: it is linear in $\log k$ and allocates equal scale-time per octave, matching the self-similar structure of the enstrophy cascade:

$$k_c(s) = \left(k_{\text{top}} 2^{mw}\right)^{1-s/s_{\text{dead}}} \kappa^{\,s/s_{\text{dead}}}, \qquad s \in [0, s_{\text{dead}}], \tag{9}$$

parked at κ for $s \geq s_{\text{dead}}$, feeding the mode clock equation 2. A linear-in-k schedule over-resolves the top octave; the geometric choice allocates equal scale-time per octave of wavenumber. Regularity is only C^0: the geometric sweep meets the parked segment with a kink at s_{dead}, where $\mathrm{d}\log k_c/\mathrm{d}s$ jumps to zero. This is harmless because the Itô SDE equation 6 requires merely measurable, locally bounded coefficients. We do not smooth the join.

Why the sweep starts at 336 and not at 119. Constraint (C1) requires $a_{\boldsymbol{k}}(0) = 1$ for *every* live mode, and by equation 2 a mode begins its clock when $k_c = k\,2^{mw}$. Because the storage truncation is an ∞-norm square, live modes reach radius $k_{\text{top}} = \sqrt{2}\,k_R^{\text{eff}} = 168$, not 119. Setting $k_c(0) = 119$ would give a mode at radius 168 the clock argument $\log_2(\sqrt{2})/w = 2$ at $s = 0$, hence $\tau \approx 0.98$ and $a \approx 0$: born destroyed, violating (C1) on 22% of the live set. Taking $k_c(0) = k_{\text{top}} 2^{mw} = 336$ removes it. The failure is silent, with no crash or test failure unless (C1) is checked explicitly, and its symptom is a permanently corrupted corner band, which is why we state it as a constraint rather than a parameter choice.

The longer sweep wastes nothing. The span is $\log_2(336/7.5) = 5.49$ octaves rather than the $\log_2(119/7.5) = 3.99$ one would write down naively, but the radius-168 modes begin transitioning immediately at $s = 0$, so no stretch is idle. What the extra 1.5 octaves buy is the corner region: sparse in variance, 22% of the live modes, and swept correctly rather than clipped.

Measured constants. Table 3 lists the constants this construction depends on. Every one is measured on the production dataset rather than assumed, and two are counterintuitive enough to flag. $\max_{k>0} E(k)$ sits at $k = 2$, at the forcing-scale turnover and *below* the power law; extrapolating the fitted law inward underestimates it by $\sim 2.6\times$. The zero mode $\boldsymbol{k} = \mathbf{0}$ must be excluded from every extremum and from the live set: $E(0) \approx 3.7 \times 10^{-15}$ is roundoff, since vorticity on a periodic domain has zero spatial mean by construction, and it is the argmin of any naive scan.

Figure 4 follows one reverse restart in field and phase, while Figure 5 shows the step-shaped forward marginals in spectral power. Together they expose the two properties used by the analysis: frozen observed coordinates and scale-ordered generation.

Spectrum matching, and what it does not do. Setting $\sigma_{\boldsymbol{k}}^2 = E(k)(1 - \bar{\alpha})$ rather than a flat σ^2 equalizes every mode's SNR trajectory exactly, whatever the shape of $E(k)$. Under the EDM preconditioning of Section E, the loss weight is $\lambda = V/\sigma^2$ and the output scale is $c^{\text{out}} = \sigma/\sqrt{V}$, so $\lambda\,(c^{\text{out}})^2 = 1$ identically, for *any* $(a, \sigma, E(k))$. An untrained network therefore has weighted residual exactly 1 per supervised mode under any schedule whatever, and the per-mode learning signal is renormalized identically whether σ was spectrum-matched or flat. **Spectrum matching cannot act through the loss.**

The trained control agrees. An otherwise identical model with isotropic σ loses by only $7.7/8.1/4.5/3.5\%$ per octave at $\eta = 0$ and less at $\eta = 1$, with *no* high-k droop. Under unconditioned sampling, the difference is not significant in any band.

Spectrum matching is nevertheless structurally required for three reasons. (i) (C3) requires $\sigma_{\boldsymbol{k}}(s_{\text{dead}})^2 = E(k)$ exactly, as that is what makes the unobserved band of the entry state a legitimate frame; a flat schedule violates it by up to four orders of magnitude at the top of the band. (ii) The composed coefficients stay finite as $a_{\boldsymbol{k}} \to 0$ because $V \to \sigma^2 > 0$, so no x_0 clamp is needed. The discrete-ladder comparison requires one whose value moves its error over four orders of magnitude, with no principled setting. (iii) A sampler started from a prior wrong by a factor ρ spends reverse steps correcting it; measured on the ladder, whose rungs terminate at $\mathcal{N}(0, I)$ while

Table 3: Measured constants, 2D Kraichnan at 256^2. Values are measured on the production dataset; those marked *derived* follow from the others.

quantity	value	note
k_0	15	not 16; the -16 slot is self-conjugate
k_R^{eff}	119	solver dealiased at 127; leaves margin below the boundary pileup
k_{top}	168	derived, $\sqrt{2}\,k_R^{\text{eff}}$; corner modes beyond radius 119 are live
κ	7.5	derived, $k_0 2^{-mw}$
$k_c(0)$	336	derived, $k_{\text{top}} 2^{mw}$; *not* k_R^{eff}
m, w	4, 0.25	$mw = 1$ octave
$\max_{k>0} E(k)$	3.72×10^3	at $k = 2$, at the turnover, not on the power law
$E(k_{\text{top}})$	2.72×10^{-5}	faintest live signal; $\epsilon = 10^{-8}$ would kill it
ϵ	10^{-12}	$E_{\min} = \epsilon \max_{k>0} E(k)$; near the geometric midpoint of a six-decade window
α	2.47	per-mode $E(k)$ exponent on $[16, 64]$; measured, never assumed
σ_{data}	2.349	ω_{rms}
Δt_{obs}	0.100	$= 0.235\,\tau$, with $\tau = 1/\omega_{\text{rms}} = 0.426$
lag-1 ρ	0.743/0.344/0.048	per octave, variance-weighted
fresh fraction	58.0%	variance-weighted, unobserved band

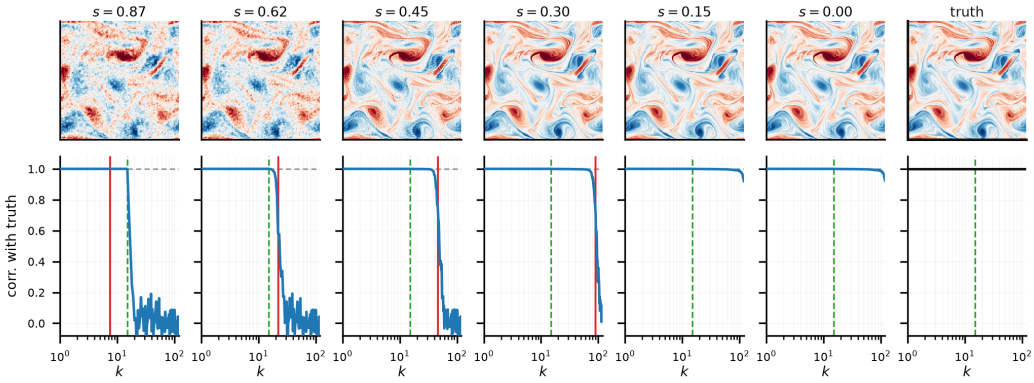

Figure 4: **Assimilation as a reverse restart.** One analysis, frame by frame: the reverse process starts from the observation at s^* (left) and integrates to $s = 0$ (right), with truth for reference. Top, the field; bottom, its per-shell correlation with truth. The green dashed line is $k_0 = 15$ and the red line is the cutoff $k_c(s)$. Three things are visible rather than asserted. The leftmost panel *is* the entry state—the observation, with climatological noise wherever the sensor is silent, not a noisy initialization to be cleaned up. Below k_0 the correlation is pinned at 1 in every panel, because $g_{\boldsymbol{k}}(s) = 0$ there and those coordinates are frozen by the dynamics. The correlation front advances in lockstep with the red cutoff: generation is ordered in scale rather than simultaneous. We plot correlation rather than $E(k)$ because spectrum matching makes the entry state's amplitude correct at every wavenumber from the start, so all six spectra would lie on truth and show nothing—what the reverse process supplies is phase. Relative error falls $0.571 \to 0.033$ across the panels.

the true per-mode variance falls $238\times$ across its band, the residual excess halves per NFE doubling with no plateau.

Schedule and clock visualizations. Figure 6 decomposes the cutoff, clocks, signal and noise coefficients, and active set. Figure 7 then contrasts the intermediate marginals produced by progressive and shared clocks.

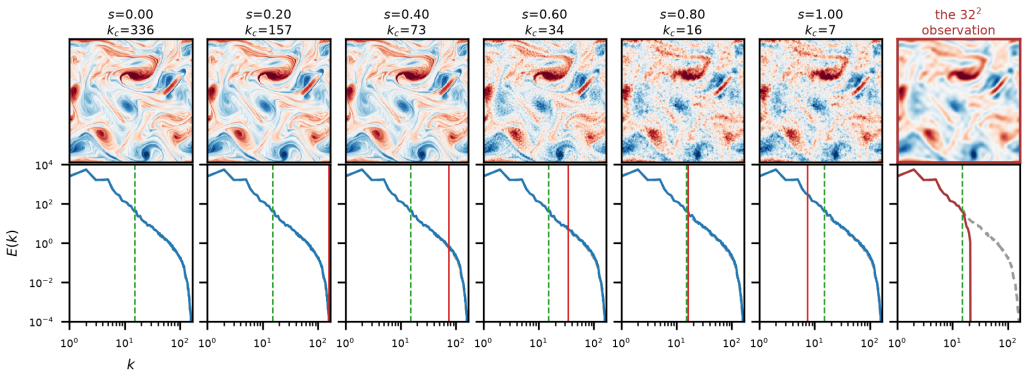

Figure 5: **The same process in $E(k)$.** Each frame is a band-limited field plus noise, forming the step-shaped marginal that makes a spectral observation a frame of the process rather than a constraint on one. A shared clock produces no such step at any interior s (Figure 7).

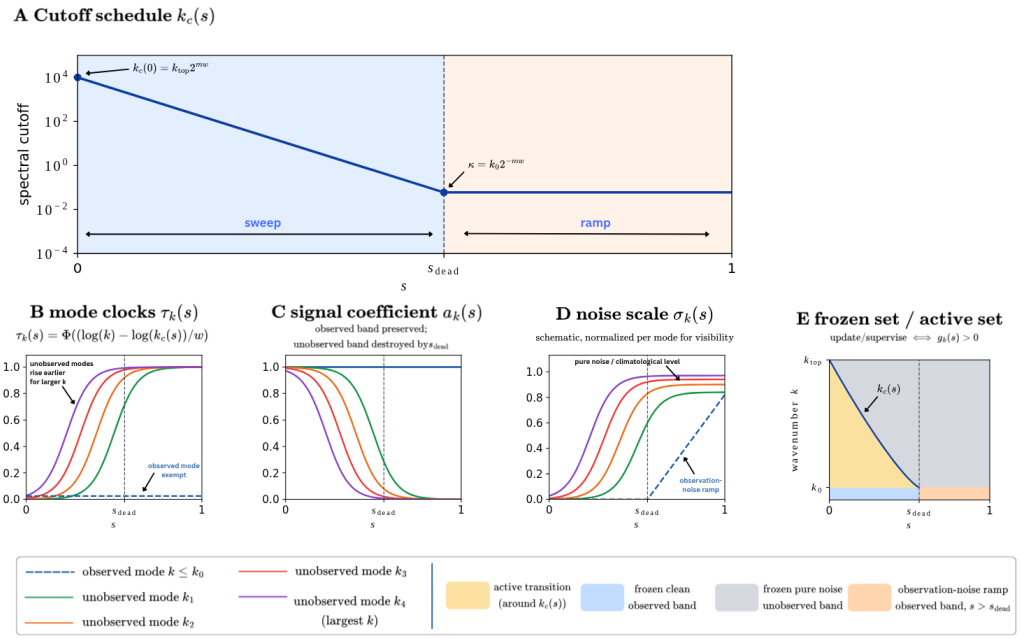

Figure 6: **Schedule anatomy.** (a) the cutoff $k_c(s)$, descending geometrically from 336 to $\kappa = 7.5$ over the sweep and parked through the ramp, with k_0 and k_{top} marked. (b) mode clocks $\tau_{\boldsymbol{k}}(s)$ for representative k: high-k clocks run early, low-k late, and each is a shifted copy of the same curve—the progressive clock. (c) signal coefficients $a_{\boldsymbol{k}}$, becoming numerically negligible before s_{dead} because the sweep overshoots to κ. (d) noise scales $\sigma_{\boldsymbol{k}}$, normalized per mode for visibility (schematic). (e) the active set: which modes carry $g_{\boldsymbol{k}}(s) > 0$, hence are supervised and updated, at each s. The two frozen regions—the observed band on the sweep and the unobserved band on the ramp—form the quadrant table of Section 4 drawn against scale-time.

D SQUARE VERSUS RADIAL GEOMETRY

Two families of sets appear in this construction and they do not nest. The exempt set $\{\|\boldsymbol{k}\|_\infty \le k_0\}$ and the storage truncation $\{\|\boldsymbol{k}\|_\infty \le k_R^{\text{eff}}\}$ are ∞-norm *squares*, because each is what a square grid resolves. The mode clock $\tau_{\boldsymbol{k}}$ and the spectrum $E(k)$ are *radial*, because each is a function of $|\boldsymbol{k}|$. Squares at both ends, radial in the middle. Three consequences follow, and all three are visible in reported numbers.

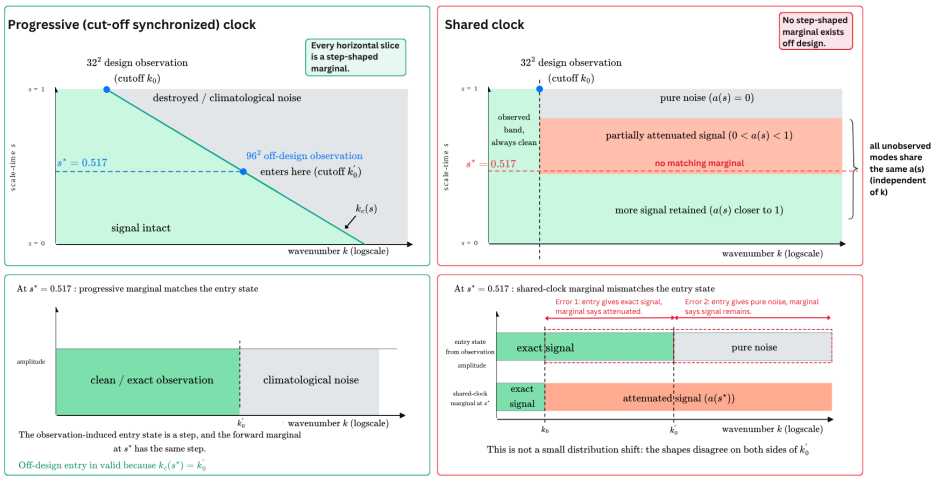

Figure 7: **Why only a progressive clock can enter off design.** Left: under a progressive clock every intermediate marginal is approximately step-shaped—intact below $k_c(s)$, destroyed well above it, and separated by a finite transition band. Thus, a 96^2 observation ($k_0' = 47$) can be used at $s^* = 0.517$. Under a shared clock every unobserved mode is partially corrupted by the same factor at every interior s, so the only band-limited frame is at s_{dead}, where $a_{\boldsymbol{k}} = 6.1 \times 10^{-17}$ annihilates anything pasted into a wider square. Right: the resulting two-error decomposition. Entering too early gives unattenuated signal where the marginal says attenuated; entering too late gives pure noise where the marginal says signal remains. The shared clock's failure is structural, not a tuning deficit. The progressive entry is useful but approximate; its measured exactness–utility curve appears in Section 7.2.

Corner modes are live. Live modes reach radius $k_{\text{top}} = \sqrt{2}\, k_R^{\text{eff}} = 168$, so 12,652 modes (22.1% of the live set of 57,121) lie beyond radius 119 in the corners of the storage square. A radial dead test at 119 would silently discard all of them. The dead test is therefore $\|\boldsymbol{k}\|_\infty > k_R^{\text{eff}}$, in the ∞-norm, and the sweep must start at 336 rather than 119 for the same reason (Section C).

Corruption is anisotropic near the observation boundary. The exempt square and the radial clock disagree on the annulus $k_0 < |\boldsymbol{k}| \le \sqrt{2}k_0$, corresponding to radii 15 to 21.2 at $k_0 = 15$. An axis-aligned mode at radius 16 is unobserved and swept; a corner mode at radius 21 is exempt forever. Membership in the exempt square decides a mode's fate *first*, and the radial clock applies only outside it. This is well defined and requires no change to the construction, but it must be stated because $\tau_{\boldsymbol{k}}$ alone does not determine fate.

Two annulus counts, both correct. The boundary annulus inside the exempt square has two valid sizes, and they answer different questions:

count	definition	the population at risk in
252	true radius $> k_0$	any *fate* test (exempt, dead, supervised), since those are ∞-norm squares and a radial test would misclassify
212	bin index $> k_0$	any *magnitude* lookup ($E(k)$, V, the whitening target), since those read the radial bin

The relationship is strict containment, $212 \subset 252$, and the 40-mode difference is exactly the radii in $[15, 15.5)$, which round down to bin 15. The rule behind it is the standing one: fate is two-dimensional, magnitude is radial. Throughout the main text we define the exempt set by the ∞-norm square and report the annulus *by radius* (252); the bin convention appears only where shell-averaged spectra are discussed, which is where binning is the operative object.

Shell averaging over a partially live shell. $E(k)$ is shell-averaged over *live* modes only, out to radius 168. Averaging a partially live shell over all of its modes understates $E(k)$ by the live fraction,

roughly $10\times$ at the outer radii. For the same reason every spectral diagnostic scans the *stored* band, to $k_R^{\text{eff}} = 119$, never to the dealias cutoff at 127: above 119 the array holds float32 quantization of an exactly-zero band ($\sim 10^{-15}$), and a diagnostic whose range is set from the dealias cutoff compares real spectrum against roundoff.

E TRAINING DETAILS

The objective. With the effective noise level $\tilde{\sigma}_{\boldsymbol{k}}(s) = \sigma_{\boldsymbol{k}}(s)/(a_{\boldsymbol{k}}(s)\sqrt{E(k)})$, the assembled denoiser is

$$\widehat{D}_{\boldsymbol{k}}(x_s, s, c) = c^{\text{skip}}_{\boldsymbol{k}}\, \widehat{z}_{s,\boldsymbol{k}} + c^{\text{out}}_{\boldsymbol{k}} \left[\text{FFT}\, F_\theta(x_s/\sigma_{\text{data}}, s, c)\right]_{\boldsymbol{k}}, \qquad \widehat{z}_{s,\boldsymbol{k}} := \frac{\widehat{x}_{s,\boldsymbol{k}}}{a_{\boldsymbol{k}}(s)\sqrt{E(k)}}, \tag{10}$$

with $c^{\text{skip}}_{\boldsymbol{k}} = (1+\tilde{\sigma}^2_{\boldsymbol{k}})^{-1}$ and $c^{\text{out}}_{\boldsymbol{k}} = \tilde{\sigma}_{\boldsymbol{k}}(1+\tilde{\sigma}^2_{\boldsymbol{k}})^{-1/2}$, trained against the clean whitened state under

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\, s\sim p(s),\, \xi} \sum_{\boldsymbol{k}:\, g_{\boldsymbol{k}}(s)>0} \lambda(\tilde{\sigma}_{\boldsymbol{k}}) \left|\widehat{D}_{\boldsymbol{k}}(x_s, s, c) - z_{\boldsymbol{k}}(0)\right|^2, \qquad \lambda(\tilde{\sigma}) = \frac{1+\tilde{\sigma}^2}{\tilde{\sigma}^2}. \tag{11}$$

EDM's scalar noise-conditioning input is subsumed: the entire profile $\tilde{\sigma}_\cdot(s)$ is a deterministic function of s, so conditioning on s carries the same information. As $a_{\boldsymbol{k}} \to 0$ at the end of a mode's sweep, $\tilde{\sigma}_{\boldsymbol{k}} \to \infty$, $c^{\text{skip}} \to 0$ and $c^{\text{out}} \to 1$: the parameterization degrades gracefully into prediction-from-noise with no special-casing.

Compose the coefficients algebraically. Both $\widehat{z}_{s,\boldsymbol{k}}$ and $\tilde{\sigma}_{\boldsymbol{k}}$ are individually singular as $a_{\boldsymbol{k}} \to 0$, while every quantity actually needed is finite. With $V_{\boldsymbol{k}} = a^2_{\boldsymbol{k}} E(k) + \sigma^2_{\boldsymbol{k}}$:

quantity	composed form	never compute as
skip term	$a\sqrt{E(k)}\,/\,V \cdot \widehat{x}_s$	$c^{\text{skip}} z$
output scale	$\sigma/\sqrt{V}$	$\tilde{\sigma}(1+\tilde{\sigma}^2)^{-1/2}$
loss weight	V/σ^2	$(1+\tilde{\sigma}^2)/\tilde{\sigma}^2$
target	$\widehat{x}_0/\sqrt{E(k)}$	—

The right-hand column is the form that requires a clamp whose value has no principled setting; the middle column needs none, because $V \to \sigma^2 > 0$ wherever $a \to 0$. This is why spectrum matching matters for the *sampler* as well as for (C3): a schedule whose terminal variance is not $E(k)$ has no such floor.

Architecture and optimization. The denoiser is a U-Net with $4{,}105{,}953$ parameters: base width 48, one residual block per level, group normalization, circular padding throughout (the domain is doubly periodic, so zero padding would impose a boundary the physics does not have), and attention at the bottleneck only. It takes two channels, the noisy state and the forecast context, and is conditioned on s and $\log k_c(s)$ through a sinusoidal embedding. Training uses Adam at 2×10^{-4} with 5000 warmup steps, batch size 16, gradient clipping at 1.0, and an EMA decay of 0.9999; the reported checkpoints are the EMA weights at $100{,}000$ steps.

Per-region loss is logged from the first step. Spectrum matching equalizes SNR per mode, which has a consequence worth watching rather than discovering late: the $12{,}652$ corner modes beyond radius 119 are 22% of the live set but carry on the order of 1% of unobserved-band variance, so under equation 11 they also receive on the order of 22% of the loss. Relatedly, a mode at radius k begins its clock when $k_c = k\, 2^{mw}$, so the first $\approx 9\%$ of the sweep in $\log k_c$ transitions corner modes *exclusively*. Neither is an error and we did not correct either; we logged exempt, cascade and corner loss separately from the first training step so that any correction would be a recorded decision.

Exposure bias, measured rather than assumed. The context is the teacher-forced forecast $F_\phi(\omega_t)$ during training but $F_\phi(\omega^a_{t-1})$ at inference, so the model meets a context distribution at test time that it did not train on. We quantified the gap rather than assuming it benign: at $k_0 = 15$ the inference-time context error sits at the $92/91/83/77$th percentile of the training support across the

four octaves. This lies inside the support but in its upper tail, and increasingly so with wavenumber. At $k_0 = 7$ the same measurement reads $95/98/97/94/98$, i.e. essentially at the edge, which is the diagnosis for that configuration's weaker performance (Section G.4).

The forecaster. The conditioning forecast is a Fourier neural operator (Li et al., 2021) trained on this dataset. It has $58{,}983{,}665$ parameters, 16 channels, $k_{\max} = 120$, and is selected on the rollout diagnostic rather than on one-step loss.

Its one-step error per octave is $0.034/0.057/0.145/0.456$. We report that beside *two* baselines, because the obvious one is misleading: persistence scores $0.753/1.137/1.373/1.415$, exceeding 1.0 in the upper three octaves, so "beats persistence" there is a bar cleared by predicting zero. The meaningful reference is the optimal linear predictor $\rho\, x_{\mathrm{prev}}$, at $0.697/0.935/0.998/1.000$; the forecaster is far inside it in every band.

The forecast *spectrum* matters as much as its error, for the reason Section 6.1 makes quantitative: an MSE-trained forecaster predicts the conditional mean, which for an unpredictable mode is ≈ 0, so it is systematically smooth. Measured, its recovered $E(k)$ ratio is $0.999/0.997/0.979/0.774$, which is essentially exact through $k = 120$ and drooping only in the corner. A conditioning input that is wrong in a structured way is worse than one that is uninformative, so this was checked before conditioning was wired in.

F Baseline configurations

Iterative refinement. We report the model of Dhingra et al. (2026), trained on this dataset at parameter parity ($4{,}171{,}537$ against our $4{,}105{,}953$, 1.6% apart), with its corrector and shared forecaster checkpoints used as released. It survives all 198 cycles of the long-horizon study. Our in-repository three-rung implementation is used only as a *discretization* ablation against continuous scale-time, with architecture and preconditioning held fixed; it is never reported as an IR baseline.

Iterative refinement: the two geometry asymmetries. Its observation is not quite ours, in its favor on both counts. Its spectral restriction retains the 32^2 grid's Nyquist line at $|\boldsymbol{k}| = 16$ and lifts it back to 256^2, where it lands outside our exempt square: 96 extra modes carrying 0.462% of observed-band energy. At the sparse $k_0 = 7$ setting its 16^2 rung sees 272 modes against our 224. We did not re-truncate either, because its corrector was trained on fields carrying those modes and removing them at inference would be a distribution shift we introduced. Inside our square the two observations agree to 1.9×10^{-7} relative, so the amplitude convention is identical and the difference is genuinely the extra modes. Both asymmetries are pinned by tests so they cannot drift.

Ensemble Kalman filter. Two EnKF variants appear. The first shares our FNO forecaster, so that it differs from our method only in the analysis step. The second, the perfect-dynamics comparator of Section 7.3, forecasts with the solver that generated the data at 384^2 with the production viscosity. Both use 20 members, inflation 1.05 and additive noise $\sigma_{\mathrm{add}} = 0.20$, following the IR configuration of Dhingra et al. (2026) and left *untuned*. This documents one reproducible configuration; it does not establish a fair optimum or an upper bound. Inflation and localization (Houtekamer & Mitchell, 2005) should instead be selected on validation data to characterize the EnKF's ceiling. Accordingly, the 25-cycle comparison in Section 7.3 reports what the paired test resolves for this configuration, and the divergence in Figure 12 is descriptive.

Deterministic super-resolution. The det-SR baseline is an EDSR-style network ($4{,}872{,}705$ parameters) trained for $40{,}000$ steps and takes the *observation only*. It is stable over 199 cycles on both systems by approaching the zero-prediction line in the unobserved band.

Cheap controls. Spectral zero-padding lifts the observation with the N_1/N_0 amplitude factor and fills the unobserved band with zeros; it is stateless, cannot diverge, and is exactly 1.0 in every unobserved band by construction, which makes it the measured climatology line rather than an assumed one. Direct insertion forecasts with the FNO and overwrites the observed band with y_t. FNO free-run is the forecaster iterated without any assimilation. Free-run and direct insertion are included to locate the divergence horizon rather than as competitive methods: knowing that an

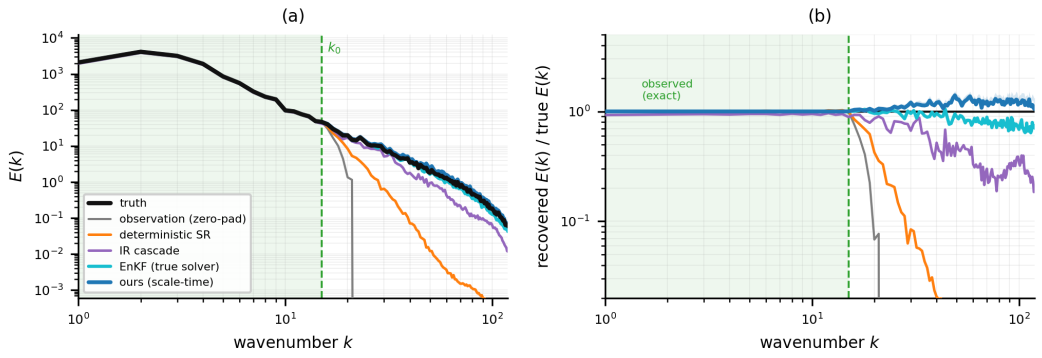

Figure 8: **Recovered power.** Absolute and truth-compensated spectra with our 5–95% member envelope. Ours tracks truth; deterministic SR loses high-k power; IR retains power without phase. Curves end at the storage truncation $k_R^{\text{eff}} = 119$.

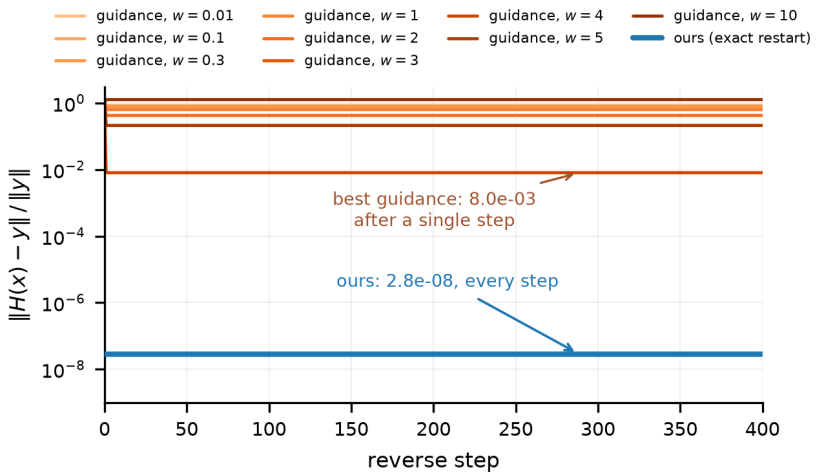

Figure 9: **Frozen-coordinate diagnostic with noiseless observations.** Observation residual $\|Hx - y^c\|/\|y^c\|$ against reverse step, where the conditioning value $y^c = y$ in this arm. Ours (blue) is flat at 2.8×10^{-8} (float32 round-trip roundoff) at all 401 states. Each guidance trace takes one update and then plateaus. The refined sweep finds 8.0×10^{-3} at $w = 4$: with $\eta = w/\sigma_\varepsilon^2 = 1$, the clean-state correction is an orthogonal projection and its missing-band error matches restart to three decimals. The nonzero sampler-state residual distinguishes that correction from hard replacement throughout the trajectory. This diagnostic tests noiseless consistency; a noisy posterior sample should not be pinned to raw y.

unassimilated forecast blows up at cycle 16 and a hard-inserted one at 23 is what makes the cycled comparison interpretable, because it separates "this method is stable" from "this problem does not punish instability".

G ADDITIONAL RESULTS

Figure 8 complements the main-text error decomposition by separating power recovery from phase accuracy across the stored spectral range.

Table 4: Clock ablation, paired over $n = 40$ shared starts at 25 cycles. Entries are the improvement over a progressive-only clock; * marks $|t| > 3$.

vs. progressive	$(15, 30]$	$(30, 60]$	$(60, 120]$	$(120, 168]$
shared clock	+17.4%*	+16.0%*	+9.1%	+3.5%
progressive + refinement	+22.6%*	+19.2%*	+11.2%*	+7.9%*
paired t, shared clock	+3.63	+4.26	+1.83	+0.63
paired t, refinement	+7.40	+7.39	+4.08	+4.42

G.1 The discretization axis

The discretization axis has three points, and we instantiate all of them with *everything else held fixed*: same architecture, same preconditioning, same conditioning path, same sampler, same forecaster. A *shared clock* moves every unobserved mode together ($\tau_{\boldsymbol{k}} = s$); a *discrete ladder* uses a few fixed cutoffs, each with its own diffusion loop; the *continuous sweep* is the construction of Section 3. Only the first and third are trained variants of one model; the ladder is our own three-rung port, retained here as a discretization ablation and never as an IR baseline (Section F).

Cost and benefit, in the same band. At the design resolution the shared clock beats the sweep by $25.4/12.4/5.4/2.3\%$ per octave at $\eta = 0$, narrowing to 8.6–2.0% at $\eta = 1$. At 96^2 the sweep beats the shared clock's best by $7.0\times$ in the first unobserved octave. Both effects land in $(15, 30]$, necessarily: transitioning last is simultaneously what exposes a band to accumulated trajectory error over the 65% of the reverse integration that follows its own transition, and what lets a later entry point rescue that same band.

Table 4 reports the paired design-resolution improvements.

Figure 10 shows that the design-resolution clock gap persists through cycling, while Figure 11 verifies that all three clock discretizations retain approximately correct spectral power.

The control that rules out a discretization artifact. The gap could be a step-budget effect, since a mode in the sweep receives only the steps falling inside its own transition. Matching *effective per-mode steps* rather than total NFE (the sweep at $M = 3200$, where each mode receives ≈ 432 steps within its transition, against the shared clock at $M = 400$) leaves the gap unchanged. Both variants also cost the same to sample (NFE 400, 1.43 s), so it is not a compute difference either.

G.2 Protocol and metric details

Observation interval. Observations arrive at $\Delta t_{\text{obs}} = 0.100 = 0.235\,\tau$, where $\tau = 1/\omega_{\text{rms}}$. The variance-weighted lag-1 correlations are $0.743/0.344/0.048$ by octave; the unweighted convention gives $0.670/0.257/0.027$. Thus the first missing octave retains forecast information while the upper octaves refresh almost completely; 58.0% of missing-band variance is fresh per interval. The observation square contains 961 modes, including 252 outside radius k_0; radial-bin bookkeeping is explained in Section D.

Why every metric is per octave. $E(k)$ spans four decades across the live set, so any metric averaged over a band is a statement about that band's low-k end. The failure of greatest concern—a method that reproduces the energetic scales while quietly stopping at fine scales—is therefore invisible in a band-averaged score and obvious in a per-octave curve. At $k = 80$ truth carries 10^{-2} against 3.7×10^3 at the forcing scale, so a 30% deficit there moves a whole-field average by a part in 10^5. We report per octave everywhere, and where we report a spectrum we also report it compensated by truth, because those are the two readings that make such a deficit visible.

Table 5 summarizes the 50-cycle outcome and distinguishes divergence from stable shrinkage to the zero-prediction solution.

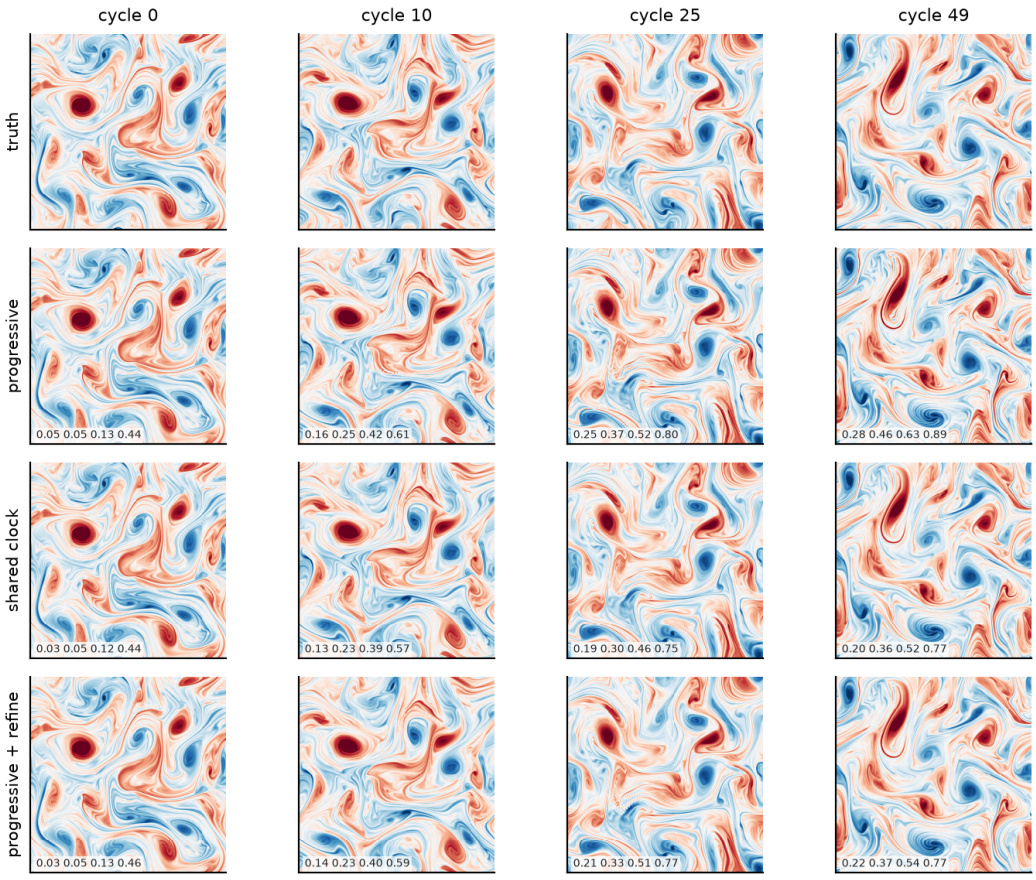

Figure 10: **Progressive against shared clock over a rollout, per octave.** The shared clock's advantage at the design resolution persists across cycles rather than being a single-analysis artifact; it is concentrated in $(15, 30]$, the band that transitions last.

Table 5: Cycled stability over 50 cycles. The failure-mode column is the point: diverging and going quiet are different failures, and a method pinned at 1.000 has stopped predicting rather than broken.

method	2D	SQG	failure mode
scale-time (ours)	reaches 50	reaches 50	error saturates, does not compound
IR cascade	stable	stable	stable but above climatology: shrinkage
deterministic SR	stable	stable	pins at 1.000; stops predicting
FNO free-run	diverges at 16	diverges	variance grows without bound
direct insertion	diverges at 23	diverges	variance grows without bound

G.3 Mismatched observation operators

The scope condition of Section 4 is testable, so we test it. Replacing the spectral sensor with a box-averaging one breaks the identification—a box average is a sinc weighting plus aliasing rather than a truncation—and the observation is then no longer a frame of the forward process. We therefore assimilate it with guidance and measure the penalty: with each arm at its own best weight in the refined sweep, the box-averaged observation costs $12.2/6.9/2.7/1.4$ across the four octaves relative to the spectral one. The gap is largest where the observation is most informative and decays with k.

The weight response identifies why. For spectral H, $w = 4$ makes the clean-state correction an exact projection and reduces first-octave error to 0.026. The box arm gains nothing there: its error is 0.325, slightly worse than 0.319 near its own optimum. A sinc weighting with aliasing is not

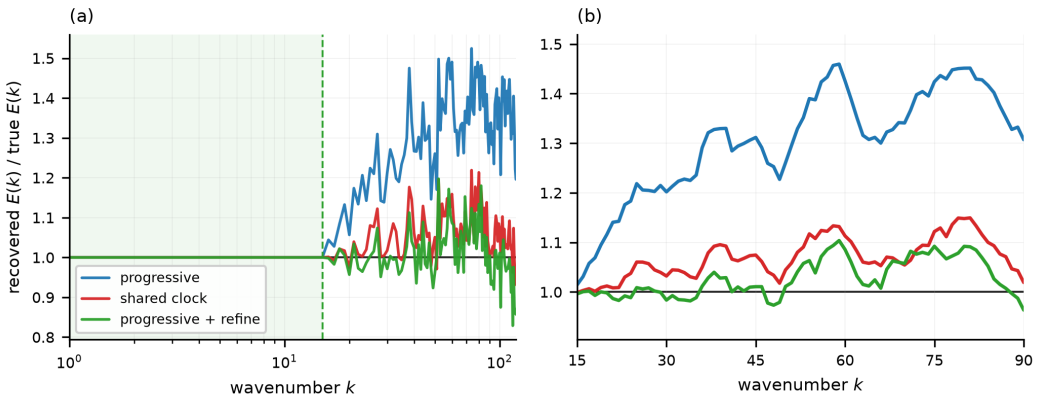

Figure 11: **Recovered $E(k)$ on the discretization axis, compensated by truth.** Shared clock, discrete ladder and continuous sweep, everything else held fixed; a perfect method is flat at 1.0. (a) the full stored band to $k_R^{\text{eff}} = 119$, with the observed square shaded. Left of k_0 every variant is exact by construction, so the clocks differ in accuracy rather than in spectral fidelity. (b) the same curves zoomed to $k \in [15, 90]$ on a linear axis, where the three curves separate. We plot the ratio rather than the absolute spectrum because all three track truth to within a factor of order one across four decades of decay, so on a log axis they lie on top of one another and the panel would say nothing.

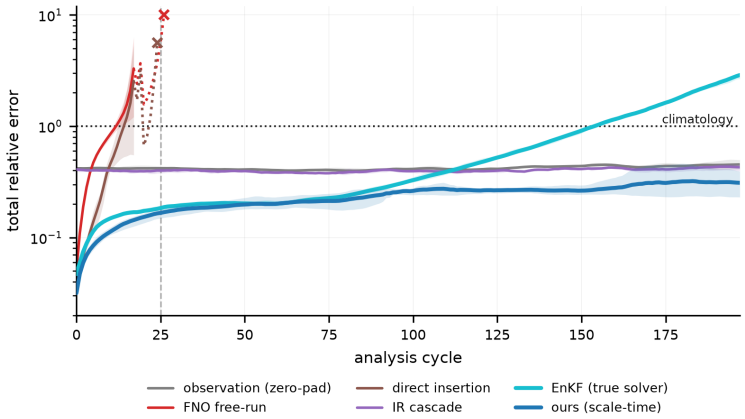

Figure 12: **Long-horizon total error.** Curves span 198 cycles over three trajectories; diverged methods end at $\times$. The dashed vertical line marks the 25-cycle paired horizon. EnKF settings are untuned, and the long-run separations are descriptive.

an orthogonal projection, so no scalar step turns this correction into one. The scope condition is therefore visible in the tuning curve, rather than only in the proof.

The same experiment on SQG. We repeat it on the second system, not to re-establish the claim but to test the implementation: $g_{\boldsymbol{k}}(s) = 0$ on the observed band is a property of the *schedule*, so it must hold identically under different physics, and a replication that failed would indicate a bug rather than a limitation. It does not fail. Guidance again plateaus after one update at every tested weight. Since both systems use $\sigma_\varepsilon = 2$, the unit-step prediction is again $w = 4$; there SQG guidance reaches 7.1×10^{-3} residual and matches restart's $0.119/0.210/0.385/0.679$ error to three decimals. This cross-system replication supports the step-size explanation rather than a flow-specific optimum.

Restart's final-state residual is exactly zero on both systems, but that number alone is not diagnostic: any method that hard-projects the final state achieves it by construction. The 2.8×10^{-8} of Figure 9 is the stronger measurement: the residual at *every* one of 401 intermediate reverse states, which is

float32 round-trip roundoff and is what makes replacement the dynamics rather than a correction applied at the end. Only the final-state number was recorded on SQG.

The box control also replicates in shape. Its refined-sweep penalty is $4.4/2.1/1.6/1.4$ on SQG and $12.2/6.9/2.7/1.4$ on 2D, decreasing monotonically toward the corner. The levels differ, so this is not a claim of flow-independent magnitude. On both systems the spectral arm improves sharply at $w = 4$ while the box arm does not (0.536 versus its best 0.529 on SQG; 0.325 versus 0.319 on 2D), which is the operator-level replication the experiment supports.

G.4 The sparse observation, $k_0 = 7$

The configuration. At $k_0 = 7$ the observation is a 16^2 grid: 225 modes, against 961 at the design resolution. The schedule is re-derived rather than reused: $\kappa = 3.5$, $\sigma^2_{\max} = 173.3$, and a sweep spanning 6.59 octaves against 5.49, because every one of those follows from k_0. The forecaster does *not* need retraining, since it acts on the full field and never sees the observation. 31.5% of live variance is now unobserved, against 17.8% at $k_0 = 15$.

What survives and what does not. The construction itself is unaffected: recovered $E(k)$ sits within 2–8% of truth in every band ($1.079/1.041/1.020/1.025/1.017$), so the sampler is still producing a spectrally correct field. Cycled, it reaches the full 50-cycle horizon with stable variance, whereas direct insertion at the same observation doubles its variance by cycle 22 and crosses climatology in error by cycle 6.

Forecast-bound crossover. A forecast-bound comparison captures what the sparse observation costs. Compare each analysis against a forecast started from *truth*, which is an upper bound on what the dynamics alone can supply. At $k_0 = 15$ our analysis *beats* that bound by 27–38%: the observation is adding information the forecast does not have. At $k_0 = 7$ it *loses* to it by 3–8×. The sparse observation has stopped being the dominant source of information, and the analysis is limited by what the forecast can carry rather than by the sampler. Section G.10 shows the same transition in the calibration diagnostics, where spread/error falls to 0.10–0.28.

This is also where our one clear loss to the cascade lives (Section 8), and the two facts are the same fact: where neither method has information, emitting calibrated but uncorrelated variance is penalized and declining to guess is not. Figure 14 shows the qualitative outcome, and Figure 15 reports the paired per-octave comparison, including the corner loss.

G.5 SQG

Why a second system, and why this one. SQG is not a repetition of the 2D experiment with different data. It differs in the property most likely to interact with a scale-indexed forward process: its eddy turnover time is *scale-dependent*, measured at $\tau \sim k^{-0.34}$ to $k^{-0.52}$ across five solver configurations, where the 2D enstrophy cascade has $\tau \sim k^0$. The asymptotic SQG prediction is $-2/3$ and none of our configurations reaches it. The fitted range spans barely two octaves and sits close to both the forcing and dissipation scales, so we report the measurement as *measured* -0.34 *to* -0.52 *against an asymptotic* $-2/3$ rather than as confirmation of the theory. We also tested and rejected the hypothesis that the shallowest group was dissipative contamination: halving ν_4 moved it by 0.004.

Spectral fit window. SQG's spectrum is steeper: the fitted per-mode slope is -3.09 over $[22, 60]$, against 2D's -2.48 on the same window. Fitting $[8, 60]$ makes SQG's two halves differ by 0.835, but $k_f = 8$ places the start of that window at the forcing scale, so its steep lower half measures the forcing rolloff rather than cascade curvature. Over $[22, 60]$, inertial for both systems, SQG's halves differ by 0.004 against 2D's 0.142: SQG is the *cleaner* power law ($R^2 = 0.9999$). The systems differ in exponent, which is the physical claim, not in power-law quality.

Results. Our error increases relative to 2D and the margin narrows. IR and deterministic SR instead have slightly lower relative errors on SQG, so the cross-system table does not define a universal difficulty ordering. At the snapshot we reach $0.445/0.784/1.031/1.227$, crossing the zero-prediction score in the upper two octaves (Table 1 and Fig. 19). IR reaches

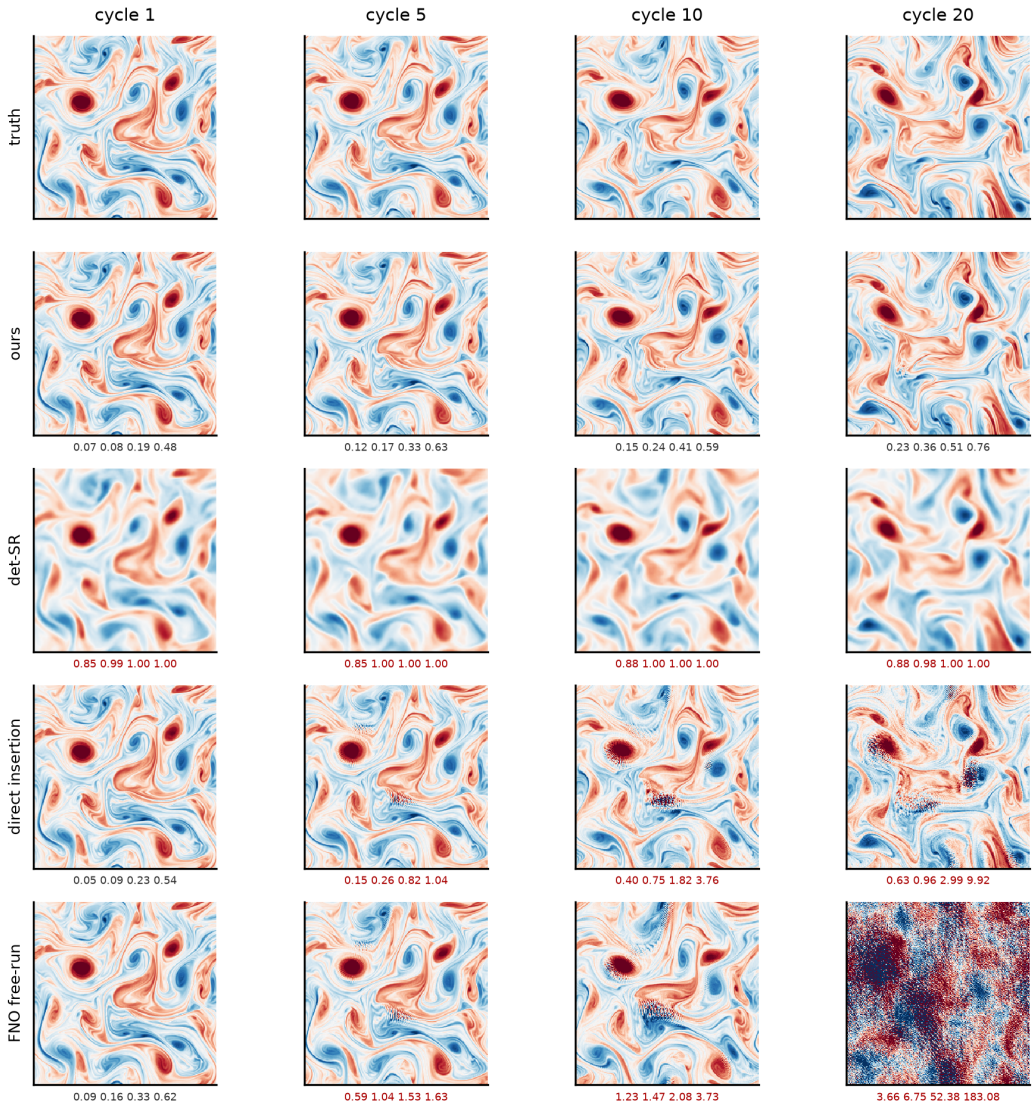

Figure 13: **2D cycled degradation, per octave.** How much of the single-analysis error is recovered at each cycle, and where it saturates. Ours saturates; the cheap baselines do not.

$0.685/0.968/1.081/1.080$: we lead in the first three bands and trail in the corner. Deterministic SR is also lower in the upper two bands, at $0.998/1.001$, by collapsing to the zero-prediction solution rather than recovering those scales. Figure 20 separates those cases.

The cheap baselines fail *earlier* than on 2D: FNO free-run diverges at cycle 12 against 16, and direct insertion at 19 against 23. An EnKF sharing our forecaster fails the variance criterion at cycle 12, where on 2D it reached 24. Its upper-octave spread/error ratios are 5.5 and 4.6, so injected spread rather than forecast skill accounts for its dispersion. Figure 21 shows how these failures develop by octave.

Long-horizon stability on SQG. Over 198 cycles on the same trajectories and seed as the 2D study (Figures 17 and 18), our per-octave error ends at $0.487/0.798/1.032/1.201$, essentially unchanged from the $0.445/0.784/1.031/1.227$ of the cycle-31 snapshot, and marginally *better* in the upper two octaves. Compare 2D over the same horizon, where the first octave drifts from 0.262 to 0.471. On SQG we start worse and stay there; the error saturates rather than compounding, which is a stronger stability statement than the 2D curve makes.

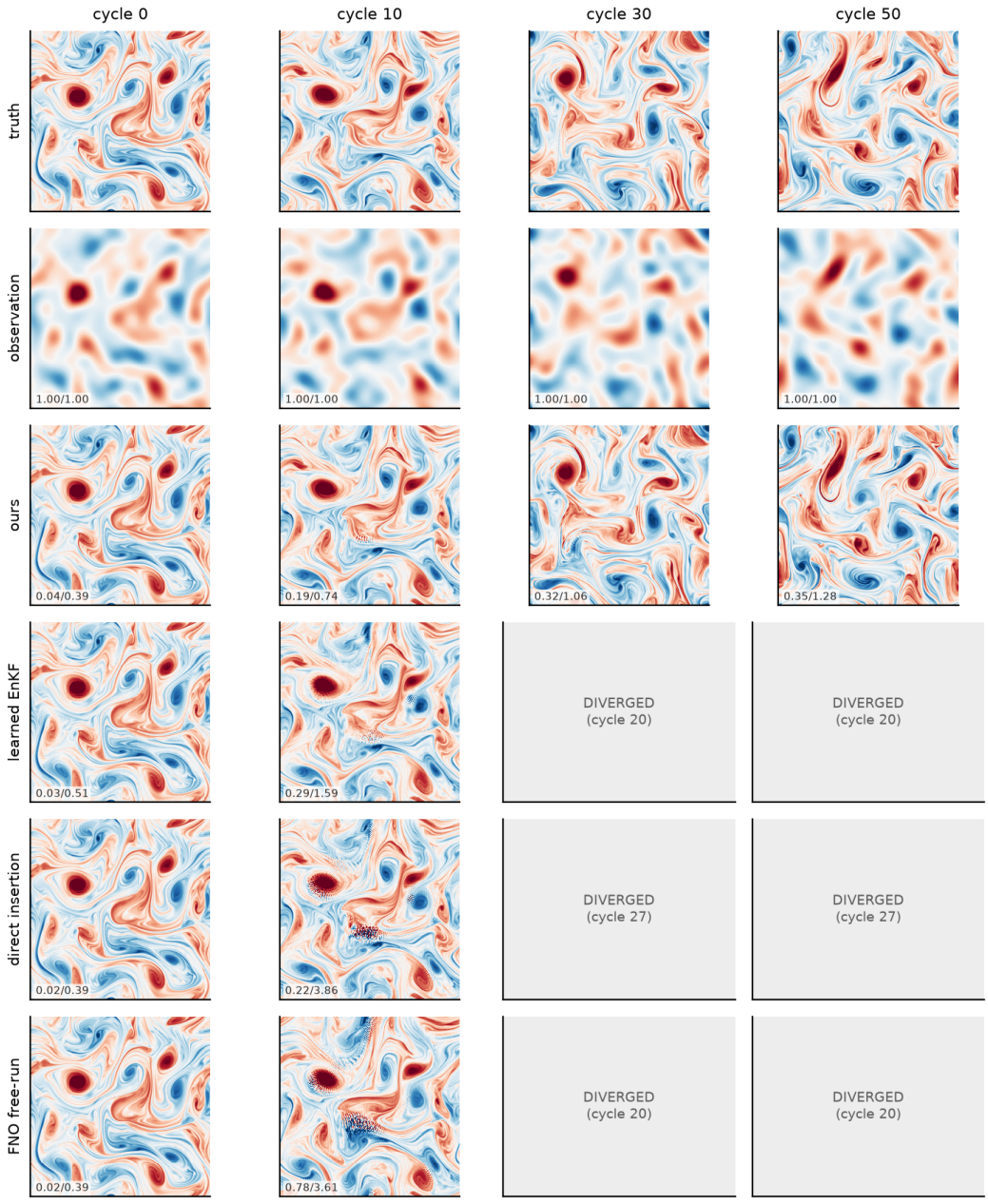

Figure 14: **The sparse observation, qualitatively.** What each method believes at $k_0 = 7$, where 31.5% of live variance is unobserved against 17.8% at the design resolution.

In total error the ordering is clearer than the per-octave table suggests: at cycle 197 we reach 0.222 against spectral zero-padding's 0.320, while free-run and direct insertion have long since diverged (cycles 12 and 19). Deterministic SR survives, and ends at $0.825/0.975/1.000/1.001$, pinned at the zero-prediction line in the upper three octaves, which is the same "stable because it stopped predicting" failure the 2D study shows. It appears in the per-octave panel but not the total-error one, whose long-horizon run recorded field RMS rather than a total.

The cascade fails the same way on different physics. For SQG, we train the IR recipe of Dhingra et al. (2026) on this dataset at the same σ setting selected for 2D and for the same $40{,}000$ steps. **We did not retune it.** Its augmentation sigmas were swept on 2D, and retuning the baseline while our

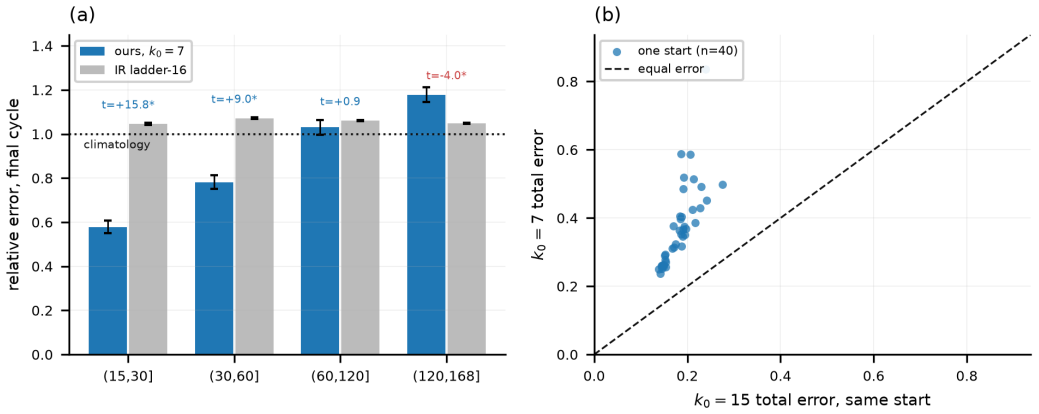

Figure 15: $k_0 = 7$ **against the cascade, paired** $n = 40$**.** Per octave at the final cycle, with paired t annotated. Ours is lower below $k = 60$ and higher in the corner at $t = -3.96$. This difference follows the mechanism of equation 5: where neither method has information, a near-zero output scores ≈ 1.0 while a calibrated sampler pays $\sqrt{1+P}$. The comparison is also confounded in the cascade's favor, its 16^2 rung carrying 272 modes against our 224.

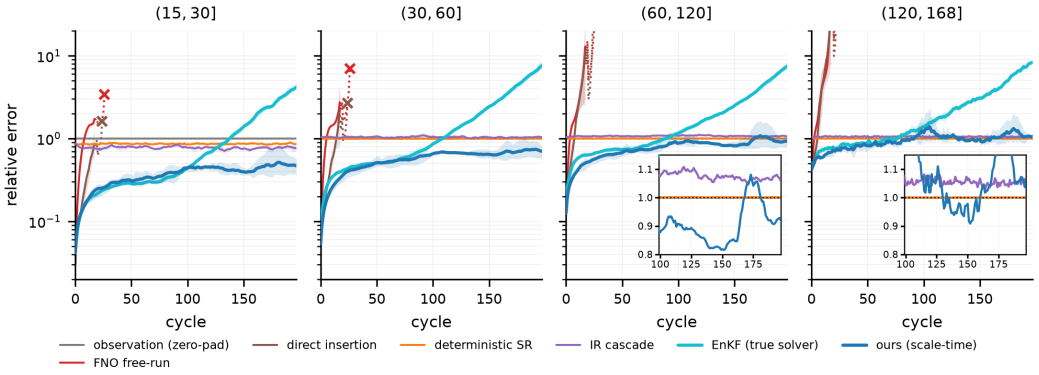

Figure 16: **The same run, per octave.** Ours stays below climatology (dotted) in every band through cycle 198 except the corner, which it reaches at 1.06; the perfect-dynamics EnKF crosses in all four. The IR cascade is flat and just above climatology from the second octave onward: stable but not predicting. Sample counts and the untuned-inflation caveat are as in Figure 12.

own method keeps its 2D-derived settings would bias the comparison in our favor. The handicap is therefore deliberate, runs against the baseline, and is stated here rather than left for a reader to infer.

Under this unchanged 2D tuning, it shows the same qualitative pattern as on 2D. It survives all 198 cycles without ever blowing up, ending at $0.731/1.006/1.073/1.092$ per octave and 0.286 total against our 0.222. Read those two rows together: the cascade is *stable* and at or above climatology in every band above the first, which is the shrinkage failure of equation 5 reproduced under different physics. A divergence criterion passes it on both systems.

That the same failure appears on a flow with scale-dependent turnover is the more useful observation, and the decomposition equation 5 makes it quantitative. The following values are integrated over each evaluation band; unlike medians of the per-shell ratios in Figure 20, they obey equation 5 exactly:

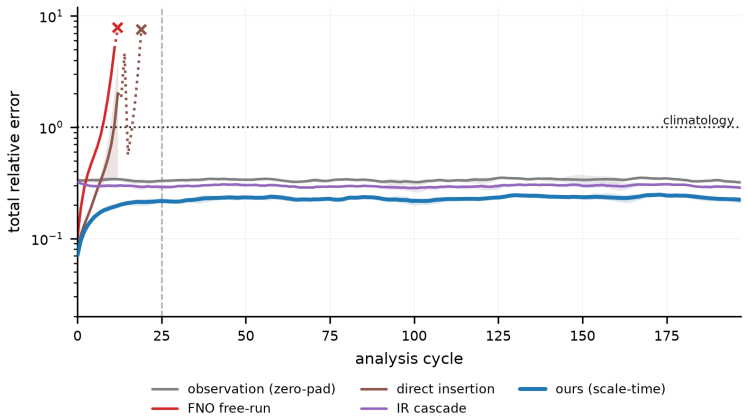

Figure 17: **SQG long horizon, 198 cycles, total error.** Same protocol, trajectories and seed as Figure 12. Ours ends at 0.222 against spectral zero-padding's 0.320; free-run and direct insertion end at their divergence cycles (12 and 19), earlier than their 2D counterparts (16, 23). Deterministic SR is absent here and present in Figure 18, because its long-horizon run recorded field RMS rather than a total.

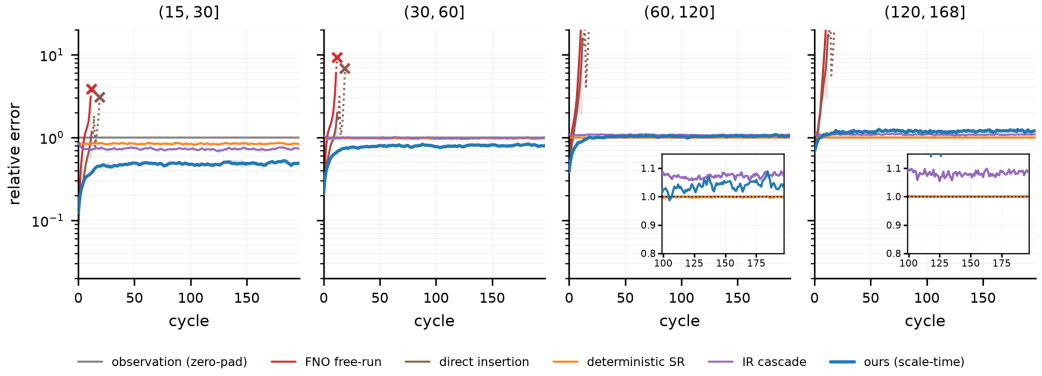

Figure 18: **SQG long horizon, per octave.** Ours saturates rather than compounding; the end state is within a few percent of the cycle-31 snapshot in every band. Deterministic SR pins at 1.000 in the upper three octaves: stable, and not predicting.

band	2D (P / C)	SQG (P / C)
$(15, 30]$	0.802 / +0.682	0.688 / +0.735
$(30, 60]$	0.541 / +0.388	0.390 / +0.363
$(60, 120]$	0.285 / +0.073	0.294 / +0.115
$(120, 168]$	0.184 / −0.078	0.228 / +0.065

In both cases the correlation decays toward zero while nonzero power remains, which drives err toward $\sqrt{1+P} > 1$. In SQG's corner, ours retains $P = 1.126$ with $C = 0.292$, whereas IR shrinks to $P = 0.228$ with $C = 0.065$; once correlation is weak, that shrinkage explains IR's lower RMSE. The repeated cascade signature is therefore stable low-power prediction, rather than recovery of the missing phase.

G.6 Off-design entry details

A model trained at 32^2 is handed a finer 96^2 observation ($k_0' = 47$) at inference, with no retraining. We enter the progressive process at $k_c(s^*) = 47$ and use the observation below that cutoff. The finite-width transition means this state is only approximately distributed as the corresponding

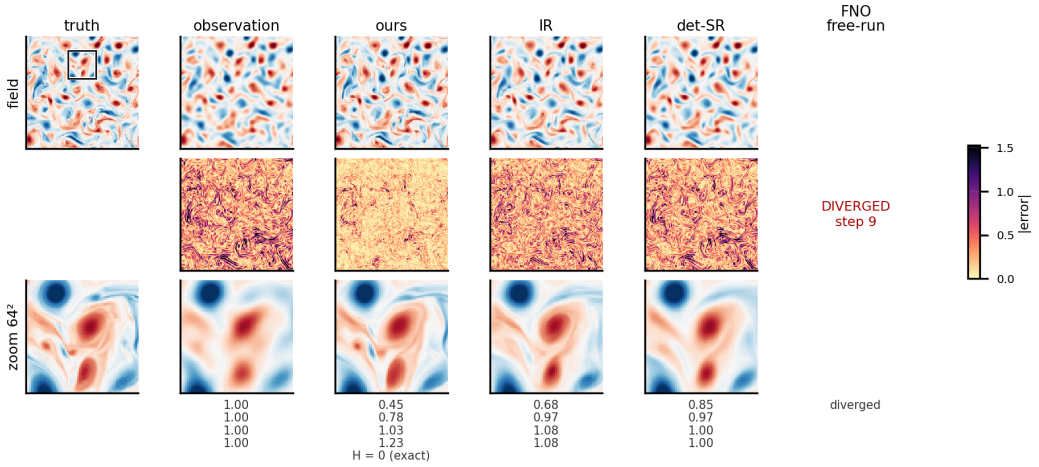

Figure 19: **SQG single analysis.** Same layout as Figure 2. The margin is narrower than on 2D and we cross the zero-prediction score in the upper two octaves. IR leads us in the corner; deterministic SR scores lower in the upper bands by collapsing toward zero power.

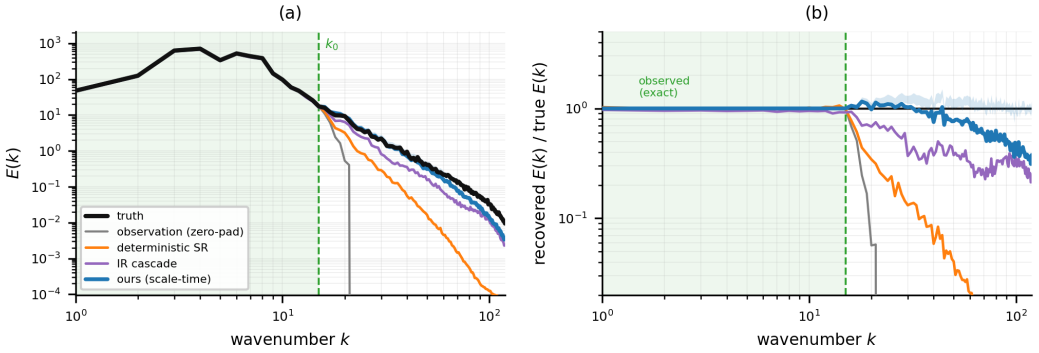

Figure 20: **SQG recovered spectrum.** Absolute $E(k)$ (left) and the same curves divided shell-wise by truth (right), truncated at the stored support $k = 119$. Ours follows truth farther into the missing band; IR and deterministic SR lose power. The band-integrated P, C values in the text, rather than medians of this shell-wise ratio, are the quantities that obey equation 5. Diverged methods are named rather than drawn.

marginal; at the most useful operating point, the first unobserved octave nevertheless improves $9.4\times$ over design-resolution entry. The curve below reports the approximation explicitly.

Two baselines are flat across the off-design entries, for different reasons. The *shared clock* is structurally unable: its only band-limited frame sits at s_{dead}, where $a_{\boldsymbol{k}}(s_{\mathrm{dead}}) = 6.1 \times 10^{-17}$, so data pasted into a wider square is multiplied by zero. Both variants, discarding the extra modes or pasting them anyway, return the design-resolution result to four decimal places, confirming the structural limitation. The *IR ladder* is unable for a different reason: 96 is not a dyadic rung, so the finer observation is restricted back to 32^2 before the corrector sees it, and its result is likewise identical to its 32^2 number to four decimals. Neither is a tuning deficit.

The cost and benefit occur in the same band. The sweep is 25.4% worse at 32^2 and $7.0\times$ better than the shared clock at 96^2, both in $(15, 30]$. It must be the same band: transitioning last is simultaneously what exposes a band to accumulated trajectory error and what lets a later entry point rescue that same band. A method that wants multi-resolution entry pays for it there and is repaid there. For fixed design-resolution entry, the shared clock is preferable.

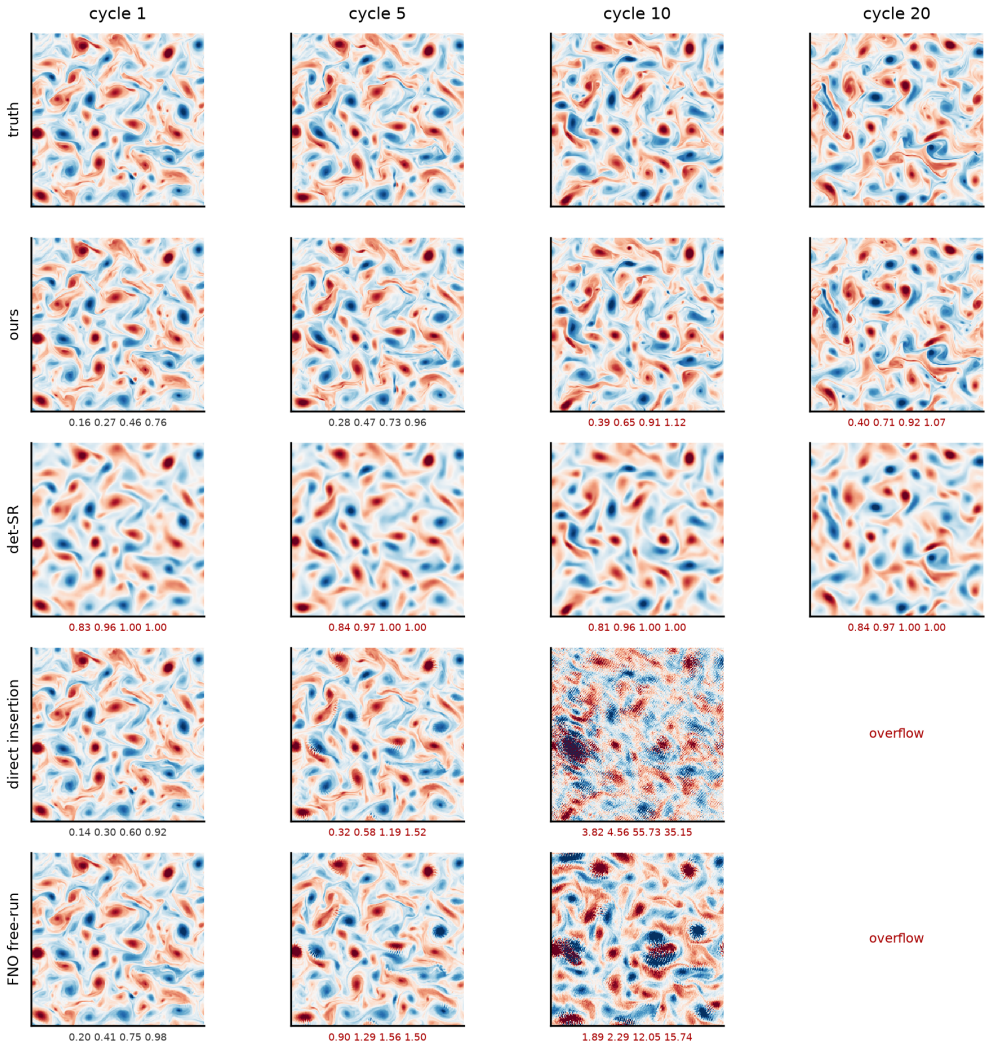

Figure 21: **SQG cycled degradation.** Same layout as the 2D panel (Figure 13).

Exactness–utility curve. The $9.4\times$ headline is the best point on a trade we can move along, and the trade is irreducible rather than a tuning failure. Entering at k_0' with *full* overshoot is exact but vacuous: the band above k_0' is already dead at that s', so equation 4 faithfully evaluates to fresh noise there and the 96^2 observation is assimilated as if it were 32^2. Zero overshoot uses the data but incurs an order-one signal-coefficient mismatch over $\sim mw$ octaves above k_0'. Measured, the first unobserved octave runs $0.0053 \to 0.0160 \to 0.0279 \to 0.0410$ as the overshoot goes $1.0 \to 1.25 \to 1.5 \to 2.0$, against 0.0498 for design-resolution entry, degrading $0/24/51/80\%$ of the way back to the reference, and reaching it to four decimals at full overshoot, which is the vacuity made quantitative. The exactness cost moves in opposition: at zero overshoot the largest retained signal coefficient above k_0' is $a_{\boldsymbol{k}} = 0.6355$, so the entry state discards correlation the true marginal retains, concentrated in roughly one octave above k_0' and negligible by $k \approx 66$.

Two limits are hard, and neither is an implementation gap. *Noisy* off-design entry is infeasible outright: on the resolved band at mid-sweep $\sigma_{\boldsymbol{k}}(s') = 0$ and $a_{\boldsymbol{k}}(s') = 1$, so the noise budget reads $0 \geq r_{\boldsymbol{k}}$ and fails for any sensor noise at all. And entry at resolutions *coarser* than trained is impossible, since (C2) exempts those modes at every s and no frame of the process has that shape.

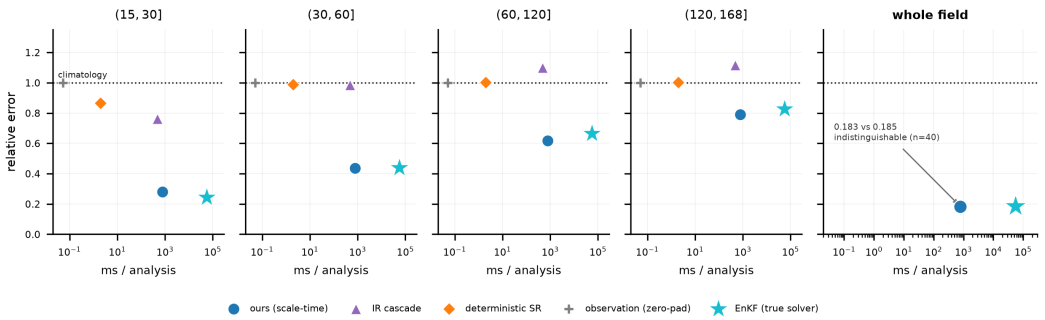

Figure 22: **Accuracy against wall-clock per analysis, per octave.** Lower-left is better. Each method sits at its own measured converged operating point, so the axis is work actually done rather than a step count – NFE is not comparable across these methods. The perfect-dynamics EnKF sits far to the right and level with us: that is the figure's point, and it is a trade rather than a ranking. Its 55 s is CPU NUMPY with a serial member loop, so a batched or distributed implementation could move it substantially left; the panel states that cost model rather than presenting the timing as an algorithmic speed ratio.

G.7 SAMPLER COST

Each method at its own converged point. Reporting a cost comparison at a shared step count would flatter whichever method converges later, so each is priced where its own accuracy has stopped improving. Ours converges at $M = 100$: the per-octave error at $M = 100, 200, 400, 800$ is flat to within 1% ($0.0445/0.0511/0.1168/0.3790$ at $M = 100$, and marginally *worse* at larger M, which is discretization noise rather than a trend). The cascade converges at 25 steps per rung, where its total error is 0.403 against 0.412 at 50 and 0.410 at 100.

The numbers, and what they are not. At those points ours costs 793 ms per analysis (100×7.93 ms at 256^2) and the cascade 483 ms (25 steps on each of three rungs, at 6.4, 5.2 and 7.8 ms). **The cascade is $1.64\times$ cheaper than we are.** It is also at or above climatology in every band above $k = 30$, so the trade is that we buy a recovered spectrum for $1.64\times$ the wall-clock; we are not faster.

NFE is not a common currency here and should not be quoted as one: two of the cascade's three rungs run below 256^2, so equal step counts are unequal work. That is why Figure 22 is drawn against wall-clock; Figure 23 displays the ranking change produced by using NFE instead. An independent snapshot in Figure 24 checks that the single-analysis comparison is not specific to the main example.

G.8 OBSERVATION-NOISE DETAILS

Sensor noise enters through the entry time alone: s^* is the scale-time whose ramp noise level matches σ_ε, so a noisier observation simply enters earlier and denoises for longer. There is no inflation factor and no likelihood weight to tune. Sweeping $\sigma_\varepsilon \in \{0.5, 1, 2, 3, 4.511\}$ moves the entry time $0.868 \to 1.0$ and the observed-band error $0.021 \to 0.044$, a smooth, tuning-free degradation (Figure 25).

Misspecification is asymmetric, and the asymmetry is the result. Assuming the wrong σ_ε costs differently in the two directions. On the band where the effect is visible, $k \in (10, 15]$, assuming half the true noise costs $3.55\times$ (0.104 against 0.029), while assuming $1.5\times$ too much costs *nothing* (0.026, marginally better). The mechanism is immediate from the construction: on the ramp $a_{\boldsymbol{k}} \equiv 1$, so entering at a higher scale-time discards no signal and only spends more reverse steps denoising a band that is already intact. Under-stating the noise, by contrast, enters below the level the observation actually carries, and the sampler then treats sensor noise as signal.

The band restriction is essential and we state it wherever either number appears. Over the *whole* observed square the same experiment reads $0.058/0.034/0.033$, i.e. $1.72\times$ rather than $3.55\times$, be-

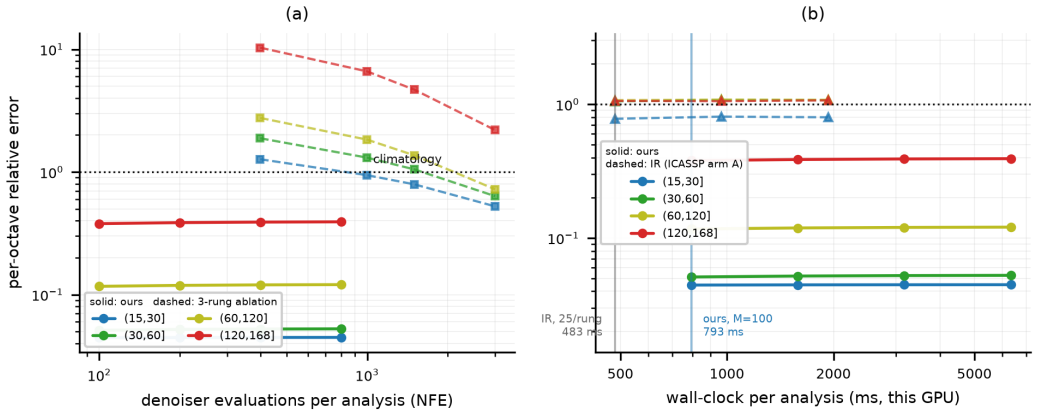

Figure 23: **Accuracy against wall-clock and against NFE.** Each method at its own measured converged operating point. The two panels disagree, and that is the point: NFE flatters the cascade because two of its three rungs run below 256^2, so equal step counts are unequal work. Wall-clock is the common currency.

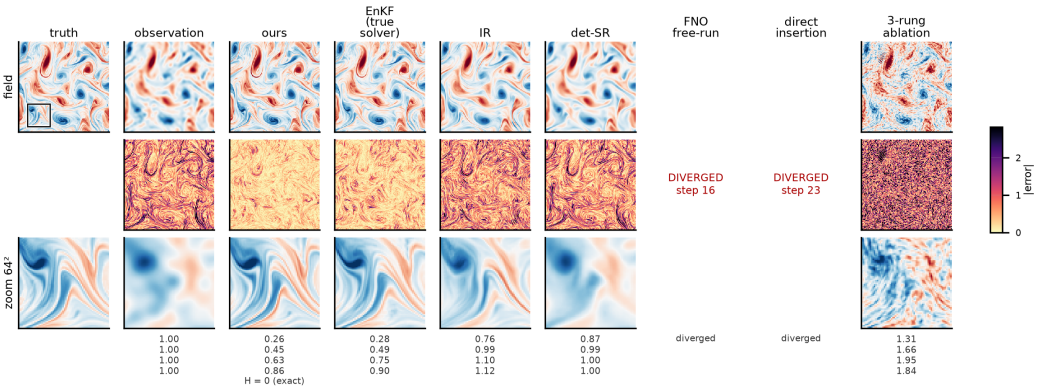

Figure 24: **A second 2D snapshot.** The same comparison as Figure 2 at an independently selected cycle, to show the single-analysis result is not a favorable draw.

cause that band's mean $|x|^2$ is dominated by $k \approx 2$ where $E(k) = 3.7 \times 10^3$ and an edge effect is averaged away. This is the same convention-11 trap the per-octave reporting exists to avoid.

G.9 Additional limitations and scope details

Sensor class. The theorem covers orthogonal projections that the forward process reaches in order. Besides spectral truncation, this can include ideal antialiased downsampling or band-limited imaging in a matching orthogonal basis. It does not include box averaging with aliasing or sparse point sampling merely because those maps are linear: their retained coordinates are not one of the nested subspaces swept out here. Designing a different forward process around such a sensor is open.

Comparison qualifications. The cascade's 32^2 restriction retains 96 additional Nyquist modes carrying 0.462% of observed-band energy; at $k_0 = 7$ it receives 272 modes against our 224. Both differences favor the cascade. Conversely, its SQG augmentation settings were selected on 2D and not retuned, which weakens that comparison. Section F gives the audit and the reasons these inputs were not altered. The EnKF uses default inflation 1.05 and additive noise 0.20 without validation tuning, so neither its short-horizon result nor long-horizon failure estimates the method's ceiling.

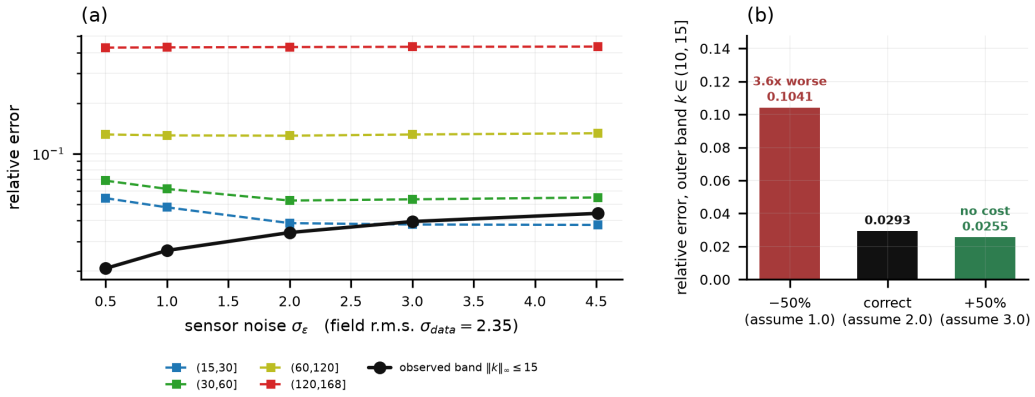

Figure 25: **Observation noise enters through the entry time alone.** (a) sweeping σ_ε moves s^* up the ramp and degrades the analysis smoothly, with no tuned parameter anywhere in the loop; the black curve is the observed band and the dashed curves the four unobserved octaves. (b) misspecification is asymmetric – assuming too little noise costs $3.55\times$ on $k \in (10, 15]$, assuming too much costs nothing, because $a_{\boldsymbol{k}} \equiv 1$ on the ramp so a later entry discards no signal. The band restriction is essential: over the whole observed square the same effect reads $1.72\times$, the low-k end dominating the average.

Measurements still missing. The shared-clock ablation was trained only without a sensor-noise ramp, and the analytic $1.49\times$ grid-adaptivity estimate has no timing benchmark. The five-trajectory test split also prevents a reliable trajectory-level uncertainty estimate. These omissions require new experiments rather than editorial qualification.

G.10 Calibration

Three regimes, one reference. Our sampler is stochastic, so it produces an ensemble at no extra architectural cost, and we report spread/skill per octave against a perfect-denoiser reference of 0.956, the value a correctly calibrated sampler would give under this diagnostic, computed rather than assumed. Three regimes appear. Observation-only (no forecast context) gives 0.83–0.89: mildly under-dispersed. Sequential assimilation at $k_0 = 15$ gives 0.70–1.00, spanning the reference. At $k_0 = 7$ it falls to 0.10–0.28 and becomes badly over-confident. The cause is identifiable: the context c_t does not change when k_0 shrinks, so the posterior the model was trained to represent is narrower than the one the sparser observation actually implies.

For comparison, the EnKF sharing our forecaster is over-dispersed in the opposite direction, at $1.61/2.22/3.62/4.75$, with CRPS an order of magnitude worse than ours in every band ($0.114/0.191/0.443/0.966$ against $0.010/0.017/0.041/0.122$).

Effective sample size across space and time. An effective sample size must discount correlation along *every* axis being pooled. Treating the $33{,}932$ modes within an octave as independent overstates the effective count by 3.5–$1294\times$ because the field has only a few hundred independent structures per octave. We therefore validate the estimator on a case with a known answer. An integral-of-|autocorrelation| estimator recovers only 3–7% of the true mode count on band-filtered white noise, because a band-pass autocorrelation oscillates and the $|\cdot|$ sum overcounts the correlation area. The variance-of-mean-square estimator $N_{\text{eff}} = 2\langle f^2\rangle^2/\mathrm{Var}(\langle f^2\rangle)$ recovers the true count to within 1–9% and is the one used here. Figure 26 collects the three spread/skill regimes and the forecast-bound comparison using this corrected effective count.

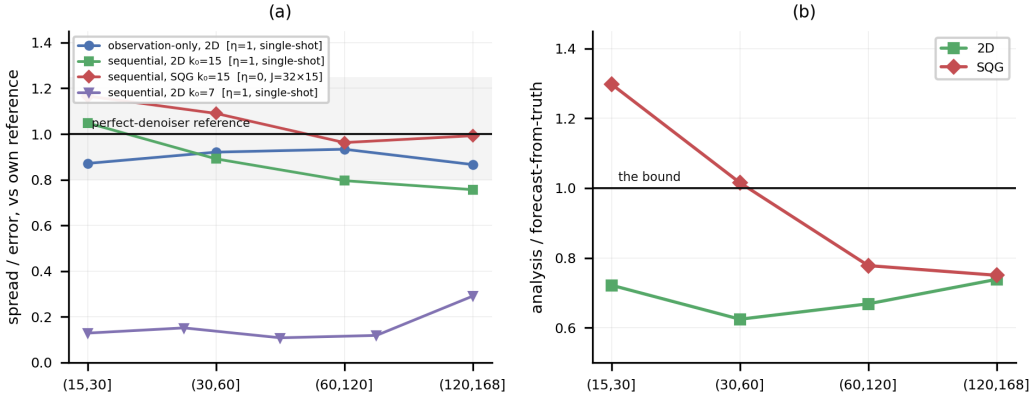

Figure 26: **Calibration, per octave.** (a) spread/skill by regime against the perfect-denoiser reference of 0.956, the value a correctly calibrated sampler gives under this diagnostic, computed rather than assumed. (b) the analysis against a forecast started from *truth*; below 1.0 the analysis beats a bound no filter can exceed from the dynamics alone. Both use the effective sample size actually measured rather than the raw mode count, which overstates the degrees of freedom by 3.5–1294×.