STABILIZING PROBABILISTIC FORECASTS VIA STOCHASTIC INTERPOLATION CASCADES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study whether generative forecasters sample the correct distribution of future states given coarse observations of a physical system. We build a benchmark pairing observed coarse-grained states with many future states generated by varying the unobserved variables, which enables direct evaluation of conditional distributions. The benchmark reveals that commonly used diffusion models and stochastic interpolants trained directly over long forecast horizons can generate samples with plausible spatial statistics but substantially mismatched conditional distribution. We then propose and analyze a stylized mathematical model of dynamics and prove the existence of a regime where learning and composing shorter transitions outperforms learning the full transition directly. Shorter transitions reduce approximation error, while repeated conditioning can discard hidden information, leading to an optimal intermediate number of stages. This analysis motivates Cascaded SI, which implements staged forecasting with stochastic interpolants. Experiments across several physical systems show substantial improvements in conditional sampling, with gains persisting when the number of supervised transitions is held fixed.

1 INTRODUCTION

Predicting the future state of a system from past observations is a fundamental problem across the sciences. A key obstacle is the fact that observation are often coarse: they provide only partial information about the state of the system, while unobserved components also influence the evolution. As a consequence, the same observed state can lead to different futures, and the forecasting problem becomes probabilistic in nature Chorin et al. (2000; 2002). An archetypal example is found in chaotic systems, most notably weather: arbitrarily close trajectories diverge as small differences in the initial conditions amplify, in what is known as the "butterfly effect" Lorenz (1963); Lorenz (1969).

The forecasting problem can then be formalized as that of searching for the conditional distribution $P(X_T \mid X_0)$ of the final states X_T given an initial state X_0. State-of-the-art approaches use diffusion models and stochastic interpolants (SIs) to generate conditional forecasts of weather and physical dynamics (Price et al., 2025; Chen et al., 2024; Kohl et al., 2026) (see Appendix A for a fuller discussion). Their generative sampling dynamics need not coincide with the physical evolution between observations (Song et al., 2021). However, it is often the case that such information is available, in the form of intermediate observations along physical trajectories (Rühling Cachay et al., 2023; Rozet & Louppe, 2023).

This raises the question of whether including these observations to learn shorter transitions can improve conditional forecasting over direct endpoint learning. Using the framework of SIs Albergo et al. (2025), we study how cascade depth affects conditional forecasting accuracy and analyze the competing effects of approximation error and information loss in a simplified model. In particular, we consider two variants of SIs (see Figure 1):

1. *Direct SIs* are classically trained on endpoint pairs (X_0, X_T) and directly learn the posterior distribution;
2. *Cascaded SIs* are trained on shorter, consecutive pairs $(X_{t_i}, X_{t_{i+1}})$, and learn the conditional transitions $P(X_{t_{i+1}} \mid X_{t_i})$ over time intervals of length $|t_{i+1} - t_i| = T/S$, where S is the number of intervals, composing them to sample the posterior distribution.

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.

Recent probabilistic forecasters use both approaches Chen et al. (2024); Price et al. (2025); Kohl et al. (2026); Mücke & Sanderse (2026), but two questions remain open. How short should the transitions be for the composed forecast to recover the conditional law at a fixed horizon? And, since the observed process is not Markov under partial observation, at what point does conditioning each transition only on the current observed state lose enough information about X_0 to outweigh the benefit of shorter transitions?

Our analysis focuses on the tradeoff between the two approaches, the cascade depth S, and the optimal value $S^\star$ suggested by the theory. In particular, our main contributions are:

- We introduce a new benchmark for conditional forecasting under partial observation spanning various physical systems (see Table 1 for an overview), chosen to test forecasting under different kinds of dynamics. We construct ensemble with a shared observed initial state and different hidden components, obtaining multiple possible future states from the same observation. In Section 3, we evaluate Direct and Cascaded SI on these conditional distributions and find that cascading can dramatically improve conditional sampling accuracy.
- We give possible explanations for the experimental results. In Section 4, we introduce a stylized rotation model with an unobserved Ornstein–Uhlenbeck (OU) process that influences the future dynamics. For a simple SI model that approximates the conditional mean with a quadratic function, we compute the endpoint distribution error exactly and prove that, when the contribution of the OU process is sufficiently small but nonzero, the error is minimized at a finite cascade depth $S^\star \geq 2$. In Appendix C, we develop a more general analysis of why cascading can help and why using too many transitions can downgrade performance.

2 A Benchmark for Conditional Forecasting

To evaluate conditional forecasters, we create a new benchmark spanning several physical systems and use a hierarchical sampling design to measure how well they recover the correct conditional distributions. Let Y_t denote the full state of a system at time t, and X_t its observed state. At the initial time, we write $Y_0 = (X_0, Z_0)$, where Z_0 denotes the unobserved component. We distinguish between (i) the *evolution* under an operator $\mathcal{G}_t$,

$$Y_t = \mathcal{G}_t(Y_0), \tag{1}$$

where $\mathcal{G}_t$ denotes the numerical evolution operator, and (ii) the *observation* through an operator $\mathcal{H}_t$ that maps the full state Y_t to the observed state X_t:[1]

$$\mathcal{H}_t(Y_t) = X_t. \tag{2}$$

In all benchmark systems, the observed fields are represented by their low-frequency Fourier modes; the precise observed and hidden variables depend on the system. An overview is provided in Table 1.

Since the observed initial state X_0 does not determine the hidden component Z_0, the same observation can correspond to multiple full initial states and hence to different observed futures. The forecasting target is therefore the conditional law

$$P(X_T \mid X_0 = x_0). \tag{3}$$

Our hierarchical sampling design provides samples from this law directly. For each conditioning ensemble e, we fix an observed initial state $x_0^{(e)}$ and vary the hidden component:

$$Y_0^{(e,i)} = \big(x_0^{(e)}, z_0^{(e,i)}\big), \qquad i = 1, \dots, n. \tag{4}$$

The numerical solver evolves each full initial state $Y_0^{(e,i)}$, and the observation operator yields

$$X_t^{(e,i)} = \mathcal{H}_t\big(Y_t^{(e,i)}\big). \tag{5}$$

The resulting endpoints $\{X_T^{(e,i)}\}_{i=1}^n$ thus form a reference ensemble for $P(X_T \mid X_0 = x_0^{(e)})$, as illustrated in Figure 1(a). The learned model receives only the observed states, whereas the simulator evolves the full states.

[1] For the heterogeneous-wave benchmark, the observed state additionally contains a finite-difference velocity proxy constructed from consecutive displacement observations; details are given in Appendix B.1.3. Appendix B.1.3 defines the proxy, and Appendix B.1.3 examines its effect.

Table 1: **Benchmark systems for conditional forecasting.** The observed forecast state and the withheld initial variation for each system. *Coarse* (resp. *fine*) refers to low-frequency (resp. high-frequency) modes.

System	Observed state	Hidden state
`euler`	Coarse vorticity field	Fine vorticity field
`sqg`	Coarse surface buoyancy	Fine surface buoyancy
`het-waves`	Coarse displacement, velocity proxy	Fine displacement, physical velocity
`nonlinear-kg`	Coarse displacement and velocity	Fine displacement and velocity
`passive-scalar`	Coarse concentration	Fine concentration

Each conditioning ensemble is assigned entirely to the training, validation, or test split, so all trajectories sharing the same observed initial condition belong to the same split. Evaluation is performed separately for each conditioning ensemble; pooling final states across different groups would instead test a marginal distribution, rather than the conditional law of interest.

We now present an overview of the different datasets; a detailed description can be found in Appendix B.1. All systems evolve on two-dimensional periodic domains.

Euler (`euler`) Euler equations describe the dynamics of adiabatic, zero-viscosity fluids. The scalar field of interest, vorticity, is transported by a velocity field determined by vorticity itself, giving rise to nonlinear dynamics. Nonlinear interactions couple spatial scales, allowing hidden fine initial components to influence the observed coarse evolution, making the system a test of conditional forecasting under uncertainty in the fine-scale initial state.

Surface-Quasi Geostrophic (`sqg`) Surface quasi-geostrophic (SQG) dynamics describe the evolution of a surface buoyancy field in a rapidly rotating, stratified fluid, with uniform potential vorticity (Held et al. (1995)). It is similar to the `euler` dataset, in that the scalar field influences the advecting velocity field, resulting in nonlinear dynamics; however, the velocity is more strongly dependent on small-scale scalar variations than in two-dimensional Euler flow. SQG therefore serves as a complementary test of how the coarse-fine scales coupling affects conditional forecasts.

Heterogeneous waves (`het-waves`) This dataset encompasses damped wave propagation in a heterogeneous medium, with a spatially varying wave speed fixed across the dataset. Spatial heterogeneity results in wave scattering, while damping attenuates the propagating disturbances; the dynamics remain linear, providing a test for conditional forecasting in a system with propagation and scattering without nonlinear interactions.

Nonlinear Klein-Gordon (`nonlinear-kg`) The nonlinear Klein–Gordon system describes a scalar field undergoing dispersive wave propagation with a local nonlinear restoring force. Its state comprises the field and its time derivative. Dispersion produces frequency-dependent propagation, while the nonlinearity couples Fourier modes, making the system a test of conditional forecasting under interacting wave dynamics.

Passive scalar (`passive-scalar`) This dataset describes the advection and diffusion of a scalar field in a steady, divergence-free cellular flow. The evolution is linear, with diffusion suppressing small-scale variations; with Gaussian hidden initial components, the conditional distributions remain Gaussian under the evolution law and the linear observation map.

3 Direct and Cascaded Conditional Forecasts

3.1 Methods and evaluation

Direct and cascaded forecasting. We compare two approaches for learning the conditional distribution $P(X_T \mid X_0 = x_0)$ defined in Section 2. Both are illustrated in Figure 1(b). Direct SI learns a single conditional stochastic interpolant from X_0 to X_T (Chen et al., 2024). Cascaded SI

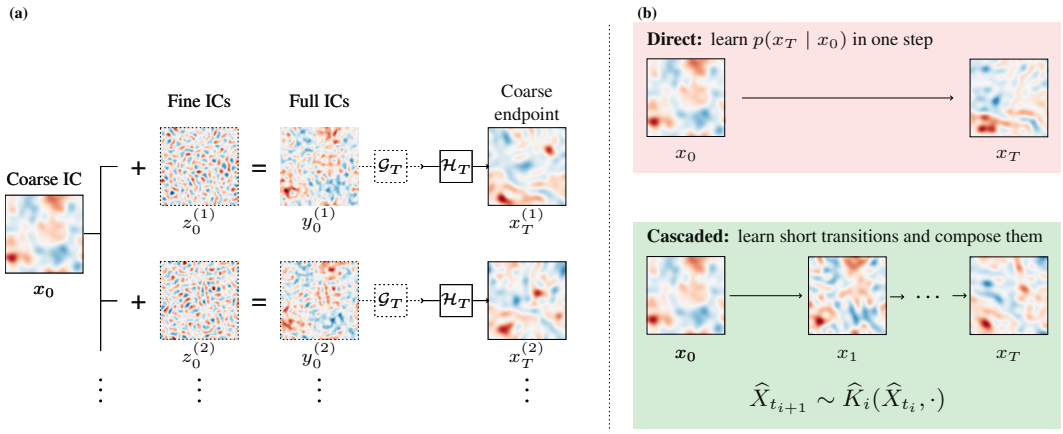

Figure 1: **Hierarchical coarse–fine data enable evaluation of posterior quality.** **(a)** Multiple unresolved fine initial states consistent with the same observed coarse state generate a reference conditional ensemble at the forecast time. **(b)** Direct forecasting learns one endpoint transition. Cascaded forecasting composes shorter learned transitions, passing each generated state to the next segment.

instead divides the same forecast horizon into S intervals, with $t_i = iT/S$, and learns the consecutive observed transitions from X_{t_i} to $X_{t_{i+1}}$. At $S = 1$, Cascaded SI reduces to Direct SI.

While Direct SI should sample from the correct conditional distribution in the ideal setting, Cascaded SI need not do so under partial observation. Let $\widehat{K}_i$ denote the learned conditional transition kernel from t_i to t_{i+1}. A single shared network represents all transitions and is conditioned on the current forecast state and the physical start and end times. At inference,

$$\widehat{X}_{t_{i+1}} \sim \widehat{K}_i(\widehat{X}_{t_i}, \cdot), \qquad \widehat{X}_0 = x_0, \quad t_i = iT/S. \tag{6}$$

Each transition therefore receives the generated state from the preceding one rather than the true intermediate state. Shorter transitions can be easier to learn, but repeatedly conditioning only on the current observed state can lose information about X_0 that is still relevant at time T. How these two effects depend on the cascade depth S is the focus of our analysis.

Together with Direct and Cascaded SI, we also include an Elucidated Diffusion Model (EDM) (Karras et al., 2022) as a diffusion-model baseline, trained conditionally on the same endpoint pairs (X_0, X_T) as Direct SI. Appendix B.2 gives the SI and EDM objectives and samplers.

Fixed training-data budget. A single trajectory provides one endpoint pair for Direct SI or EDM, but S consecutive transition pairs for Cascaded SI with depth S. Using the same number of trajectories would therefore give Cascaded SI S times more supervised training pairs. We instead fix the training-data budget across methods and depths. The benchmark contains 80 training conditioning ensembles with 200 candidate trajectories each. Direct SI and EDM use one endpoint pair from each of the 200 trajectories per ensemble, giving $B = 16{,}000$ supervised pairs in total. Cascaded SI at depth S uses only $200/S$ trajectories per ensemble, but takes all S consecutive transitions from each trajecotry, again giving $B = 16{,}000$ supervised pairs.

Every training run uses 30,000 optimizer updates. Increasing the cascade depth therefore does not provide additional supervised pairs or optimization steps; instead, Cascaded SI is trained on fewer distinct physical trajectories and incurs a higher inference cost because forecasting requires S sequential sampler integrations, one for each segment.

Model selection and sampling. Within each system, all methods use the same U-Net architecture. Hyperparameters, checkpoints, and cascade depth $S^\star \in \{1, 2, 5, 10, 20, 40\}$ are selected by conditional energy distance on validation ensembles (Székely & Rizzo, 2013; Gneiting & Raftery, 2007). Sampler resolution is calibrated on the same validation data; test ensembles are used only for final evaluation. Appendix B.2 specifies the searches, boundary extensions, and sampling tolerances.

Table 2: **Conditional forecasting performance.** Test csFID-10k, reported as mean ± SEM across ten held-out conditioning ensembles for one training seed. $S^\star$ is selected on validation data. For passive scalar, $S^\star = 1$, so the Direct SI and Cascaded SI entries correspond to the same fitted model. †Wave EDM and the NLKG Direct SI and EDM baselines retain unresolved hyperparameter-search boundary flags; see Appendix B.2.

	csFID-10k ↓			
System	Direct SI	EDM	Cascaded SI	$S^\star$
`euler`	57.26 ± 3.35	56.32 ± 3.74	$\mathbf{16.32 \pm 1.14}$	10
`sqg`	138.48 ± 23.09	118.62 ± 21.67	$\mathbf{51.24 \pm 8.32}$	10
`het-waves`	115.86 ± 6.76	$123.04 \pm 7.07^\dagger$	$\mathbf{36.66 \pm 1.78}$	5
`nonlinear-kg`	$98.48 \pm 6.18^\dagger$	$92.37 \pm 5.43^\dagger$	$\mathbf{16.24 \pm 1.55}$	20
`passive-scalar`	8.67 ± 0.28	$\mathbf{7.88 \pm 0.31}$	8.67 ± 0.28	1

Primary evaluation metric: csFID-10k. For each initial observation x_0 in the test set, we draw 10,000 endpoint samples from the learned forecaster and compare them with 10,000 reference endpoints generated by the physical simulator from the same conditioning state. Because a forecast can reproduce realistic structures while placing them at incorrect spatial locations, we use spatial FID (Nash et al., 2021) as our primary evaluation metric. In contrast to standard Heusel et al. (2017), it retains spatial information in the feature representation. We compute this metric, which we refer to as **csFID-10k**, separately for each conditioning ensemble and report the mean ± standard error of the mean (SEM) across the ten test ensembles. Appendix B.2.1 gives the metric construction, and Appendix B.3 provides a detailed validation of this choice.

3.2 Conditional endpoint distributions

Table 2 reports the test performance of the models selected entirely on validation data. For all three nonlinear systems (`euler`, `sqg`, `nonlinear-kg`), the validation-selected Cascaded SI achieves substantially lower csFID-10k than both Direct SI and EDM.

On Euler, Cascaded SI reduces csFID-10k by a factor of about 3.5 relative to both Direct SI and EDM at $S^\star = 10$, and outperforms both endpoint baselines in every test conditioning ensemble. The improvement extends to SQG and nonlinear Klein–Gordon, showing that it is not specific to Euler: it occurs both in nonlinear active-scalar dynamics and in nonlinear dispersive wave dynamics.

Heterogeneous waves show that nonlinearity is not required for cascading to help. The dynamics are linear, yet Cascaded SI reduces csFID-10k by more than a factor of three relative to both Direct SI and EDM at $S^\star = 5$. Passive scalar is also linear, but behaves very differently. Its conditional endpoint distribution is Gaussian with an affine endpoint mean, and validation selects $S^\star = 1$, so Cascaded SI reduces to Direct SI; EDM achieves the lowest score. Hence, whether cascading is beneficial cannot be explained simply by whether the underlying dynamics are linear or nonlinear.

Figure 2 shows one held-out test condition for Euler and heterogeneous waves. In both systems, Direct SI and EDM can generate individually plausible samples while their conditional mean and pointwise standard deviation differ substantially from the reference ensemble generated by the physical simulator. Cascaded SI matches both statistics much more closely. Additional qualitative comparisons for SQG, nonlinear Klein–Gordon, and passive scalar are provided in Appendix B.4; the first two show the same dramatic improvement under cascading, while passive scalar reflects the absence of a benefit at the validation-selected $S^\star = 1$. We next examine how conditional forecasting performance changes as the cascade depth S varies.

3.3 Cascading depth

Figure 3 shows how test csFID-10k changes with cascade depth S, with the model, checkpoint, and sampler resolution at every depth selected on validation data. The test curves are used only for evaluation and do not affect the selected $S^\star$.

In four of the five systems, moving from direct forecasting ($S = 1$) to a moderate cascade substantially improves the conditional forecast. The dependence on depth is nevertheless non-monotone. Euler and

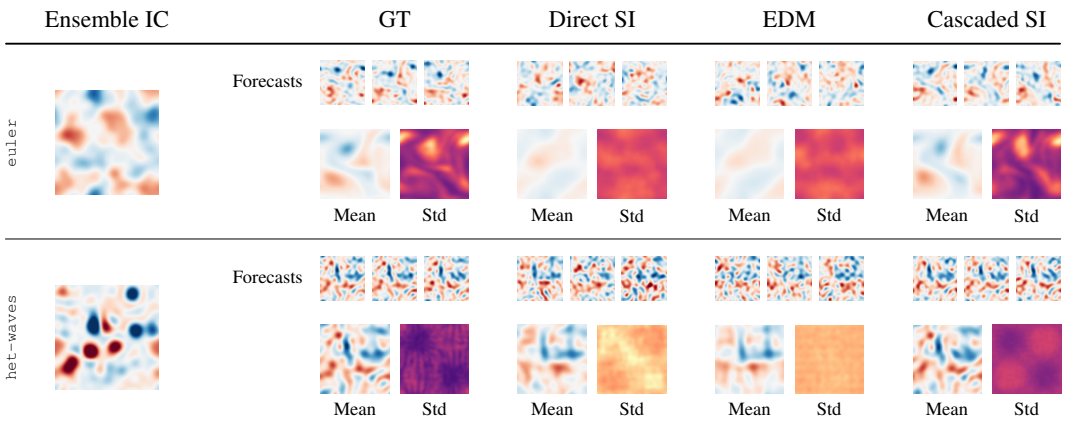

Figure 2: **Conditional endpoint distributions.** One held-out test condition for Euler and heterogeneous waves. GT denotes the simulator-generated ground truth. For each method, three endpoint samples, the conditional mean, and the pointwise standard deviation are shown, estimated from 10,000 samples.

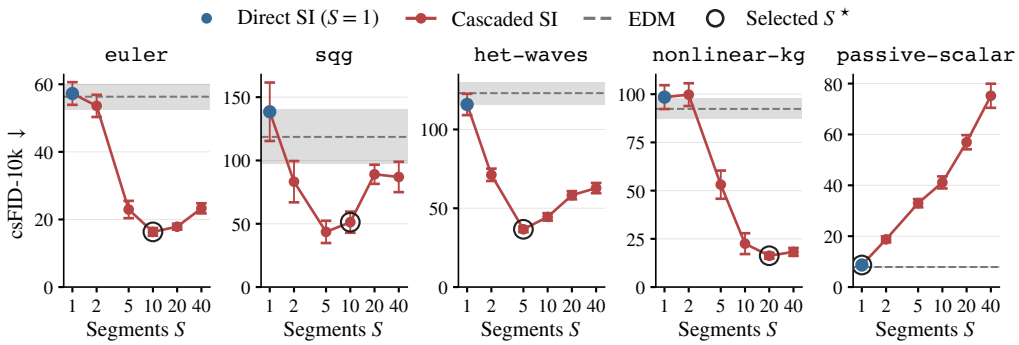

Figure 3: **Intermediate-depth cascades improve conditional forecasting in four of five systems.** Test csFID-10k versus cascade depth S; $S = 1$ is Direct SI, outlined markers show the validation-selected $S^\star$, and dashed lines show EDM. Error bars indicate ± 1 SEM across ten test conditioning ensembles. Numerical values are given in Appendix B.2.1.

heterogeneous waves improve sharply up to intermediate depths and then deteriorate as the horizon is divided further. SQG shows the same broad behavior, with its lowest test score at $S = 5$ and the independently validation-selected $S^\star = 10$. Nonlinear Klein–Gordon favors a deeper cascade, improving strongly up to $S = 20$ before slightly worsening at $S = 40$.

Passive scalar behaves differently: its test error increases monotonically once the forecast is cascaded, and validation accordingly selects $S^\star = 1$. Thus, neither direct forecasting nor increasingly fine factorization is uniformly optimal. The preferred depth is system-dependent and, in the systems where cascading helps, occurs at an intermediate number of segments. Section 4 examines one mechanism through which approximation error and hidden memory produce a finite optimal depth.

4 Understanding the Cascading Trade-off

The experiments show that shorter transitions can make the learned forecast more accurate, but this improvement does not continue indefinitely. We give an explicit stylized example in which the target endpoint distribution, the distribution produced by the generative model, and the distance between them can all be calculated exactly.

Proofs of the results in this section, together with a more general analysis of why cascading can help and why using too many transitions can hurt, are given in Appendix C.

4.1 A ROTATION–OU MODEL

We consider a two-dimensional observed state $Y_t \in \mathbb{R}^2$. Its deterministic dynamics are a rotation whose angular velocity is controlled by a conditioning variable $c \sim \mathrm{Unif}[-1, 1]$. In addition, it is driven by two sources of uncertainty: an unobserved Ornstein–Uhlenbeck (OU) process $H_t \in \mathbb{R}^2$, which is correlated over time, and independent white noise. The variable c is observed and retained throughout the forecast, while H_t is never observed.

Let $Y_0 \sim \mathcal{N}(0, s_0^2 I_2)$ and $H_0 \sim \mathcal{N}(0, I_2)$ be independent. For independent Brownian motions B, W, we consider

$$\begin{aligned} \mathrm{d}H_t &= -\kappa H_t \,\mathrm{d}t + \sqrt{2\kappa}\,\mathrm{d}B_t, \\ \mathrm{d}Y_t &= \omega c J Y_t \,\mathrm{d}t + \beta_\rho R(\omega c t) H_t \,\mathrm{d}t + \sigma_\rho \,\mathrm{d}W_t, \qquad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad R(\theta) = e^{\theta J}. \end{aligned} \tag{7}$$

The first term in the dynamics of Y_t rotates the observed state with angular velocity ωc. Thus, in the absence of noise, starting from $Y_0 = y$ gives

$$Y_T = R(\omega c T) y.$$

The parameter ω controls how strongly the endpoint depends on c: as ω increases, the map $c \mapsto R(\omega c T)y$ becomes increasingly oscillatory. The hidden process H_t is stationary and forgets its value gradually at rate κ. As a result, its values remain correlated across time, including across consecutive transitions.

We use ρ to control how much of the uncertainty at time T comes from these two sources, while keeping the target endpoint distribution fixed. For a fixed $q > 0$ and $0 \le \rho < 1$, define

$$V_H(t) = 2\left[\frac{t}{\kappa} - \frac{1 - e^{-\kappa t}}{\kappa^2}\right], \quad \beta_\rho^2 = \frac{\rho q}{V_H(T)}, \quad \sigma_\rho^2 = \frac{(1-\rho)q}{T}. \tag{8}$$

With this choice, the hidden OU process contributes $\rho q I_2$ to the covariance at time T, while the white noise contributes $(1-\rho)qI_2$. Their sum is always qI_2. Indeed, expressing the dynamics in the rotating coordinates $R(-\omega c t)Y_t$ gives

$$Y_T \mid (Y_0 = y, c) \sim \mathcal{N}\big(R(\omega c T)y, qI_2\big). \tag{9}$$

Thus, the target conditional distribution is the *same for every* ρ. Changing ρ changes how uncertainty is carried through time, but not the conditional distribution that must be reproduced at time T. We denote the Gaussian law in (9) by $K_T^{\mathrm{OU}}(y, c)$.

The reason this model is useful can already be seen from its conditional mean. Direct SI must approximate the complete map $(c, y) \longmapsto R(\omega c T)y$ in a single transition. Cascaded SI with depth S instead learns transitions containing the smaller rotation

$$R\left(\frac{\omega c T}{S}\right) y$$

and composes them. These shorter rotations are easier to approximate with a simple function class. There is, however, a competing effect. In the true dynamics, the same hidden OU process persists throughout the interval $[0, T]$. A cascaded forecast does not carry H_t from one transition to the next. Each new transition is conditioned only on the current observed state Y_{t_i} and on c. When $\rho > 0$, repeatedly conditioning in this way can therefore lose information carried by H_t, even if every short-time conditional transition is known exactly.

4.2 THE ERROR OF A SIMPLE SI MODEL

We now choose a simple SI model for which the generated endpoint distribution can be computed explicitly. This lets us measure the two effects described above.

Let $t_i = iT/S$, $\delta = T/S$. For the rotation–OU process, the exact conditional distribution over one such interval is Gaussian:

$$Y_{t_{i+1}} \mid (Y_{t_i} = y, c) \sim \mathcal{N}\big(r_i R(\omega c\delta)y, q_i I_2\big). \tag{10}$$

The scalar r_i determines how the current observation y enters the conditional mean, while q_i is the remaining conditional variance. Both are determined exactly by the rotation–OU process.

To isolate the error due to approximating the dependence on (c, y), we assume that the generative model knows r_i and q_i exactly. It only has to approximate the rotation $R(\omega c\delta)y$. For this problem, we consider a simple approximation class: polynomials $p(c, y)$ of total degree at most two. At each transition, the exact conditional mean $r_i R(\omega c\delta)y$ is replaced by $r_i p(c, y)$, while the conditional variance $q_i I_2$ is kept exact.

The explicit formulas of coefficients r_i and q_i are given in Appendix C.7. Set $\nu = \omega\delta$. Since $R(\nu c)y = \cos(\nu c)y + \sin(\nu c)Jy$, and y is centered and isotropic, the best approximation in squared error among polynomials $p(c, y)$ of total degree at most two has the form

$$p_\nu(c, y) = M_\nu(c)y.$$

Because p_ν is linear in y, only terms that are linear in y contribute to the optimal quadratic approximation. Since $p(c, y)$ has total degree at most two, their coefficients can depend at most linearly on c. We therefore project $\cos(\nu c)$ onto the constants and $\sin(\nu c)$ onto the linear functions of c. Since $c \sim \mathrm{Unif}[-1, 1]$, this gives

$$M_\nu(c) = j_0(\nu)I_2 + 3j_1(\nu)cJ, \qquad j_0(\nu) = \frac{\sin\nu}{\nu}, \qquad j_1(\nu) = \frac{\sin\nu - \nu\cos\nu}{\nu^2}. \tag{11}$$

For a single transition, the approximation depends only linearly on c. After S transitions, however, the conditional mean contains the product $M_\nu(c)^S$. Multiplying these matrices generates higher-order powers of c. A cascade can therefore represent a more complicated dependence on c, even though each individual transition uses the same quadratic approximation class.

Conditional on $(Y_0 = y, c)$, the generated covariance remains isotropic. We denote its endpoint value by $\widehat{q}_S(c)I_2$; its recurrence is given in Appendix C.7, Eq. (81).

Let

$$A_S = \prod_{i=0}^{S-1} r_i.$$

After S transitions, the generated endpoint distribution conditional on $(Y_0 = y, c)$ is therefore

$$\widehat{K}_S(y, c) = \mathcal{N}\big(A_S M_\nu(c)^S y, \widehat{q}_S(c)I_2\big). \tag{12}$$

We compare (12) with the target law (9) using the average *squared conditional Wasserstein distance*. Since both distributions are isotropic Gaussians,

$$\begin{aligned}\mathcal{E}_S(\rho) &= \mathbb{E}_{Y_0,c}\, W_2^2\big(K_T^{\mathrm{OU}}(Y_0, c), \widehat{K}_S(Y_0, c)\big) \\ &= \frac{1}{2}\int_{-1}^{1}\Big[s_0^2\|R(\omega Tc) - A_S M_\nu(c)^S\|_F^2 + 2\big(\sqrt{q} - \sqrt{\widehat{q}_S(c)}\big)^2\Big]\,\mathrm{d}c.\end{aligned} \tag{13}$$

The first term measures the error in the conditional mean. The second measures the error in the conditional covariance.

For Direct SI, $S = 1$. In this case the covariance is approximated exactly, but a single quadratic approximation cannot in general reproduce the dependence of the full rotation $R(\omega cT)y$ on c. In fact, if $\omega \neq 0$,

$$\mathcal{E}_1(\rho) = 2s_0^2\left[1 - j_0(\omega T)^2 - 3j_1(\omega T)^2\right] > 0. \tag{14}$$

Importantly, this error does not depend on ρ. We have therefore fixed both the target endpoint distribution and the error of Direct SI while varying how much information is carried through time by the hidden OU process.

4.3 WHY AN INTERMEDIATE DEPTH IS OPTIMAL

We now examine how the error $\mathcal{E}_S(\rho)$ changes with the cascade depth S.

If $\rho = 0$, the hidden OU process does not influence Y_t, and making the transitions shorter reduces the approximation error. In fact, the error tends to zero as $S \to \infty$.

When $\rho > 0$, the situation is different. As $S \to \infty$, repeated conditioning on the current observed state leads to a limiting distribution that differs from the target (9). We denote its squared conditional Wasserstein error by $\mathcal{C}_{\rho,\infty}$.

With $v_t = s_0^2 + \beta_\rho^2 V_H(t) + \sigma_\rho^2 t$, define

$$a_\rho(t) = \frac{\beta_\rho^2(1 - e^{-\kappa t})}{\kappa v_t}, \qquad A_\infty = \exp\left(\int_0^T a_\rho(t)\,\mathrm{d}t\right),$$
$$q_\infty = v_T - s_0^2 A_\infty^2, \qquad \mathcal{C}_{\rho,\infty} = 2s_0^2(A_\infty - 1)^2 + 2\big(\sqrt{q} - \sqrt{q_\infty}\big)^2. \tag{15}$$

The first term is the error in the conditional mean, while the second is the error in the conditional covariance. In particular, when $\rho > 0$, even a perfect fit to every short observed-state transition would not remove this limiting error: the problem comes from repeatedly conditioning on an incomplete state, not from approximating the individual short transitions.

The following theorem shows that the competition between the two effects can produce an optimal finite cascade depth.

Theorem 1. *Fix $T, \kappa, q, s_0 > 0$ and $\omega \neq 0$ in (7)–(13). For the SI model defined above, which approximates the transition mean with a quadratic function:*

1. *If $\rho = 0$, then $\mathcal{E}_S(0) = O(S^{-2})$ as $S \to \infty$.*
2. *For every fixed $0 < \rho < 1$, there is a $\Lambda_\rho > 0$ such that*
$$\mathcal{E}_S(\rho) = \mathcal{C}_{\rho,\infty} - \Lambda_\rho \frac{T}{S} + O(S^{-2}). \tag{16}$$
3. *There is a $\rho_* > 0$ such that, for each $0 < \rho < \rho_*$, a global minimizer over positive integer depths exists at a finite $S^\star \geq 2$, and*
$$\mathcal{E}_{S^\star}(\rho) < \min\left\{\mathcal{E}_1(\rho), \mathcal{C}_{\rho,\infty}\right\}. \tag{17}$$

The theorem shows that these two effects can produce a finite optimal cascade depth. When $\rho = 0$, the approximation error vanishes as $S \to \infty$. When $\rho > 0$, arbitrarily fine cascading instead retains a nonzero limiting error. For sufficiently small but nonzero ρ, the minimum therefore occurs at a finite depth $S^\star \geq 2$, with lower error than both Direct SI and the infinitely deep cascade. In Appendix C, we further prove that $S^\star(\rho) \asymp \rho^{-2}$ as $\rho \downarrow 0$. Thus, as the contribution of the hidden OU process becomes weaker, the optimal cascade can become deeper.

5 CONCLUSION

We studied one-step and cascaded stochastic interpolants for conditional forecasting under partial observation, focusing on how forecast error changes with cascade depth and across systems. Our analysis identifies two competing effects: approximation error from learning longer transitions, and Markovianity error from memory loss as the cascade becomes deeper. We show that cascaded SIs perform better on the more involved systems, while the optimal depth varies across systems. We also release a benchmark with reference ensembles for testing whether a forecaster recovers the conditional distribution of future states.

Limitations. Cascading increases inference cost because each additional stage requires another sampler integration. The benefit of cascading may also depend on model capacity. Finally, the theory establishes a mechanism for a finite optimal depth but does not predict the depth selected in the PDE experiments.

AI use statement. We used generative AI tools throughout this project to assist with research ideation and experimental design, code development and debugging, interpretation of results, retrieval and discovery of related work, and drafting and polishing the manuscript. AI tools also assisted in designing and implementing parts of the pipeline used to generate the synthetic benchmark datasets; the reported data themselves were produced by the numerical physical simulators described in Appendix B.1. For the theoretical component, generative AI tools were used to suggest mathematical formulations, intermediate claims, proof ideas, and draft proof text. All mathematical claims and proofs were checked and revised by the authors, all cited literature was verified against the original sources, and AI-assisted code and experimental outputs were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

Reproducibility statement. Section 2 and Appendix B.1 specify the benchmark construction, observation operators, numerical solvers, and data splits. Section 3.1 and Appendix B.2 describe the fixed-budget training protocol, model architecture, objectives, hyperparameter searches, validation-based model and depth selection, and sampler calibration used for Table 2 and Figure 3. Appendices B.2.1 and B.3 specify the evaluation metrics, and Appendix C contains the assumptions and proofs for the theoretical results. Exact experiment configurations, code, and benchmark datasets will be shared anonymously during discussion phase and released publicly upon acceptance.

REFERENCES

Ryan Abernathey et al. pyqg: Python quasigeostrophic model, 2022. URL https://zenodo.org/records/6563667.

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Gerard Brunick and Steven Shreve. Mimicking an itô process by a solution of a stochastic differential equation. *The Annals of Applied Probability*, 23(4), August 2013. ISSN 1050-5164. doi: 10.1214/12-aap881. URL http://dx.doi.org/10.1214/12-AAP881.

Keaton J. Burns, Geoffrey M. Vasil, Jeffrey S. Oishi, Daniel Lecoanet, and Benjamin P. Brown. Dedalus: A flexible framework for numerical simulations with spectral methods. *Phys. Rev. Res.*, 2:023068, Apr 2020. doi: 10.1103/PhysRevResearch.2.023068. URL https://link.aps.org/doi/10.1103/PhysRevResearch.2.023068.

Yifan Chen, Mark Goldstein, Mengjian Hua, Michael Samuel Albergo, Nicholas Matthew Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and föllmer processes. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6728–6756. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24n.html.

Alexandre J Chorin, Ole H Hald, and Raz Kupferman. Optimal prediction and the mori–zwanzig representation of irreversible processes. *Proceedings of the National Academy of Sciences*, 97(7): 2968–2973, 2000.

Alexandre J Chorin, Ole H Hald, and Raz Kupferman. Optimal prediction with memory. *Physica D: Nonlinear Phenomena*, 166(3-4):239–257, 2002.

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. URL https://doi.org/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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. doi: 10.52202/068431-1926. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf.

Georg Kohl, Li-Wei Chen, and Nils Thuerey. Benchmarking autoregressive conditional diffusion models for turbulent flow simulation. *Neural Networks*, 199:108641, 2026. doi: 10.1016/j.neunet.2026.108641.

Lizao Li, Rob Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Seeds: Emulation of weather forecast ensembles with diffusion models. *arXiv preprint arXiv:2306.14066*, 2023.

Edward N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2): 130–148, March 1963. doi: 10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2.

Edward N. Lorenz. The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307, 1969. doi: https://doi.org/10.1111/j.2153-3490.1969.tb00444.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1969.tb00444.x.

Chao Ma, Jianchun Wang, and Weinan E. Model reduction with memory and the machine learning of dynamical systems. *Communications in Computational Physics*, 25(4):947–962, 2019. doi: 10.4208/cicp.OA-2018-0269. URL https://doi.org/10.4208/cicp.OA-2018-0269.

Nikolaj T. Mücke and Benjamin Sanderse. Physics-aware generative models for turbulent fluid flows through energy-consistent stochastic interpolants. *Computers & Fluids*, 314:107098, 2026. ISSN 0045-7930. doi: https://doi.org/10.1016/j.compfluid.2026.107098. URL https://www.sciencedirect.com/science/article/pii/S0045793026001404.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter Battaglia. Generating images with sparse representations. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7958–7968. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nash21a.html.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation, 2022. URL https://arxiv.org/abs/2104.11222.

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

Tobias Rohner and Siddhartha Mishra. Efficient computation of large-scale statistical solutions to incompressible fluid flows. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706394. doi: 10.1145/3659914.3659922. URL https://doi.org/10.1145/3659914.3659922.

François Rozet and Gilles Louppe. Score-based data assimilation, 2023. URL https://arxiv.org/abs/2306.10574.

Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. DYffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=WRGldGm5Hz.

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

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2013.03.018. URL https://www.sciencedirect.com/science/article/pii/S0378375813000633.

L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, Apr 2016. URL http://arxiv.org/abs/1511.01844.

A Related work

Generative models for probabilistic forecasting. Recent work has used diffusion models and stochastic interpolants to represent distributions over future states rather than produce a single deterministic forecast. Chen et al. (Chen et al., 2024) formulate forecasting directly as sampling from the conditional distribution of a future state given the current state, using stochastic interpolants to define a generative path between them. Related diffusion-based approaches have been developed for weather forecasting, including SEEDS (Li et al., 2023) and GenCast (Price et al., 2025), and for complex physical systems and turbulent flows (Kohl et al., 2026; Mücke & Sanderse, 2026). Our work considers the same probabilistic forecasting objective, but focuses on how the physical time interval used for learning affects the accuracy of the resulting conditional distribution.

Using intermediate times and composing forecasts. Several forecasting methods make use of intermediate physical states or generate trajectories by repeatedly applying learned transitions. GenCast, for example, learns a conditional distribution over a fixed physical time step and samples successive states autoregressively (Price et al., 2025). Conditional diffusion models have similarly been applied autoregressively to turbulent-flow simulation (Kohl et al., 2026). DYffusion (Rühling Cachay et al., 2023) is particularly related to our setting, since it uses intermediate physical times to build a dynamics-informed diffusion process. These works show that intermediate temporal information can be useful for generative forecasting; we instead compare, for a fixed final forecast time T, learning $P(X_T \mid X_0)$ directly with learning and composing transitions over intervals of length T/S. We then vary S systematically while controlling the amount of supervised training data. This exposes a non-monotone dependence on cascade depth: making the transitions shorter can improve the forecast, but dividing the interval increasingly finely need not continue to help.

Partial observation and memory. The second effect in our analysis is specific to forecasting from incomplete observations. Projection of a dynamical system onto a subset of its variables generally produces dynamics with memory: unresolved variables can retain information about the past that is not contained in the current observation. This phenomenon is classical in optimal prediction and the Mori–Zwanzig formalism (Chorin et al., 2000; 2002). It has a direct consequence for cascaded probabilistic forecasts. Even if each short transition $P(X_{t_{i+1}} \mid X_{t_i})$ were known exactly, their composition need not equal $P(X_T \mid X_0)$ when the observed process is not Markov. Our theoretical analysis makes this loss of information explicit and studies its competition with the approximation error of learning longer transitions. In the rotation–OU model, these two effects produce a finite optimal cascade depth.

Evaluating conditional forecast distributions. Evaluating generative models is difficult because realistic individual samples do not by themselves imply that the correct distribution has been learned (Theis et al., 2016). Probabilistic forecasting methods therefore use ensemble-based scores, distributional statistics, and diagnostics of forecast uncertainty (Li et al., 2023; Price et al., 2025; Kohl et al., 2026). Our benchmark is designed specifically for conditional evaluation under partial observation. For each observed initial state x_0, we generate many full initial states that share x_0 but differ in their hidden components, and evolve each of them with the physical simulator. The resulting reference ensemble gives direct samples from $P(X_T \mid X_0 = x_0)$. This allows us to distinguish a model that generates plausible states from one that recovers the conditional distribution associated with the particular observation supplied to the forecaster.

B Experiments

B.1 Dataset construction and numerical solvers

For each observed initial state, we simulate several trajectories by varying the hidden initial conditions. We call this set of trajectories an *ensemble*. All trajectories in an ensemble therefore start from the same observation but can reach different final states (see Fig. 4).

The benchmark uses 80 ensembles for training, 10 for validation, and 10 for testing, with different observed initial states in each split. Each test ensemble provides $10{,}000$ simulated endpoints for comparison with the model's predictions.

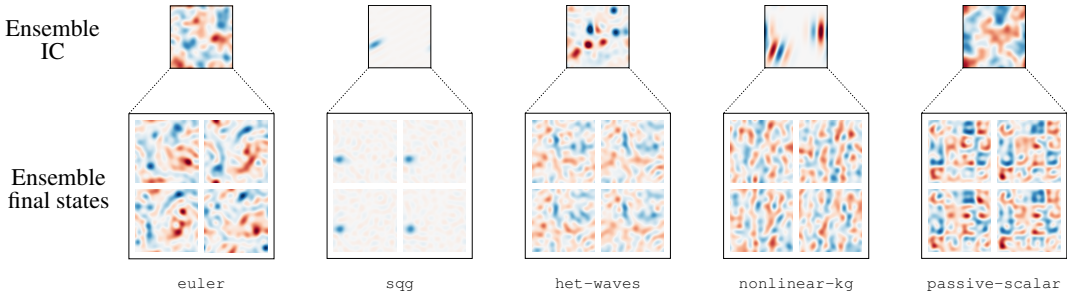

Figure 4: **Representative conditional ensembles from the benchmark.** For each system, the top panel shows one observed initial condition, and the four lower panels show example final states. Each dataset contains many such initial-condition ensembles.

The model observes only Fourier modes $0 < |k| \leq 8$, represented by 196 real coordinates per scalar field. Table 3 summarizes the simulation settings and observed variables.

Table 3: **Benchmark dataset specifications.** Parentheses give the number of real coordinates observed by the model. Each benchmark trajectory contains 41 equally spaced snapshots. Horizons are reported in physical time.

Dataset	Simulation grid	Horizon T	Observed variables
`euler`	1024^2	1	Vorticity (196)
`sqg`	256^2	20	Buoyancy (196)
`het-waves`	512^2	1	Displacement and velocity proxy (392)
`nonlinear-kg`	128^2	12	Displacement and velocity (392)
`passive-scalar`	512^2	1	Concentration (196)

B.1.1 Two-dimensional Euler

On the unit periodic square, vorticity evolves according to

$$\partial_t \omega + u \cdot \nabla \omega = 0, \qquad u = \nabla^\perp \Delta^{-1} \omega. \tag{18}$$

Here $\nabla^\perp = (-\partial_y, \partial_x)$. The initial field has independently drawn low and high Fourier bands. Filtered Gaussian white noise with spectral amplitude proportional to $|k|^{-2}$ on $1 < |k| \leq 8$ determines the condition; an independent draw with amplitude proportional to $|k|^{-3}$ on $8 < |k| \leq 64$ is added for each member. The bands are normalized separately to velocity RMS values 1 and 0.2.

We use the double-precision *azeban* pseudospectral solver Rohner & Mishra (2024), with a 1024^2 grid, 1536^2 padding, smooth-cutoff spectral viscosity ($s = 1$, $\epsilon = 0.4$), and adaptive SSP–RK3 integration at CFL 0.5. The 41 states are archived every 0.025 through $T = 1$. A comparison of 1024^2 and 2048^2 simulations for one initial-condition ensemble supports distributional convergence of the observed endpoint states at the benchmark resolution.

B.1.2 Surface quasi-geostrophic flow

On the periodic square $[0, 2\pi)^2$, surface buoyancy obeys

$$\partial_t b + J(\psi, b) = 0, \qquad \widehat{\psi}(k) = \frac{\widehat{b}(k)}{|k|} \quad (k \neq 0). \tag{19}$$

Here $J(f, g) = \partial_x f \, \partial_y g - \partial_y f \, \partial_x g$. Each condition contains one to four signed elliptical vortices, with uniform centers and orientations, amplitude magnitudes in $[0.7, 1.2]$, aspect ratios in $[1.5, 4]$, and widths in $[2\pi/9, 2\pi/6]$. Centers are separated by at least 1.2. Projection onto $0 < |k| \leq 8$ defines the observed initial state. Each ensemble member adds a real Gaussian Fourier field on $8 < |k| \leq 32$, with coefficient variance proportional to $|k|^{-5/3}$, normalized to spatial RMS 0.10.

We use `pyqg.SQGModel` (Abernathey et al., 2022) (version `0.7.3.dev23`, commit `9abad33`), with $f_0 = N = H = 1$, $\beta = 0$, and no linear drag. The solver uses a 256^2 Fourier grid, third-order Adams–Bashforth stepping with $\Delta t = 0.005$, and its built-in high-wavenumber filter, which dissipates buoyancy variance. We archive 41 states every 0.5 through $T = 20$ and rescale forecast time to $[0, 1]$. A comparison of 256^2 and 512^2 simulations, with the time step halved at the finer resolution, supports distributional convergence of the observed endpoint states.

B.1.3 Heterogeneous waves

Displacement u and velocity v evolve on the unit periodic square:

$$\partial_t u = v, \qquad \partial_t v = \nabla \cdot (c^2 \nabla u) - 0.2v, \qquad c^2(x, y) = 1 + 0.5 \sin(2\pi x) \sin(2\pi y). \tag{20}$$

The medium is fixed. Each condition combines an anisotropic Gaussian background (correlation lengths 0.06 and 0.03, scale 0.2) with 18 random boxes, triangles, or disks. Shapes overwrite earlier values where they overlap; centers and orientations are uniform, sizes lie in $[0.08, 0.20]$, and amplitudes in $[-3, 3]$. Smoothing with Gaussian width $2/128$ and projection onto $0 < |k| \leq 8$ define the shared initial displacement.

Each ensemble member adds masked Gaussian noise projected onto $8 < |k| \leq 16$. The mask depends on the condition's shape interiors; before masking, the band filter gives expected spatial mean square 3.3^2. Initial velocity consists of three Gaussian kicks with uniform centers, width 0.03, and independent strengths $\mathcal{N}(0, 165^2)$, projected onto $|k| > 8$. Both perturbations therefore leave the observed initial state unchanged.

A double-precision PyTorch solver uses centered spatial differences and RK4 on a 512^2 grid with $\Delta t = 2.5 \times 10^{-4}$. We archive 41 states every 0.025 through $T = 1$. The observed state combines coarse displacement with the velocity proxy

$$\hat{v}_{t_n} = P_8 \frac{u_{t_n} - u_{t_{n-1}}}{0.025} \quad (n > 0), \qquad \hat{v}_0 = 0, \tag{21}$$

where P_8 extracts the 196 real observed Fourier coordinates. Training targets use this proxy; during generation, the model predicts both channels and passes them to the next segment. An independent `py-pde` implementation (Abernathey et al., 2022) verifies the spatial operator. Time-step and grid-refinement checks support numerical accuracy of the observed states.

Wave state diagnostic. We compare deterministic regressors using displacement alone or the augmented state $(u, \hat{v})$. They have nearly equal parameter counts and use mean-squared error, 8,000 training updates, and learning rates 3×10^{-4} and 10^{-3}, with checkpoint selection on eight development conditions. Each depth receives $B = 16{,}000$ supervised transitions. We recursively compose the predictions and measure their correlation with the simulated conditional endpoint mean on ten validation conditions (Table 4). The supplied initial proxy uses a forward difference and therefore gives additional initial information. Setting that proxy to zero at inference tests the same fitted augmented model without this information.

Table 4: **Wave state diagnostic.** Mean correlation with the true conditional endpoint mean over ten validation conditions. Both augmented columns use the same fitted model. The benchmark initializes the proxy to zero and subsequently carries generated values.

		Displacement and velocity proxy	
S	Displacement only	Initial proxy supplied	Initial proxy zero
1	0.731	0.760	0.647
5	0.108	0.949	0.871
10	0.094	0.953	0.888
20	0.024	0.946	0.848

B.1.4 Nonlinear Klein–Gordon

On the periodic square $[0, 2\pi)^2$, displacement and velocity obey the defocusing nonlinear Klein–Gordon equation:

$$\partial_t u = v, \qquad \partial_t v = \Delta u - u - 64u^3. \tag{22}$$

Each condition consists of two travelling Gaussian wave packets, projected onto $0 < |k| \leq 8$ in both fields. Centers, directions, and phases are uniform; amplitudes lie in $[0.5, 1]$, carrier wavenumber magnitudes in $[3, 5]$, and the envelope width is 0.8. Each ensemble member adds hidden displacement and velocity on $8 < |k| \leq 24$. Their Fourier coefficients are proportional to $g_k/\sqrt{1 + |k|^2}$ and g'_k, respectively, with independent complex Gaussian draws and reality enforced. Both hidden fields are scaled by the same member-specific factor to give displacement RMS 0.20. We observe 196 real Fourier coordinates of u and 196 of $v/4$.

Dedalus 3.0.5 (Burns et al., 2020) uses a 128^2 Fourier grid, factor-two dealiasing for the cubic term, and the third-order IMEX scheme RK443 with $\Delta t = 0.0025$. States are archived every 0.1; retaining every third state gives 41 snapshots through $T = 12$. Spatial and time-step refinement checks support numerical convergence.

B.1.5 Passive scalar

On the unit periodic square, concentration evolves under a fixed flow:

$$\partial_t q + v \cdot \nabla q = \kappa \Delta q, \tag{23}$$

$$v(x, y) = \big(V \sin(2\pi K x) \cos(2\pi K y), -V \cos(2\pi K x) \sin(2\pi K y)\big), \tag{24}$$

with $K = 2$, $V = 0.5$, and $\kappa = 6.184154 \times 10^{-6}$. Each condition is drawn on a 256^2 generator grid from Gaussian white noise filtered with amplitude $|k|^{-2}$ on $1 < |k| \leq 8$ and normalized to spatial RMS 1. Each ensemble member adds independent Gaussian noise filtered with amplitude $|k|^{-3}$ on $8 < |k| \leq 64$. A fixed multiplier gives expected spatial mean square 1.14491, without normalizing individual members. The hidden field is therefore Gaussian, and linear evolution preserves a Gaussian conditional endpoint distribution.

A double-precision Fourier–Galerkin solver uses integrating-factor RK4 on a 512^2 grid with 768^2 padding and $\Delta t = 0.005/6$. We archive states every 0.005 through $T = 1$ and retain every fifth state for the 41-snapshot benchmark. Independent simulations of five trajectories with Dedalus (Burns et al., 2020) agree with the observed dataset states as the Dedalus time step is refined.

B.2 Training, selection, and numerical sampling

Supervision and architecture. The matched-budget comparison uses 80 canonical training conditions and 200 candidate trajectories per condition. At depth $S \in \{1, 2, 5, 10, 20, 40\}$, $200/S$ trajectories supply all S consecutive transitions, keeping the budget at $B = 16{,}000$. Direct SI is the $S = 1$ member of this family. Normalization is fitted on training data and shared across depths and methods.

The U-Net acts on 32×32 reconstructions of the observed modes, with channel widths $64, 128, 256$, circular 3×3 convolutions, SiLU activations, two average-pooling stages, and skip connections. Its output is projected back onto the observed modes. SI receives the interpolated state, conditioning state, auxiliary time, and physical segment start/end times. Euler uses 1,515,265 parameters. Passive scalar and waves also receive spatial-coordinate features because their flow or medium is spatially inhomogeneous. The wave model predicts both displacement and velocity-proxy channels.

Objectives and samplers. For a normalized pair (a, b), draw $\tau \sim \mathrm{Unif}[0, 1]$ and $z \sim \mathcal{N}(0, I)$, and set

$$I_\tau = (1 - \tau)a + \tau b + \sigma(1 - \tau)\sqrt{\tau}\, z, \qquad R_\tau = b - a - \sigma\sqrt{\tau}\, z. \tag{25}$$

The drift regresses R_τ from $(I_\tau, \tau, a, t_i, t_{i+1})$ using the sum of $\frac{1}{2}\|b_\theta\|^2 - b_\theta \cdot R_\tau$ for antithetic draws $z, -z$. Euler–Maruyama samples

$$dZ_\tau = b_\theta(Z_\tau, \tau, a, t_i, t_{i+1})\, d\tau + \sigma(1 - \tau)\, dW_\tau, \qquad Z_0 = a. \tag{26}$$

At the next physical segment, a is replaced by the generated endpoint. SI gradient norms are clipped at 1.

EDM uses the conditional denoising objective and preconditioning of Karras et al. (2022), with $\sigma_{\mathrm{data}} = 1$, $P_{\mathrm{std}} = 1.2$, and a searched log-noise location P_{mean}. Its deterministic Heun sampler uses $\sigma_{\min} = 0.002$, $\sigma_{\max} = 80$, and schedule exponent 7; n numerical steps require $2n - 1$ network evaluations. EDM uses no gradient clipping or learning-rate ramp-up.

Optimization and selection. Models train for 30,000 Adam updates with batch size 256, constant learning rate, no weight decay, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 10^{-8}$. Training uses float32, without mixed precision or TF32. Exponential-moving-average weights have a 500,000-example half-life and ramp-up ratio 0.05 and are used for validation and test. Checkpoints are scored every 1,000 updates and at updates 100, 200, 300, 500, and 750.

Checkpoint and hyperparameter selection minimize the mean validation energy-distance U-statistic (Székely & Rizzo, 2013); $S^\star$ minimizes it across depths. For generated states $A = \{a_i\}_{i=1}^n$ and true states $R = \{r_j\}_{j=1}^m$, we use

$$\widehat{\mathrm{ED}}_U(A, R) = \frac{2}{nm}\sum_{i,j}\|a_i - r_j\| - \frac{1}{n(n-1)}\sum_{i\neq i'}\|a_i - a_{i'}\| - \frac{1}{m(m-1)}\sum_{j\neq j'}\|r_j - r_{j'}\|. \quad (27)$$

For sampler-resolution calibration, we instead use the corresponding V-statistic

$$\widehat{\mathrm{ED}}_V(A, R) = \frac{2}{nm}\sum_{i,j}\|a_i - r_j\| - \frac{1}{n^2}\sum_{i,i'}\|a_i - a_{i'}\| - \frac{1}{m^2}\sum_{j,j'}\|r_j - r_{j'}\|, \quad (28)$$

which includes self-pairs and is more stable for relative convergence checks near zero.

Selection uses 100 true and 100 generated states per condition in normalized model coordinates, with float64 accumulation. Exact ties select the earlier checkpoint. One training seed is used; sampling random numbers are shared across configurations within each condition.

Initial learning rates are $\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$, SI noise scales are $\{0.25, 0.5, 1, 2, 4, 8, 16, 32\}$, and EDM uses $P_{\mathrm{mean}} \in \{-3.2, -2.8, \ldots, -0.4\}$. One boundary-extension round is allowed. At the lower or upper edge, learning rates extend by $\{3\times10^{-5}, 10^{-5}\}$ or $\{3\times10^{-3}, 10^{-2}\}$, SI noise by $\{0.125, 0.0625\}$ or $\{64, 128\}$, and P_{mean} by $\{-3.6, -4\}$ or $\{0, 0.4\}$. Unresolved search boundaries retain a flag.

Sampler resolution. For each depth and method, a fixed probe is evaluated using $\widehat{\mathrm{ED}}_V$ on the step ladder $20, 40, 80, 160, 320, 640$. The upper resolution of the first doubling with relative score change below 5% is retained. Probes use the final checkpoint at learning rate 3×10^{-4}, with SI noise 8 or EDM $P_{\mathrm{mean}} = -1.6$. The selected model is checked again at twice its assigned resolution. A failed check permits one resolution escalation and validation reselection; a repeated failure is flagged. Sampler calibration is therefore separate from model selection by $\widehat{\mathrm{ED}}_U$.

Selected configurations. Table 5 reports the frozen selections for all five systems. Sampler steps are reported per segment; network function evaluations (NFE) count the full forecast.

B.2.1 Conditional distributional evaluation

csFID-10k reconstructs unnormalized coarse fields on a 128×128 grid. Signed values are clipped symmetrically at the training-fitted 0.9999 quantile of absolute amplitude, linearly mapped to uint8, and replicated into three identical channels. There is no colormap or per-image normalization. For heterogeneous waves and nonlinear Klein–Gordon, csFID evaluates displacement only, although the forecasting state also includes a velocity channel. Images undergo one bicubic resize to 299×299 in `clean-fid` 0.1.35 (Parmar et al., 2022). The first seven channels of Inception's `mixed_6/conv` 17×17 activation map are flattened in spatial–channel order, giving 2023 features without spatial pooling. Pooled cFID uses the same preprocessing and Inception weights with 2048-dimensional `pool3` features.

For each condition, the Fréchet distance is computed between the feature means and covariances of 10,000 true and 10,000 generated endpoints. The conditional score averages these distances with equal condition weights. Reported uncertainty is the sample standard deviation of condition-level scores divided by the square root of the number of conditions. For the main comparison, this is SEM over ten test conditions for one training seed. Secondary state-space diagnostics use fixed subsets of 200 true and 200 generated endpoints unless otherwise stated.

Numerical values from the depth sweep presented in Section 3.3 are reported in Table 6.

Table 5: Selected configurations and sampling cost.

System	Method	S	Learning rate	Noise parameter	Update	Sampler steps	NFE
Euler	Direct SI	1	10^{-4}	$\sigma = 8$	5,000	40	40
	EDM	1	10^{-3}	$P_{\text{mean}} = -1.6$	300	40	79
	Cascaded SI	10	3×10^{-4}	$\sigma = 8$	17,000	80	800
`sqg`	Direct SI	1	10^{-4}	$\sigma = 16$	4,000	40	40
	EDM	1	3×10^{-4}	$P_{\text{mean}} = -3.2$	4,000	40	79
	Cascaded SI	10	10^{-3}	$\sigma = 16$	11,000	160	1,600
Waves	Direct SI	1	10^{-3}	$\sigma = 32$	6,000	40	40
	EDM	1	3×10^{-5}	$P_{\text{mean}} = -0.4$	3,000	40	79
	Cascaded SI	5	3×10^{-3}	$\sigma = 2$	30,000	40	200
`nonlinear-kg`	Direct SI	1	10^{-3}	$\sigma = 128$	15,000	40	40
	EDM	1	10^{-3}	$P_{\text{mean}} = -4$	17,000	40	79
	Cascaded SI	20	10^{-4}	$\sigma = 1$	30,000	320	6,400
Passive scalar	Direct/Cascaded SI	1	3×10^{-3}	$\sigma = 1$	6,000	40	40
	EDM	1	10^{-3}	$P_{\text{mean}} = -1.2$	16,000	40	79

Table 6: **Conditional forecasting performance across cascade depths.** Test csFID-10k, mean $\pm$ SEM across ten conditioning ensembles for one training seed. Bold indicates the lowest SI score within each system; EDM is shown for comparison. For SQG at $S = 20$, one nonfinite generated endpoint was excluded from one conditioning ensemble, leaving 9,999 generated samples for that score.

Segments S	`euler`	`sqg`	`het-waves`	`nonlinear-kg`	`passive-scalar`
1 (Direct SI)	57.26 ± 3.35	138.48 ± 23.09	115.86 ± 6.76	98.48 ± 6.18	$\mathbf{8.67 \pm 0.28}$
2	53.60 ± 3.28	83.25 ± 16.31	71.23 ± 3.91	99.75 ± 5.89	18.75 ± 1.17
5	22.94 ± 2.55	$\mathbf{43.57 \pm 8.79}$	$\mathbf{36.66 \pm 1.78}$	53.10 ± 7.30	32.97 ± 1.61
10	$\mathbf{16.32 \pm 1.14}$	51.24 ± 8.32	44.51 ± 2.34	22.50 ± 5.40	41.14 ± 2.36
20	17.86 ± 0.87	89.08 ± 7.56	58.28 ± 2.58	$\mathbf{16.24 \pm 1.55}$	56.95 ± 2.79
40	23.29 ± 1.50	86.98 ± 12.00	62.81 ± 3.25	18.23 ± 2.05	75.21 ± 4.74
EDM	56.32 ± 3.74	118.62 ± 21.67	123.04 ± 7.07	92.37 ± 5.43	7.88 ± 0.31

B.3 Validation of the spatial feature metric

The conditioning field fixes spatial registration: realistic structures at incorrect locations can still constitute an incorrect conditional forecast. We therefore use spatial FID (sFID) (Nash et al., 2021), which retains intermediate spatial Inception features, alongside pooled FID as a diagnostic. The representations have 2023 and 2048 coordinates, respectively; all image preprocessing and Inception weights are shared. Appendix B.2.1 specifies their construction and conditional aggregation.

Agreement with state-space diagnostics. On Euler, pooled cFID separates Direct SI and EDM more strongly than csFID does (Table 7). The paired EDM-minus-Direct difference over the ten test conditions is -9.35 ± 1.31 for cFID and -0.94 ± 0.90 for csFID (mean $\pm$ SEM). The closer csFID scores agree with the similar sliced Wasserstein distances, conditional-mean errors, and spread ratios of these two endpoint models. Cascaded SI has the lowest cFID, csFID, sliced Wasserstein distance, and conditional-mean error.

Sensitivity to spatial rearrangement. We split each true 10,000-member endpoint bank into disjoint sets A, B of 5000 samples and compare $d(A, B)$ with $d(F(A), B)$. Transformations act exactly on Fourier coefficients, before image construction. A translation of d pixels shifts both coordinates by d on the 128×128 reconstruction. Euler averages ten resplits within each condition; passive scalar uses one split. Table 8 reports ratios of across-condition means to the corresponding truth-to-truth mean. On Euler, spatial features respond more strongly relative to this baseline. Pooled features also respond strongly on passive scalar, so their sensitivity to spatial rearrangement is

Table 7: **Euler metric comparison.** Mean ± SEM over ten test conditions. Feature metrics use 10,000 true and generated endpoints per condition. State-space diagnostics use 200 of each in normalized Fourier coordinates: sliced W_2 uses 100 fixed random projections, mean error is the Euclidean distance between ensemble means, and spread is the ratio of mean pairwise distances (generated/true, excluding self-pairs). Lower is better except for spread, whose target is one.

Metric	Cascaded SI ($S = 10$)	Direct SI	EDM
cFID	8.96 ± 0.89	24.30 ± 2.08	14.94 ± 1.21
csFID	16.32 ± 1.14	57.26 ± 3.35	56.32 ± 3.74
Sliced W_2	0.188 ± 0.012	0.390 ± 0.026	0.389 ± 0.025
Conditional-mean error	1.94 ± 0.17	5.22 ± 0.34	5.17 ± 0.32
Spread ratio	0.931 ± 0.013	1.060 ± 0.023	1.065 ± 0.021

dataset-dependent. These probes measure sensitivity, not a calibrated distance between spatial transformations. Resplits reuse the same bank and are not independent ensemble replicates.

Table 8: **Response to spatial perturbations.** Each entry is the ratio of the perturbed score to the truth-to-truth score, using means over ten conditioning states. Passive scalar has only two distinct outputs because Cascaded SI selects $S^\star = 1$.

	Euler		Passive scalar	
Transformation	cFID	csFID	cFID	csFID
Translation $(2, 2)$ px	1.04	1.16	2.52	4.57
Translation $(16, 16)$ px	2.62	5.68	28.50	30.80
Rotation 90°	3.65	7.55	11.17	26.07
Reflection in x	2.84	7.65	9.21	25.24

Sample-size stability. Recomputing both metrics on nested ensembles of $N = 2000, 5000, 10000$ preserves the ordering of the mean scores on Euler and passive scalar. On Euler, the EDM/Direct-SI score ratios are $0.995, 0.988, 0.984$ for csFID and $0.706, 0.646, 0.615$ for cFID. Thus the smaller separation under spatial features persists throughout this range.

B.4 Additional qualitative conditional forecasts

Figure 5 shows an example test condition for SQG, nonlinear Klein–Gordon, and passive scalar. For SQG and nonlinear Klein–Gordon, Cascaded SI outperforms Direct SI and EDM by a large amount. For passive scalar, validation selects $S^\star = 1$, so Direct SI and Cascaded SI coincide.

C How cascading changes forecast error

A forecast over a fixed horizon T can be learned directly, in a single transition, or through a cascade of shorter transitions. The experiments in the main text show that shorter transitions can substantially improve the conditional forecast, but that the improvement can stop as the cascade becomes deeper. In this section, we study two mechanisms that can produce this behavior. First, a fixed approximation class may represent a short transition well while failing to represent the dependence that develops over a longer interval. Composing several simple maps can then represent dependence that a single map in the same class cannot. Second, under partial observation, repeatedly conditioning only on the current observed state can discard information carried by the unresolved state. We study these two effects separately before combining them in the rotation–OU model analyzed in the main text.

To keep the analysis explicit, we use forecast classes built from polynomial functions of the observed state. These provide a tractable setting in which we can calculate when the dependence of the forecast on its input leaves a fixed function class and how composing shorter fitted transitions changes the resulting error. We then study exact short conditional kernels to isolate the effect of repeated conditioning under partial observation. Finally, the rotation–OU model puts both effects in the same calculation and gives an example in which an intermediate cascade depth is optimal. The memory

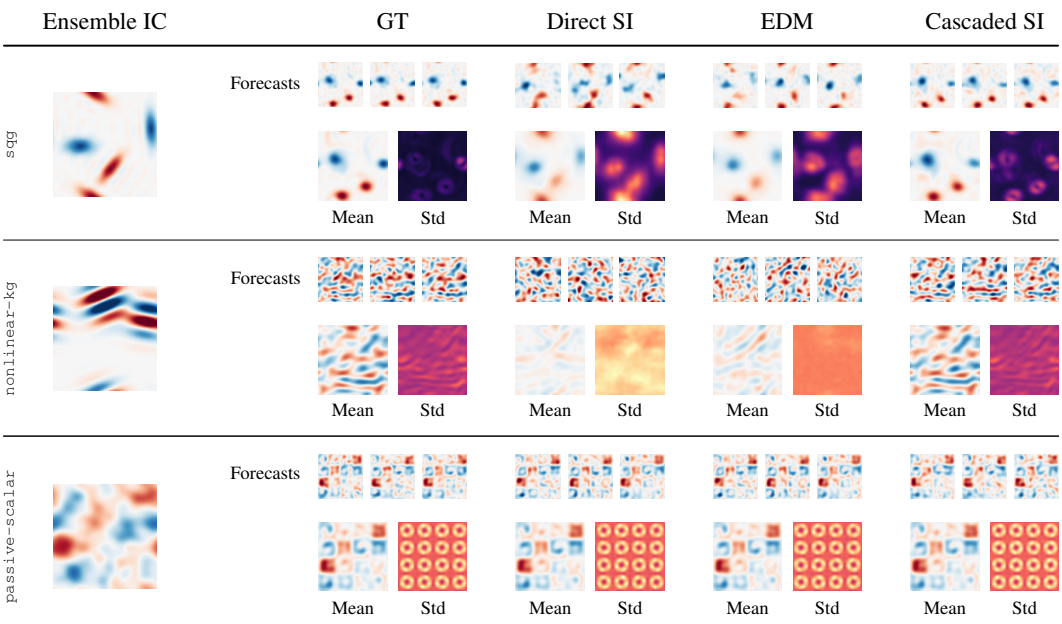

Figure 5: **Additional conditional endpoint distributions.** Test conditions for SQG, nonlinear Klein–Gordon, and passive scalar. For each method, three endpoint samples are shown together with the conditional mean and pointwise standard deviation estimated from 10,000 generated endpoints; reference statistics are computed from 10,000 simulator-generated endpoints from the same conditioning state.

effect that appears under partial observation is classical in optimal prediction Chorin et al. (2002); Ma et al. (2019).

C.1 When a fixed approximation class becomes insufficient

Our first goal is to understand how approximation error appears when the conditional forecast is restricted to a fixed class of functions. We begin with a general model of a partially observed dynamical system and identify the first term in the short-time expansion of the forecast that cannot be represented by the chosen class.

Let $Z_t \in \mathbb{R}^d$ denote the full physical state. We assume that it evolves according to the smooth autonomous system

$$\dot{Z}_t = V(Z_t),$$

where $V : \mathbb{R}^d \to \mathbb{R}^d$. We do not observe the full state Z_t. Instead, we observe

$$X_t = PZ_t \in \mathbb{R}^m,$$

where $P : \mathbb{R}^d \to \mathbb{R}^m$ is a fixed observation map.

Let $\chi : \mathbb{R}^m \to \mathbb{R}^p$ specify the quantity that we wish to forecast from the observed state. For example, χ may select a single observed Fourier coefficient, a collection of observed coefficients, or the entire observed state, in which case $\chi = \mathrm{Id}$. For each forecast horizon t, define

$$(\mathcal{T}_t\chi)(x) = \mathbb{E}[\chi(X_t) \mid X_0 = x]\,. \tag{29}$$

Thus, $(\mathcal{T}_t\chi)(x)$ is the conditional mean of the future quantity $\chi(X_t)$ given the observed initial state $X_0 = x$.

Because only part of the physical state is observed, the operators $\mathcal{T}_t$ need not satisfy the semigroup identity

$$\mathcal{T}_{t+s}\chi = \mathcal{T}_t(\mathcal{T}_s\chi).$$

Indeed, the unresolved variables may carry information from one time interval to the next that cannot be recovered from the current observed state alone. This issue will become important when we study cascades under partial observation below.

We first examine how the conditional forecast changes over a short physical time interval. For any smooth function $g : \mathbb{R}^d \to \mathbb{R}^p$ of the full state, the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t} g(Z_t) = Dg(Z_t)V(Z_t).$$

We therefore define the derivative of g along the full dynamics by

$$(\mathcal{L}_V g)(z) = Dg(z)V(z).$$

Higher-order derivatives along the dynamics are defined recursively by

$$\mathcal{L}_V^0 g = g, \qquad \mathcal{L}_V^{j+1} g = \mathcal{L}_V(\mathcal{L}_V^j g).$$

It follows that

$$\frac{\mathrm{d}^j}{\mathrm{d}t^j} g(Z_t)\Big|_{t=0} = \mathcal{L}_V^j g(Z_0).$$

We apply this identity to the full-state function $g = \chi \circ P$. After conditioning on the observed initial state, define

$$a_{j,\chi}(x) = \mathbb{E}\Big[\mathcal{L}_V^j(\chi \circ P)(Z_0) \mid X_0 = x\Big]. \tag{30}$$

Each $a_{j,\chi}$ is a function of the observed initial state x. However, the derivative inside the conditional expectation is computed using the full state Z_0 and the full dynamics V. The conditional expectation averages this derivative over the unresolved components of Z_0.

The theorem below gives precise conditions under which the functions $a_{j,\chi}$ are the coefficients in the short-time expansion of $\mathcal{T}_t\chi$. It also controls the error made by truncating that expansion.

Let

$$\mu = \mathrm{Law}(X_0)$$

denote the distribution of the observed initial state. Choose a closed linear space

$$\mathcal{F} \subset L^2(\mu; \mathbb{R}^p)$$

representing the fixed class of forecasting functions under consideration. The class $\mathcal{F}$ is chosen before selecting the forecast horizon and before fitting a predictor to future targets. Since $\mathcal{F}$ is closed, let $\Pi_{\mathcal{F}}$ denote the orthogonal projection from $L^2(\mu; \mathbb{R}^p)$ onto $\mathcal{F}$.

We define $n_\star$ to be the first order for which the corresponding coefficient cannot be represented by $\mathcal{F}$:

$$n_\star = \min\{j \geq 0 : a_{j,\chi} \notin \mathcal{F}\}.$$

If every coefficient $a_{j,\chi}$ belongs to $\mathcal{F}$, we set

$$n_\star = \infty.$$

In particular, if $n_\star = n < \infty$, then

$$a_{0,\chi}, \ldots, a_{n-1,\chi} \in \mathcal{F}, \qquad a_{n,\chi} \notin \mathcal{F}.$$

Thus, $a_{n,\chi}$ is the first coefficient in the short-time expansion that cannot be represented by the chosen approximation class.

Theorem 2. *Let $t_0 > 0$ and suppose that $n = n_\star < \infty$. Assume that, almost surely, the trajectory Z_t is defined on $[0, t_0]$ and the map*

$$t \longmapsto \chi(PZ_t)$$

is $(n+1)$ times continuously differentiable on this interval. Assume also that

$$\mathbb{E}\left[\left|\mathcal{L}_V^j(\chi \circ P)(Z_0)\right|^2\right] < \infty \qquad \textit{for every } j = 0, \ldots, n.$$

Define

$$M_{n+1} := \sup_{0 \leq u \leq t_0} \left(\mathbb{E}\left[\left|\mathcal{L}_V^{n+1}(\chi \circ P)(Z_u)\right|^2\right]\right)^{1/2},$$

and assume that $M_{n+1} < \infty$. *Finally, let*

$$c_n = \|(I - \Pi_{\mathcal{F}})a_{n,\chi}\|_{L^2(\mu)}.$$

Then $c_n > 0$ *and, for every* $t \in [0, t_0]$,

$$\left|\operatorname{dist}_{L^2(\mu)}(\mathcal{T}_t\chi, \mathcal{F}) - \frac{c_n t^n}{n!}\right| \le \frac{M_{n+1}t^{n+1}}{(n+1)!}. \tag{31}$$

Proof of Theorem 2. Taylor's integral formula along each trajectory gives, with $g = \chi \circ P$,

$$g(Z_t) = \sum_{j=0}^{n} \frac{t^j}{j!}\mathcal{L}_V^j g(Z_0) + \frac{1}{n!}\int_0^t (t-u)^n \mathcal{L}_V^{n+1} g(Z_u)\,\mathrm{d}u.$$

Condition on X_0 and use Minkowski's inequality and the L^2 contraction of conditional expectation. The conditional remainder has norm at most $M_{n+1}t^{n+1}/(n+1)!$. Applying $I - \Pi_{\mathcal{F}}$ removes all terms before n and does not increase the remainder. The reverse triangle inequality proves (31). Since $\mathcal{F}$ is closed and $a_{n,\chi} \notin \mathcal{F}$, its distance c_n is positive. □

C.2 A QUADRATIC DYNAMICAL MODEL AND A FIXED FORECAST CLASS

We now apply the general result above to a concrete class of dynamical systems. We consider quadratic dynamics, forecast the entire observed state, and use quadratic functions of the initial observation as our fixed forecast class. In the notation of the previous subsection, this means that

$$\chi = \mathrm{Id} \qquad \text{and} \qquad \mathcal{F} = \mathcal{P}_2.$$

Our goal is to show that, even when the full dynamics are quadratic, the conditional mean can develop a nonzero cubic dependence on the observed initial state.

This model isolates the nonlinear mechanism relevant to the active-scalar systems in our benchmark. Finite-dimensional Galerkin versions of Euler, for example, contain quadratic interactions between modes. These interactions can make the observed forecast depend on increasingly complicated combinations of the initial variables, and can transfer information between observed and unresolved modes. The result below is general for quadratic systems satisfying the stated assumptions. We also give Lorenz–96 as a standard additional example, although it is not part of the benchmark.

Let the full physical state satisfy

$$\dot{Z}_t = V(Z_t), \qquad V(z) = Q(z) + Az + f, \qquad Z_t \in \mathbb{R}^d, \tag{32}$$

where Q is homogeneous and quadratic, Az is linear, and f is constant. We assume that the quadratic term satisfies

$$z^\top Q(z) = 0 \qquad \text{for every } z \in \mathbb{R}^d. \tag{33}$$

This means that the quadratic term alone does not change $|Z_t|^2$. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}t}|Z_t|^2 = 2Z_t^\top Q(Z_t) = 0$$

when the dynamics contains only Q. This property holds in the examples that motivate the analysis: the quadratic term conserves enstrophy in a Galerkin truncation of two-dimensional Euler and the squared norm in Lorenz–96. The linear term Az and the forcing term f need not preserve this quantity. We use Eq. (33) both to control the growth of the trajectories and to prove that the cubic term identified below is nonzero.

As in the previous subsection, we observe

$$X_t = PZ_t \in \mathbb{R}^m.$$

We now assume that the rows of P are orthonormal:

$$PP^\top = I_m.$$

Define

$$E = P^\top.$$

Then

$$PE = I_m \qquad \text{and} \qquad |Ex| = |x| \quad \text{for every } x \in \mathbb{R}^m.$$

Thus, E maps an observed state x back into the full state space while preserving its observed coordinates and its norm.

We assume that the initial state can be written as

$$Z_0 = EX_0 + \eta, \qquad P\eta = 0, \qquad \mathbb{E}|Z_0|^8 < \infty. \tag{34}$$

Here, EX_0 is the observed part of the initial state and η is the unresolved part. The condition $P\eta = 0$ means that η is invisible in the observation $X_0 = PZ_0$. We assume that X_0 and η are independent. This matches the coarse–fine initialization used in the Euler benchmark, where the observed and hidden Fourier modes are sampled independently. Independence is assumed only at the initial time; the dynamics may couple the two parts at later times. The eighth-moment condition is used to control the remainder in the short-time expansion.

We keep the notation

$$\mu = \mathrm{Law}(X_0)$$

for the distribution of the observed initial state. We assume that μ has a density that is strictly positive on a nonempty open set. This ensures that a polynomial which is nonzero as a function cannot vanish μ-almost everywhere.

Let $\Phi_t(z)$ denote the full state at time t when the initial state is z. Given an observed initial state $X_0 = x$, define

$$K_t(x,\, \mathrm{d}x') = \mathrm{Law}(P\Phi_t(Z_0) \mid X_0 = x)\,. \tag{35}$$

Thus, $K_t(x, \cdot)$ is the conditional distribution of the future observed state X_t. Its conditional mean is

$$m_t(x) = \int x' K_t(x,\, \mathrm{d}x') = \mathbb{E}[X_t \mid X_0 = x]. \tag{36}$$

Under squared error, m_t is the best possible forecast of X_t based only on X_0. Since it predicts X_t from X_0 over the full interval, without using intermediate physical states, it is the direct conditional mean forecast in this analysis.

We now restrict this direct forecast to a fixed class of quadratic functions. Let $\mathcal{P}_2$ be the space of vector-valued polynomials $p : \mathbb{R}^m \to \mathbb{R}^m$ of total degree at most 2, and let Π_2 denote the orthogonal projection onto $\mathcal{P}_2$ in $L^2(\mu)$. Define

$$e_2(t) = \inf_{p \in \mathcal{P}_2} \|m_t - p\|_{L^2(\mu)} = \|m_t - \Pi_2 m_t\|_{L^2(\mu)}\,. \tag{37}$$

Thus, $e_2(t)$ measures how well a quadratic function of X_0 can represent m_t when the whole physical interval is predicted in a single transition.

This is the quantity that appears in Theorem 2 when

$$\chi = \mathrm{Id} \qquad \text{and} \qquad \mathcal{F} = \mathcal{P}_2.$$

Indeed,

$$(\mathcal{T}_t\, \mathrm{Id})(x) = \mathbb{E}[X_t \mid X_0 = x] = m_t(x),$$

and hence

$$e_2(t) = \mathrm{dist}_{L^2(\mu)}\left(\mathcal{T}_t\, \mathrm{Id}, \mathcal{P}_2\right).$$

The square $e_2(t)^2$ also has a direct interpretation in terms of prediction error. For any quadratic predictor $p \in \mathcal{P}_2$,

$$\mathbb{E}\big[|X_t - p(X_0)|^2\big] = \mathbb{E}\big[|X_t - m_t(X_0)|^2\big] + \|m_t - p\|^2_{L^2(\mu)}.$$

Taking the infimum over $p \in \mathcal{P}_2$ gives

$$\inf_{p \in \mathcal{P}_2} \mathbb{E}\big[|X_t - p(X_0)|^2\big] = \mathbb{E}\big[|X_t - m_t(X_0)|^2\big] + e_2(t)^2.$$

The first term is the prediction error that remains because X_0 does not fully determine X_t. The second term is the additional error caused by restricting the conditional mean forecast to quadratic functions.

The next subsection identifies the first part of the conditional mean that lies outside $\mathcal{P}_2$. We then apply Theorem 2 to obtain an explicit short-time estimate for $e_2(t)$.

C.3 THE CUBIC TERM THAT QUADRATIC FORECASTS CANNOT REPRESENT

We now identify the first coefficient in the short-time expansion of the direct conditional mean m_t that can lie outside $\mathcal{P}_2$. This will allow us to apply Theorem 2.

For $\chi = \mathrm{Id}$, the zeroth-order coefficient is

$$a_{0,\mathrm{Id}}(x) = x,$$

which belongs to $\mathcal{P}_2$. The first-order coefficient is

$$a_{1,\mathrm{Id}}(x) = \mathbb{E}[PV(Z_0) \mid X_0 = x]\,.$$

Since V is quadratic and $Z_0 = Ex + \eta$, this is a polynomial of degree at most two in x. Here we use the independence of X_0 and η, which ensures that the moments of η appearing in the conditional expectation do not depend on x. Therefore,

$$a_{1,\mathrm{Id}} \in \mathcal{P}_2.$$

The second-order coefficient is

$$a_{2,\mathrm{Id}}(x) = \mathbb{E}[P\,DV(Z_0)V(Z_0) \mid X_0 = x]\,.$$

Unlike the previous coefficients, it can contain terms of degree three in x. Its degree-three part comes from applying the quadratic term Q twice and is given by

$$C_3(x) = P\,DQ(Ex)Q(Ex). \tag{38}$$

Here, $DQ(Ex)$ is the derivative of Q evaluated at Ex, and the final application of P keeps only the observed coordinates. All the remaining terms in $a_{2,\mathrm{Id}}$ have degree at most two. Therefore,

$$a_{2,\mathrm{Id}} = C_3 + b_2 \qquad \text{for some } b_2 \in \mathcal{P}_2,$$

and hence

$$(I - \Pi_2)a_{2,\mathrm{Id}} = (I - \Pi_2)C_3.$$

We measure the part of C_3 that cannot be represented by a quadratic function by

$$c_3 = \|(I - \Pi_2)C_3\|_{L^2(\mu)}\,. \tag{39}$$

Since the zeroth- and first-order coefficients belong to $\mathcal{P}_2$, proving that $c_3 > 0$ will show that $n_\star = 2$. The subscript in c_3 refers to the fact that C_3 has polynomial degree three, even though it appears at second order in physical time.

Theorem 3. *Under the assumptions of the previous subsection, suppose that $Q(Ex)$ is not identically zero as a function of x. Then*

$$n_\star = 2 \qquad \textit{and} \qquad c_3 > 0.$$

Moreover, for every fixed $t_0 > 0$, there is a finite and explicitly bounded constant M_3 such that, for every $t \in [0, t_0]$,

$$\left| e_2(t) - \frac{c_3 t^2}{2} \right| \le \frac{M_3 t^3}{6}. \tag{40}$$

Consequently,

$$e_2(t)^2 = \frac{c_3^2}{4} t^4 + O(t^5). \tag{41}$$

The coefficient c_3 depends on the quadratic dynamics Q, the observation map P, and the distribution μ of the observed initial state. It does not depend on the distribution of the independent unresolved initial state η. The unresolved variables can still affect the lower-degree terms and the remainder.

The condition $Q(Ex) \not\equiv 0$ has a simple meaning: an initial state containing only observed modes must activate at least one quadratic interaction. To see why this implies $c_3 > 0$, differentiate

$$z^\top Q(z) = 0$$

in the direction $Q(z)$. This gives

$$|Q(z)|^2 + z^\top DQ(z)Q(z) = 0.$$

Setting $z = Ex$ and using $E = P^\top$ gives

$$x^\top C_3(x) = -|Q(Ex)|^2. \tag{42}$$

If $Q(Ex)$ is nonzero for some x, the right-hand side is strictly negative at that point. Therefore, C_3 is not identically zero.

Finally, C_3 is a homogeneous polynomial of degree three, whereas every element of $\mathcal{P}_2$ has degree at most two. Since μ has a density that is positive on a nonempty open set, a nonzero cubic polynomial cannot agree with a quadratic polynomial μ-almost everywhere. It follows that

$$c_3 > 0.$$

Notice that the condition $Q(Ex) \not\equiv 0$ does not require

$$PQ(Ex) \not\equiv 0.$$

The first quadratic interaction may therefore lie entirely in the unresolved coordinates. Identity (42) shows that, after the dynamics acts again, this interaction still produces a nonzero term in the observed forecast.

The conclusion $n_\star = 2$ applies to the forecast of the entire observed state. For a scalar quantity $\chi(x) = u^\top x$, the corresponding term is $u^\top C_3(x)$. Therefore, the same conclusion holds only if

$$u^\top C_3 \not\equiv 0.$$

Thus, some individual coordinates, including conserved or uncoupled quantities, may not contain this term.

For a general initial distribution, evaluating c_3 requires an $L^2(\mu)$ projection. When the observed initial state is Gaussian, this projection can be computed explicitly; see Corollary 4 below. The argument below also shows that the conclusion is unchanged by invertible affine changes of the observed coordinates.

Corollary 4 (An explicit coefficient for Gaussian conditioning states)**.** *If $X_0 \sim \mathcal{N}(a, \Sigma)$ with $\Sigma > 0$, write*

$$[C_3(\Sigma^{1/2} z)]_k = \sum_{i,j,\ell=1}^{m} T_{kij\ell} z_i z_j z_\ell, \tag{43}$$

with T symmetric in its last three indices. Then

$$c_3^2 = 6 \sum_{k,i,j,\ell} T_{kij\ell}^2, \qquad e_2(t)^2 = \frac{3}{2} \|T\|_F^2 \, t^4 + O(t^5). \tag{44}$$

For a fixed unit output direction u, the corresponding coefficient is obtained by replacing $T_{kij\ell}$ by $\sum_k u_k T_{kij\ell}$.

Proof of Corollary 4. Write $X_0 = a + \Sigma^{1/2} Z$, where $Z \sim \mathcal{N}(0, I_m)$. An invertible affine coordinate change preserves polynomial degree. The shift by a changes C_3 only by terms of degree at most two after its leading homogeneous part is extracted. For standard Gaussian inputs, the part of $Z_i Z_j Z_\ell$ orthogonal to degrees at most two is

$$:Z_i Z_j Z_\ell: \; = Z_i Z_j Z_\ell - \delta_{ij} Z_\ell - \delta_{i\ell} Z_j - \delta_{j\ell} Z_i.$$

Their covariance is the sum of the six matchings between the three indices in the first factor and those in the second factor. One direct derivation differentiates the identity

$$\mathbb{E}\big[e^{s^\top Z - |s|^2/2} e^{t^\top Z - |t|^2/2}\big] = e^{s^\top t}$$

three times in s and three times in t, at zero. Contracting against the symmetric tensor T gives

$$\mathbb{E}\left|\sum_{i,j,\ell} T_{kij\ell} :Z_i Z_j Z_\ell:\right|^2 = 6 \sum_{i,j,\ell} T_{kij\ell}^2.$$

Summation in k proves the first identity; Theorem 3 gives the stated leading term in $e_2(t)^2$. Contracting the output index gives the directional version. □

Affine changes of observed coordinates. Let B be invertible, $\widetilde{\mu} = (B \cdot + b)_{\#}\mu$, and $\widetilde{f}(u) = f(B^{-1}(u-b))$. Composition with $B^{-1}(\cdot - b)$ is an isometry from $L^2(\mu)$ onto $L^2(\widetilde{\mu})$ and a bijection of degree-r polynomial spaces. Taking the infimum over those spaces gives exact equality of the approximation errors, including their leading coefficients. If outputs are additionally transformed to $Cf + d$, with C invertible and $r \geq 0$, candidates correspond as $p \mapsto Cp + d$. The inequalities

$$\sigma_{\min}(C)|f - p| \leq |C(f - p)| \leq \sigma_{\max}(C)|f - p|$$

give the bounds after integration and minimization. An isometric output transformation gives equality. Thus membership of the vector coefficients, and their first escape index, is unchanged. This argument transports the input law and output metric consistently; arbitrary rescaling while holding a different law fixed is a different comparison.

Examples and comparisons. The results above apply differently to the systems considered in the paper and to related examples:

- **Galerkin Euler.** The quadratic vorticity interaction conserves enstrophy and therefore satisfies Eq. (33). Any observed subspace for which $Q(Ex) \not\equiv 0$ has the nonzero cubic term described above.
- **Surface quasi-geostrophic flow.** A finite Galerkin truncation of SQG also has quadratic transport. The conclusion of Theorem 3 applies when its assumptions hold, in particular when $Q \circ E \not\equiv 0$ for the chosen observed modes.
- **Lorenz–96.** The quadratic part conserves the squared Euclidean norm, while damping and forcing appear in the lower-order terms. For $d \geq 4$, the cubic term is nonzero when the full state is observed. Under partial observation, the result applies whenever $Q(Ex) \not\equiv 0$. Lorenz–96 is included here as a standard quadratic comparison rather than as one of our benchmark systems.
- **Conditional rotation.** After including the conditioning variable in the state, the vector field is
  $$(0, \omega c J y),$$
  which is quadratic and preserves the norm of y. Its finite-time conditional mean is
  $$R(\omega t c) y,$$
  which has nonpolynomial dependence on the conditioning variable c. This example shows explicitly how repeated quadratic interactions can produce a much more complicated finite-time forecast.
- **Prescribed passive scalar.** This benchmark provides a linear comparison. Its evolution is linear in the initial scalar field, so it does not create the same higher-degree dependence. Under independent coarse–fine initialization, its conditional mean remains affine at every forecast horizon.

The heterogeneous-wave benchmark is also linear in the full physical state, although partial observation of that state still makes the forecast probabilistic. The nonlinear Klein–Gordon benchmark contains a cubic nonlinearity and is therefore not covered by the quadratic theorem. The general first-escape result in Theorem 2 can still be applied whenever its regularity and moment assumptions hold.

The passive-scalar statement is made precise by the following proposition.

Proposition 5. *Suppose that the full state evolves linearly according to*

$$Z_t = U_t Z_0,$$

where $U_t : \mathbb{R}^d \to \mathbb{R}^d$ is the linear solution operator at time t. Assume that X_0 and η are independent. Then

$$m_t(x) = PU_t Ex + PU_t\, \mathbb{E}[\eta]$$

for every forecast horizon t. Hence both the affine and quadratic approximation errors are zero. If η is Gaussian, then the conditional distribution of the future observed state is Gaussian, with an affine mean and a covariance that does not depend on x.

The result follows by substituting $Z_0 = EX_0 + \eta$ into $Z_t = U_t Z_0$ and conditioning on $X_0 = x$; see the proof below.

Proof of the cubic estimate and the linear comparison. Under (33), the radial derivative of a solution of (32) is bounded above by $\|A\|_{\mathrm{op}}|Z| + |f|$. Applying the same argument to the time-reversed equation gives, for every finite real t,

$$|\Phi_t(z)| \le e^{a|t|}(|z| + b|t|), \qquad a = \|A\|_{\mathrm{op}}, \quad b = |f|. \tag{45}$$

One can justify the radial inequality at zero by differentiating $\sqrt{|Z|^2 + \varepsilon}$ and sending ε to zero. Local existence and continuation for a polynomial vector field then give global existence. For $A = f = 0$, the norm is exactly conserved.

Set $\mathcal{A} = DV\,V$ and $\mathcal{B} = D\mathcal{A}\,V$. They are polynomials of degree at most three and four. Along a trajectory they are the second and third time derivatives of the state. Therefore

$$P\Phi_t(z) = Pz + tPV(z) + \tfrac{1}{2}t^2 P\mathcal{A}(z) + \int_0^t \tfrac{1}{2}(t-s)^2 P\mathcal{B}(\Phi_s(z))\,\mathrm{d}s. \tag{46}$$

For a fixed $t_0 > 0$, the following is one explicit permissible bound:

$$\begin{aligned} K_3 &= \sup_{z\in\mathbb{R}^d} \frac{|P\mathcal{B}(z)|}{(1+|z|)^4} < \infty, \qquad\qquad G = e^{at_0}(1+bt_0), \\ M_3 &= K_3 G^4 \big[\mathbb{E}(1+|Z_0|)^8\big]^{1/2}. \end{aligned} \tag{47}$$

All quantities are determined by the polynomial coefficients, P, the initial eighth moment, and t_0. Indeed, $1 + |\Phi_s(z)| \le G(1+|z|)$ on this interval. Minkowski's inequality in (46) bounds the remainder in L^2 by $M_3 t^3/6$. Conditional expectation is an L^2 contraction, so

$$\begin{aligned} m_t(x) &= x + ta_1(x) + \tfrac{1}{2}t^2 a_2(x) + \mathcal{E}_t(x), \\ a_1(x) &= \mathbb{E}[PV(Z_0) \mid X_0 = x], \\ a_2(x) &= \mathbb{E}[P\mathcal{A}(Z_0) \mid X_0 = x], \end{aligned} \tag{48}$$

with $\|\mathcal{E}_t\|_{L^2(\mu)} \le M_3 t^3/6$.

Proof of Theorem 3. Independence in (34) implies $a_1 \in \mathcal{P}_2$. The homogeneous cubic part of $DV(z)V(z)$ is $DQ(z)Q(z)$; all remaining terms have degree at most two. Substituting $z = Ex + \eta$, every term with at least one hidden factor has at most two visible factors. Its expectation is therefore a polynomial of degree at most two. Thus

$$a_2(x) = C_3(x) + p_2(x), \qquad p_2 \in \mathcal{P}_2.$$

Differentiate $z^\top Q(z) = 0$ in direction $Q(z)$: $|Q(z)|^2 + z^\top DQ(z)Q(z) = 0$. At $z = Ex$ this is (42), because $E = P^\top$. Hence $Q \circ E \not\equiv 0$ implies $C_3 \not\equiv 0$. A nonzero homogeneous cubic cannot agree with a quadratic polynomial on an open set. Positivity of the density there and closedness of the finite-dimensional polynomial subspace imply $c_3 > 0$.

Applying $I - \Pi_2$ to (48) leaves

$$(I - \Pi_2)m_t = \tfrac{1}{2}t^2(I - \Pi_2)C_3 + (I - \Pi_2)\mathcal{E}_t.$$

The projection is a contraction; the reverse triangle inequality proves (40). Squaring proves (41). For example $e_2(t) \ge c_3 t^2/4$ whenever $M_3 t \le 3c_3/2$. □

Proof of Proposition 5. Substitute $Z_0 = EX_0 + \eta$ into $Z_t = U_t Z_0$ and condition on $X_0 = x$. Independence gives $\mathbb{E}[\eta \mid X_0 = x] = \mathbb{E}\eta$. If η is Gaussian, $PU_t\eta$ is Gaussian with covariance $PU_t\,\mathrm{Cov}(\eta)U_t^\top P^\top$, which is independent of x. □

Verification for the quadratic examples. For a finite Fourier–Galerkin truncation of two-dimensional Euler, the vorticity field obeys $\partial_t\omega = -u\cdot\nabla\omega$ and $\nabla\cdot u = 0$. Consequently,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|\omega\|_{L^2}^2 = -\frac{1}{2}\int u\cdot\nabla(\omega^2)\,\mathrm{d}x = 0.$$

In orthonormal real Fourier coordinates, the Galerkin vector field is homogeneous quadratic and satisfies $z^\top Q(z) = 0$. The condition $Q \circ E \not\equiv 0$ still needs to be checked for the chosen observed

subspace. For example, observed modes $(1,1)$ and $(2,0)$ interact to produce mode $(3,1)$ with a nonzero coefficient in a truncation that retains this mode. The output of this first interaction may itself be unresolved: Theorem 3 only requires it to be nonzero in the full Galerkin state.

For Lorenz–96 with cyclic indices $i = 1, \dots, d$, $d \geq 4$, write

$$V_i(z) = (z_{i+1} - z_{i-2})z_{i-1} - z_i + F, \qquad Q_i(z) = (z_{i+1} - z_{i-2})z_{i-1}.$$

Relabeling indices gives

$$\sum_i z_i Q_i(z) = \sum_i z_{i-1} z_i z_{i+1} - \sum_i z_{i-2} z_{i-1} z_i = 0.$$

Since $Q \not\equiv 0$, the certificate (42) proves that its full-state cubic term is nonzero. A particular coarse projection requires the additional check $Q \circ E \not\equiv 0$. The damping and forcing have lower degree and cannot cancel that term. For the conditional rotation example, $Q(c, y) = (0, \omega c J y)$ is likewise quadratic and satisfies $(c, y)^\top Q(c, y) = 0$.

C.4 Why a cascade can reduce approximation error

We now isolate the first effect in the cascading trade-off: shorter transitions can reduce approximation error. The previous subsection showed that a single quadratic function may fail to represent the conditional mean over the full forecast horizon. We now show that the same quadratic class can give a more accurate endpoint forecast when it is fitted over a shorter interval and then composed repeatedly.

Partial observation creates a separate difficulty because each stage of a cascade may discard information carried by unresolved variables. We set that issue aside in this subsection and assume that the full state is observed. Thus there is no unresolved component, and the exact forecast over an interval of length h is the deterministic flow map

$$z \longmapsto \Phi_h(z).$$

Let

$$B_R = \{z \in \mathbb{R}^d : |z| \leq R\},$$

and let ν be a probability distribution on B_R whose density, relative to the uniform distribution, is bounded above and below by positive constants. We use ν to measure approximation errors in this subsection.

Let Π_2^ν denote the orthogonal projection in $L^2(\nu)$ onto the space $\mathcal{P}_2(\mathbb{R}^d; \mathbb{R}^d)$ of quadratic functions from the full state space to itself. For a physical transition of length h, define

$$F_h^{(2)} = \Pi_2^\nu \Phi_h.$$

Thus, $F_h^{(2)}$ is the best quadratic approximation of the exact flow map Φ_h under the distribution ν. Its approximation error is

$$e_2^{\text{full}}(h) = \|(I - \Pi_2^\nu)\Phi_h\|_{L^2(\nu)}.$$

The superscript "full" distinguishes this error from $e_2(t)$, which was defined earlier for a forecast based only on the observed state.

Fix a final forecast horizon T. A direct quadratic forecast uses the single map

$$F_T^{(2)}.$$

A cascade of depth S divides the same physical horizon into intervals of length

$$\delta = \frac{T}{S}$$

and repeatedly applies the short-transition approximation $F_\delta^{(2)}$. Its endpoint map is therefore

$$\left(F_\delta^{(2)}\right)^{\circ S}.$$

The following theorem compares these two forecasts. It shows that a single quadratic map over the full horizon has a nonzero approximation error, while the endpoint error obtained by composing the fitted short-transition map converges to zero as the transition length decreases.

Theorem 6. *Suppose first that*

$$V = Q \not\equiv 0$$

and that Q satisfies Eq. (33)*. Then, for every $T > 0$,*

$$e_2^{\mathrm{full}}(T) > 0.$$

Therefore, the exact flow map Φ_T cannot be represented exactly by a quadratic function.

There are constants $D_2, \Lambda_2 < \infty$, independent of S, such that, for all sufficiently large S,

$$\sup_{z \in B_R} \left| \left(F_\delta^{(2)}\right)^{\circ S}(z) - \Phi_T(z) \right| \le D_2 \frac{e^{\Lambda_2 T} - 1}{\Lambda_2} \delta. \tag{49}$$

When $\Lambda_2 = 0$, the quotient is understood to be T. Consequently, for sufficiently large S, the cascade of short quadratic fits has smaller endpoint approximation error than every direct quadratic function.

For the more general dynamics

$$V(z) = Q(z) + Az + f,$$

the same convergence bound holds on a forward-invariant ball. If $Q \not\equiv 0$, the direct approximation error is positive for every sufficiently small nonzero T.

The bound in (49) tends to zero as the transition length δ tends to zero. Thus, composing short quadratic fits can approach the exact endpoint map even though no single quadratic function can represent that map exactly.

The same result holds for every fixed polynomial degree $r \ge 2$, with an endpoint-error bound of order δ^{r-1}. The proof of this more general statement, together with a finite-sample result for quadratic fitting, is given below in the proof of Theorem 6.

Proof of Theorem 6. Write $A_n = \mathcal{L}_V^n \mathrm{Id}$, with $A_0(z) = z$. Since $\deg V \le 2$, $\deg A_n \le n + 1$. The homogeneous component of degree $n + 1$ is $\mathcal{L}_Q^n \mathrm{Id}$.

Lemma 7. *If $Q \not\equiv 0$ is homogeneous quadratic and norm preserving, then $\mathcal{L}_Q^n \mathrm{Id} \not\equiv 0$ for every integer $n \ge 1$.*

Proof. Otherwise the nth time derivative of every Q-trajectory would vanish identically. Each trajectory would be a polynomial in time of degree at most $n - 1$. It exists for all real time and has constant norm. Every bounded real polynomial is constant, so all trajectories would be constant. Their derivative at zero would imply $Q \equiv 0$, a contradiction. □

On a forward-invariant ball, Taylor's formula gives

$$\Phi_t = \sum_{n=0}^{r} \frac{t^n}{n!} A_n + \mathcal{E}_t, \qquad \sup_{B_R} |\mathcal{E}_t| \le \frac{t^{r+1}}{(r+1)!} \sup_{B_R} |A_{r+1}|.$$

For $n < r$, $A_n \in \mathcal{P}_r$. Lemma 7 shows that A_r has a nonzero homogeneous component of degree $r + 1$. Consequently

$$\left| e_r^{\mathrm{full}}(t) - \frac{c_r t^r}{r!} \right| \le \frac{t^{r+1}}{(r+1)!} \sup_{B_R} |A_{r+1}|, \quad c_r = \|(I - \Pi_r^\nu) A_r\|_{L^2(\nu)} > 0. \tag{50}$$

This establishes the small-time claim for $V = Q + Az + f$, including Lorenz–96. It uses a finite Taylor remainder, not convergence of an infinite Lie series at the terminal horizon.

For the pure field $V = Q$, the direct gap is positive at every $T > 0$. Indeed, if $e_r^{\mathrm{full}}(T) = 0$, the finite-dimensional space is closed, so $\Phi_T = p$ for a polynomial p on B_R, first almost everywhere and then everywhere by continuity and the density assumption. Norm conservation implies the polynomial identity $|p(z)|^2 = |z|^2$ on all of $\mathbb{R}^d$. If p had degree greater than one, its leading squared norm could not vanish. Hence p is affine. But $\Phi_T(0) = 0$ and $D\Phi_T(0) = I$, so $p = \mathrm{Id}$. The second variational equation at the equilibrium zero is

$$D^2 \Phi_T(0)[a, b] = T D^2 Q(0)[a, b] \not\equiv 0,$$

which contradicts $p = \mathrm{Id}$.

Convergence of the fitted short-transition rule. Finite-dimensional norm equivalence gives a constant $C_{\Pi,r}$ with

$$\|p\|_{C^1(B_{2R})} \le C_{\Pi,r}\|p\|_{L^2(\nu)}, \qquad p \in \mathcal{P}_r. \tag{51}$$

Here the norm on the right is a genuine norm on polynomial coefficients. Let $M_r = \sup_{B_R} |A_r|$. The Taylor polynomial

$$T_{r-1,\delta} = \sum_{n=0}^{r-1} \frac{\delta^n}{n!} A_n$$

belongs to $\mathcal{P}_r$. The best $L^2(\nu)$ approximation satisfies

$$F_\delta^{(r)} = T_{r-1,\delta} + \Pi_r^\nu(\Phi_\delta - T_{r-1,\delta}), \qquad \|\Phi_\delta - T_{r-1,\delta}\|_{L^2(\nu)} \le M_r\delta^r/r!.$$

For $0 < \delta \le 1$, permissible constants are

$$D_r = (1 + C_{\Pi,r})M_r/r!,$$
$$\Lambda_r = \sum_{n=1}^{r-1} \frac{1}{n!} \sup_{B_{2R}} \|DA_n\|_{\mathrm{op}} + C_{\Pi,r}M_r/r!. \tag{52}$$

They yield

$$\sup_{B_R} |F_\delta^{(r)} - \Phi_\delta| \le D_r\delta^r, \qquad \mathrm{Lip}(F_\delta^{(r)}|_{B_{2R}}) \le 1 + \Lambda_r\delta.$$

Let $z_i = \Phi_{i\delta}(z_0)$ and $\widehat{z}_i = (F_\delta^{(r)})^{\circ i}(z_0)$, starting in B_R. As long as $\widehat{z}_i \in B_{2R}$,

$$|\widehat{z}_{i+1} - z_{i+1}| \le (1 + \Lambda_r\delta)|\widehat{z}_i - z_i| + D_r\delta^r.$$

Summing the recursion gives the bound in Theorem 6, with $r = 2$; the same argument gives order δ^{r-1} for any fixed $r \ge 2$. If its right-hand side is at most $R/2$, every iterate lies in $B_{3R/2}$, so induction justifies all the estimates. This condition holds for sufficiently large S because $r \ge 2$. Combining with the positive direct gaps above proves Theorem 6.

Imperfect fits and a finite-sample bound. For $r = 2$, let $\widehat{F}_\delta \in \mathcal{P}_2$ and $\eta_\delta = \|\widehat{F}_\delta - F_\delta^{(2)}\|_{L^2(\nu)}$. If $\mathrm{Lip}(\widehat{F}_\delta|_{B_{2R}}) \le 1 + \widehat{\Lambda}\delta$, the same proof gives

$$\sup_{B_R} |\widehat{F}_\delta^{\circ S} - \Phi_T| \le \frac{e^{\widehat{\Lambda}T} - 1}{\widehat{\Lambda}} \left(D_2\delta + \frac{C_{\Pi,2}\eta_\delta}{\delta} \right), \tag{53}$$

provided the right-hand side is at most $R/2$. For example, $\eta_\delta \le C\delta^2$ implies the required stability with a constant $\widehat{\Lambda}$ independent of $\delta \le 1$, by (51). An unqualified small one-step error is insufficient to establish this rollout bound.

For a concrete statistical version, let $p = \binom{d+2}{2}$ and let φ be an orthonormal scalar basis of $\mathcal{P}_2$ in $L^2(\nu)$. Set $\kappa_\varphi = \sup_{B_R} |\varphi|^2$. Observe n independent pairs

$$Z_i \sim \nu, \qquad U_i = \Phi_\delta(Z_i) + \zeta_i, \quad \mathbb{E}[\zeta_i \mid Z_i] = 0, \quad \mathbb{E}[|\zeta_i|^2 \mid Z_i] \le v_{\mathrm{obs}}^2.$$

For every $\varepsilon \in (0, 1)$, with probability at least $1 - 4\kappa_\varphi p/n - \varepsilon$, the empirical Gram matrix is invertible and least squares satisfies

$$\eta_\delta \le 2\sqrt{\frac{\kappa_\varphi(D_2^2\delta^4 + v_{\mathrm{obs}}^2)}{n\varepsilon}}. \tag{54}$$

To verify this, let $G_n = n^{-1}\sum_i \varphi_i\varphi_i^\top$. Independence gives $\mathbb{E}\|G_n - I\|_F^2 \le \kappa_\varphi p/n$, so Markov's inequality bounds the failure of $\|G_n - I\|_{\mathrm{op}} \le 1/2$ by $4\kappa_\varphi p/n$. The mean-zero matrix

$$Z_n = \frac{1}{n}\sum_i [\Phi_\delta(Z_i) - F_\delta^{(2)}(Z_i) + \zeta_i]\varphi_i^\top$$

has second moment at most $\kappa_\varphi(D_2^2\delta^4 + v_{\mathrm{obs}}^2)/n$. On the Gram event the coefficient error is $Z_nG_n^{-1}$, of norm at most $2\|Z_n\|_F$. Orthonormality identifies its Frobenius norm with the $L^2(\nu)$ norm of the prediction error. A second Markov inequality proves (54).

In the noiseless case the fitting error is proportional to $\delta^2/\sqrt{n}$ at fixed confidence. With fixed observation noise, the propagated contribution can instead grow like $v_{\mathrm{obs}}/(\delta\sqrt{n})$. This result concerns independent full-state regression pairs. Several transitions drawn from one trajectory do not constitute independent pairs, and physical hidden-state uncertainty is not identified with this observation-noise model.

C.5 Connection to the stochastic-interpolant learner

The previous subsection compared direct and cascaded forecasts at the level of their endpoint maps. The experiments in the main text implement these forecasts using stochastic interpolants. We now connect the approximation results above to a specific stochastic-interpolant learner. In the deterministic full-state setting of Theorem 6, this gives an explicit setting in which Cascaded SI has smaller endpoint approximation error than Direct SI.

We begin by describing the stochastic-interpolant training problem. Fix a physical transition length h. Given the initial observation $X_0 = x$, let

$$C = X_h, \qquad C \mid X_0 = x \sim K_h(x, \cdot),$$

denote the future state that we wish to generate. The physical transition length h is different from the interpolation variable $\tau \in [0, 1]$ used by the stochastic interpolant. We use the following Wiener stochastic interpolant:

$$Y_\tau = (1-\tau)x + \tau C + \sigma_b(1-\tau)\sqrt{\tau}\,\xi, \qquad R_\tau = C - x - \sigma_b\sqrt{\tau}\,\xi, \tag{55}$$

where $\xi \sim \mathcal{N}(0, I)$ is independent of (X_0, C) and $\sigma_b > 0$ controls the amount of noise. At $\tau = 0$, the interpolant starts from x, while it approaches the endpoint C as τ approaches 1.

The pair (Y_τ, R_τ) defines the regression problem used to train the stochastic interpolant. Let π denote the distribution from which the interpolation time τ is sampled during training, independently of (X_0, C, ξ). For a candidate drift b, define the training objective by

$$\mathcal{J}(b) = \frac{1}{2}\mathbb{E}\Big[|b(Y_\tau, \tau; X_0) - R_\tau|^2\Big], \tag{56}$$

where the expectation is taken over X_0, C, ξ, and $\tau \sim \pi$. We assume that the expectations appearing below are finite.

Under this objective, the optimal drift is

$$b^\star(y, \tau; x) = \mathbb{E}[R_\tau \mid Y_\tau = y, X_0 = x]\,.$$

Since

$$R_\tau = \frac{C - Y_\tau}{1-\tau},$$

the optimal drift can also be written as

$$b^\star(y, \tau; x) = \frac{\mathbb{E}[C \mid Y_\tau = y, X_0 = x] - y}{1-\tau}. \tag{57}$$

Moreover, for any candidate drift b,

$$\mathcal{J}(b) - \mathcal{J}(b^\star) = \frac{1}{2}\mathbb{E}\Big[|b(Y_\tau, \tau; X_0) - b^\star(Y_\tau, \tau; X_0)|^2\Big]. \tag{58}$$

Thus, the difference from the minimum training loss is exactly one half of the mean-square error between the candidate drift b and the optimal drift $b^\star$.

The identities above follow directly from the stated interpolant. For $\tau < 1$,

$$C - Y_\tau = (1-\tau)(C - x - \sigma_b\sqrt{\tau}\,\xi) = (1-\tau)R_\tau.$$

Conditional expectation gives (57). Since $b^\star$ is the conditional expectation of R_τ given (Y_τ, τ, X_0), the cross term in

$$|b - R_\tau|^2 = |b - b^\star|^2 + |b^\star - R_\tau|^2 + 2(b - b^\star)^\top(b^\star - R_\tau)$$

has zero expectation. This proves (58).

Before relating this objective to Direct SI and Cascaded SI, we connect the approximation of the conditional mean to the error in the generated endpoint distribution.

Proposition 8. *Let $\widehat{K}_h(x,\cdot)$ be any approximation of the true conditional distribution $K_h(x,\cdot)$. Define the mean of this approximate distribution by*

$$\widehat{m}_h(x) = \int c\,\widehat{K}_h(x,\,\mathrm{d}c).$$

Assume that

$$\int_{\mathbb{R}^m}\int_{\mathbb{R}^m} |c|^2 \left[K_h(x,\,\mathrm{d}c) + \widehat{K}_h(x,\,\mathrm{d}c)\right] \mu(\,\mathrm{d}x) < \infty.$$

Then

$$\int_{\mathbb{R}^m} W_2^2\Big(K_h(x,\cdot), \widehat{K}_h(x,\cdot)\Big)\,\mu(\,\mathrm{d}x) \geq \|m_h - \widehat{m}_h\|^2_{L^2(\mu)}\,. \tag{59}$$

Proof of Proposition 8. For each x, let $(C,\widehat{C})$ be any coupling of $K_h(x,\cdot)$ and $\widehat{K}_h(x,\cdot)$. Jensen's inequality gives $\mathbb{E}|C-\widehat{C}|^2 \geq |m_h(x)-\widehat{m}_h(x)|^2$. Take the infimum over couplings and integrate over x. □

Proposition 8 shows that an error in the conditional mean necessarily produces an error in the generated conditional distribution. In particular, if $\widehat{m}_h \in \mathcal{P}_2$, then

$$\int W_2^2\Big(K_h(x,\cdot), \widehat{K}_h(x,\cdot)\Big)\,\mu(\,\mathrm{d}x) \geq e_2(h)^2.$$

Therefore, the approximation error identified in Theorem 3 is also a lower bound on the endpoint distribution error of any learner whose conditional mean is restricted to be quadratic.

This first result concerns only the endpoint distributions. It does not yet connect the approximation error $e_2(h)$ to the stochastic-interpolant training objective $\mathcal{J}$. To obtain such a relation, we consider a particular family of conditional distributions for which both the optimal drift and the generated endpoint distribution can be calculated exactly.

Proposition 9. *Suppose that*

$$C \mid X_0 = x \sim \mathcal{N}(m_h(x), q_h I)\,, \qquad q_h \geq 0.$$

Assume that the interpolation-time law π is supported on $[0,\tau_*]$ for some $\tau_* < 1$. Consider the family of stochastic-interpolant drifts $b_{h,p}$ obtained by keeping the exact Gaussian dependence on y and τ and replacing the true conditional mean $m_h(x)$ with a candidate function $p(x)$. There is a constant $w_h > 0$ such that

$$\mathcal{J}(b_{h,p}) - \mathcal{J}(b_{h,m_h}) = \frac{1}{2} w_h \,\|p - m_h\|^2_{L^2(\mu)}\,. \tag{60}$$

When $b_{h,p}$ is integrated exactly, the generated endpoint distribution is

$$\mathcal{N}(p(x), q_h I)\,.$$

Consequently,

$$W_2^2(\mathcal{N}(m_h(x), q_h I), \mathcal{N}(p(x), q_h I)) = |m_h(x) - p(x)|^2. \tag{61}$$

Thus, for this family, approximating the conditional mean controls both the stochastic-interpolant training loss and the endpoint distribution error. In particular, choosing $p = \Pi_2 m_h$ gives

$$\mathcal{J}(b_{h,\Pi_2 m_h}) - \mathcal{J}(b_{h,m_h}) = \frac{1}{2} w_h e_2(h)^2,$$

and

$$\int W_2^2(\mathcal{N}(m_h(x), q_h I), \mathcal{N}(\Pi_2 m_h(x), q_h I))\,\mu(\,\mathrm{d}x) = e_2(h)^2.$$

We can now connect this result to the direct and cascaded forecasts from the previous subsection. There, the full state is observed and the dynamics are deterministic. In that setting, the exact endpoint is

$$C = \Phi_h(x),$$

and hence

$$m_h = \Phi_h.$$

There is no conditional randomness in the endpoint. This is the case $q_h = 0$ in Proposition 9; the zero covariance simply means that the endpoint distribution is concentrated at $\Phi_h(x)$.

Since $w_h > 0$, the best quadratic function p minimizes

$$\|p - \Phi_h\|^2_{L^2(\nu)}.$$

The optimal quadratic function is therefore precisely

$$p = \Pi_2^\nu \Phi_h = F_h^{(2)}.$$

Moreover, the stochastic interpolant associated with p produces the deterministic endpoint $p(x)$. The fitted quadratic maps studied in Theorem 6 are therefore exactly the endpoint maps produced by this stochastic-interpolant learner.

We can now state the resulting comparison between Direct SI and Cascaded SI.

Corollary 10. *Under the assumptions of Theorem 6, consider the stochastic-interpolant learner in Proposition 9 with $q_h = 0$ and with $p \in \mathcal{P}_2(\mathbb{R}^d; \mathbb{R}^d)$. The endpoint map produced by Direct SI is*

$$F_T^{(2)}.$$

If the horizon is divided into S intervals of length

$$\delta = \frac{T}{S},$$

the endpoint map produced by Cascaded SI is

$$\left(F_\delta^{(2)}\right)^{\circ S}.$$

For all sufficiently large S, Cascaded SI has smaller endpoint approximation error than every Direct SI whose endpoint map is quadratic.

Corollary 10 shows, in this specified full-state setting, that the approximation benefit proved in Theorem 6 carries directly to the corresponding stochastic-interpolant learners. This result isolates approximation error. It does not yet include the information loss that can occur when the state passed between stages of a cascade is only partially observed.

Proof of Proposition 9. Suppose $C \mid x \sim \mathcal{N}(m(x), qI)$, $q \geq 0$. Define

$$D_\tau = \tau q + \sigma_b^2(1-\tau)^2, \quad K_\tau = q/D_\tau, \quad H_\tau = \sigma_b^2(1-\tau)/D_\tau,$$

and, for a candidate $p(x)$, set

$$b_p(y, \tau; x) = \frac{K_\tau - 1}{1-\tau} y - K_\tau x + H_\tau p(x). \tag{62}$$

At $q = 0$, this is $(p(x) - y)/(1-\tau)$. For $q > 0$, Gaussian conditioning proves $b_m = b^\star$, since

$$\mathbb{E}[C \mid y, x] = m(x) + K_\tau[y - (1-\tau)x - \tau m(x)].$$

The same formula holds directly for $q = 0$. Thus $b_p - b_m = H_\tau(p - m)$.

Take an interpolation-time law π independent of (x, C, ξ) with $0 < w_q := \int H_\tau^2 \pi(\,\mathrm{d}\tau) < \infty$ and finite regression losses. One common law supported in $[0, \tau_*]$, $\tau_* < 1$, suffices for all $q \geq 0$. The analytic drift formula is specified on all of $[0, 1)$; only p is fitted. Equation (60) follows with $w_h = w_{q_h}$.

The affine SDE

$$\mathrm{d}\overline{Y}_\tau = b_p(\overline{Y}_\tau, \tau; x)\,\mathrm{d}\tau + \sigma_b(1-\tau)\,\mathrm{d}W_\tau, \qquad \overline{Y}_0 = x, \tag{63}$$

has, on every compact subinterval of $[0, 1)$, a unique Gaussian solution. Its mean and covariance are

$$(1-\tau)x + \tau p(x), \qquad [\tau^2 q + \sigma_b^2 \tau(1-\tau)^2] I.$$

Substitution into the linear mean and covariance equations verifies both expressions and their initial conditions. Taking $\tau \uparrow 1$ gives the endpoint law $\mathcal{N}(p(x), qI)$. Equal-covariance Gaussians are translates: the translation coupling has cost $|p - m|^2$, and Jensen's inequality gives the matching lower bound. This proves Proposition 9. When $q > 0$, the corresponding conditional KL is $|p(x) - m(x)|^2/(2q)$; when $q = 0$, differing point masses have infinite KL and Wasserstein distance remains the appropriate finite metric.

C.6 What repeated conditioning in a cascade forgets

Corollary 10 shows that a cascade of shorter transitions can reduce approximation error when the full state is observed. We now return to the partially observed setting. Here cascading introduces a second effect: at the end of each stage, the next transition is conditioned only on the current observed state. Information carried by the unresolved variables, and by their relation to earlier observations, may therefore be lost.

We first study this effect using exact short-transition distributions. There is therefore no learning error in the analysis below. Any difference between the direct conditional law and the cascade is caused only by repeated conditioning on the observed state.

Exact short-transition distributions. We extend the notation introduced above by defining, for $0 \le s < t \le T$,

$$K_{s,t}(x,\cdot) = \mathrm{Law}(X_t \mid X_s = x), \qquad \mu_t = \mathrm{Law}(X_t).$$

Thus,

$$K_{0,h} = K_h \qquad \text{and} \qquad \mu_0 = \mu.$$

The kernel $K_{s,t}(x,\cdot)$ is the exact conditional distribution of the observed state at time t, given only the observed state $X_s = x$.

For two conditional distributions K and L, we use the notation

$$\mathfrak{D}_\mu(K,L) = \left[\int W_2^2(K(x,\cdot), L(x,\cdot))\, \mu(\mathrm{d}x)\right]^{1/2}.$$

Thus, $\mathfrak{D}_\mu(K,L)$ is their average conditional Wasserstein distance when the input state has distribution μ.

Consider a partition

$$0 = t_0 < t_1 < \cdots < t_S = T,$$

and write

$$K_i = K_{t_i,t_{i+1}}, \qquad i = 0,\dots,S-1.$$

The cascade built from the exact short-transition distributions is

$$K^{(S)} = K_0 K_1 \cdots K_{S-1}. \tag{64}$$

Starting from $X_0 = x$, this composition first samples from $K_0(x,\cdot)$ and then repeatedly samples the next observed state conditional only on the current one. We measure the difference between this exact-kernel cascade and the direct conditional distribution $K_{0,T}$ by

$$C_S = \mathfrak{D}_{\mu_0}\left(K_{0,T}, K^{(S)}\right). \tag{65}$$

Because all the kernels in $K^{(S)}$ are exact, C_S contains no learning or approximation error. It measures only the error caused by replacing the true conditional dependence from X_0 to X_T with repeated conditioning on the current observed state.

The short-transition kernels always satisfy

$$\mu_s K_{s,t} = \mu_t.$$

Consequently, their composition produces the correct unconditional distribution μ_{t_i} at every time in the partition. However, it need not produce the correct distribution conditional on $X_0 = x$. In other words, all one-time marginals can be correct even when the dependence between the initial and final observations is wrong (Brunick & Shreve, 2013).

The following theorem makes this loss of information precise. As the cascade becomes finer, the composition of the exact observed-state conditional transitions approaches a deterministic evolution driven by the conditional mean observed velocity. Consequently, the memory error C_S approaches the error made by replacing the true observed trajectory with this deterministic evolution.

The result does not require the quadratic dynamics considered in the previous subsections. It applies to any smooth partially observed deterministic system satisfying the conditions below.

Theorem 11. *Suppose that the full state evolves according to*

$$\dot{Z}_t = V(Z_t)$$

on a compact invariant set $D \subset \mathbb{R}^d$, and that V is twice continuously differentiable on a neighborhood of D. As before, let

$$X_t = PZ_t.$$

Define the conditional mean of the observed velocity by

$$\overline{v}(t,x) = \mathbb{E}[PV(Z_t) \mid X_t = x]\,.$$

Assume that $\overline{v}$ has a bounded extension to $[0,T] \times \mathbb{R}^m$ that is Lipschitz continuous in x with constant L and Lipschitz continuous in t with constant H. Let $\psi_{s,t}(x)$ denote the solution at time t of

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi_{s,t}(x) = \overline{v}(t, \psi_{s,t}(x))\,, \qquad \psi_{s,s}(x) = x.$$

Define

$$M = \sup_{z \in D} |PV(z)|, \qquad A_0 = \frac{1}{2} \sup_{z \in D} |P\,DV(z)V(z)|,$$

and

$$B_0 = A_0 + \frac{1}{2}(H + LM).$$

For the equally spaced partition with

$$\delta = \frac{T}{S},$$

let

$$r_S = e^{LT}\left(2M\sqrt{T\delta} + B_0 T\delta\right). \tag{66}$$

Then

$$\mathfrak{D}_{\mu_0}\left(K^{(S)}, \delta_{\psi_{0,T}(\cdot)}\right) \le r_S, \tag{67}$$

where $\delta_{\psi_{0,T}(\cdot)}(x,\cdot)$ is the point mass at $\psi_{0,T}(x)$. Moreover,

$$|C_S - \Gamma_T| \le r_S, \qquad \Gamma_T^2 = \mathbb{E}\big[|X_T - \psi_{0,T}(X_0)|^2\big]\,. \tag{68}$$

The deterministic flow generated by $\overline{v}$ has the correct one-time marginals:

$$(\psi_{0,t})_{\#}\mu_0 = \mu_t \qquad \textit{for every } t \in [0,T].$$

Finally, define

$$v_0 = \mathbb{E}[\operatorname{tr} \operatorname{Cov}\left(PV(Z_0) \mid X_0\right)]\,.$$

Then

$$C_S \ge \left(T\sqrt{v_0} - A_0T^2 - r_S\right)_+, \tag{69}$$

where $(a)_+ = \max\{a, 0\}$.

The main conclusion follows from Eq. (68). Since $\delta = T/S$, we have $r_S \to 0$ as $S \to \infty$, and therefore

$$C_S \longrightarrow \Gamma_T.$$

Thus, as the cascade becomes increasingly fine, the error caused by discarding unresolved information approaches the error of replacing the true observed trajectory by the deterministic conditional-mean flow

$$x \longmapsto \psi_{0,T}(x).$$

Equation (67) explains why. The composition of increasingly short observed-state conditional transitions converges to the point mass

$$\delta_{\psi_{0,T}(x)}.$$

Conditional on the initial observation $X_0 = x$, the infinitely fine exact-kernel cascade therefore becomes deterministic. It follows the average observed velocity $\overline{v}(t,x)$ and loses the variability caused by the unresolved state.

The unresolved variables persist along the true physical trajectory. Their effects over different physical intervals are therefore correlated. By contrast, after every stage of the cascade, the exact observed-state kernel retains only the new observed state. It forgets which unresolved state produced that transition and averages again over all unresolved states compatible with the current observation.

For a smooth deterministic system, the unresolved contribution to one short increment has size of order δ. After S such independently reconditioned transitions, the accumulated variance is of order

$$S\delta^2 = T\delta,$$

which converges to zero as $\delta \to 0$. This is why the random effect of the unresolved variables disappears and only the conditional mean velocity remains.

Importantly, the limiting deterministic flow still has the correct one-time distributions:

$$(\psi_{0,t})_{\#}\mu_0 = \mu_t.$$

It can therefore reproduce the distribution of X_t at every individual time while failing to preserve the correct dependence between X_0 and X_T. The error Γ_T measures this lost dependence.

The quantity v_0 measures how much the observed velocity can vary among full states with the same initial observation. If $v_0 > 0$, the unresolved state affects the initial observed velocity. The lower bound in Eq. (69) then shows that, for sufficiently small fixed $T > 0$ and sufficiently fine cascades, the memory error remains strictly positive. Thus, arbitrarily short transitions do not in general recover the correct conditional endpoint distribution.

Proof of Theorem 11. The extension of $\bar{v}$ can be assumed to have norm at most M: project any given extension onto the closed Euclidean ball of radius M. This changes no conditional velocity on its support and does not increase its spatial or temporal Lipschitz constants.

Along a physical trajectory, $|X_{t+\delta} - X_t| \le M\delta$. Taylor's formula followed by conditioning gives, for μ_{t_i}-almost every x,

$$\begin{aligned} &\int (x' - x) K_i(x, \mathrm{d}x') = \delta\bar{v}(t_i, x) + r_i(x), && |r_i(x)| \le A_0\delta^2, \\ &\int |x' - x|^2 K_i(x, \mathrm{d}x') \le M^2\delta^2. \end{aligned} \tag{70}$$

Construct a Markov chain $\widetilde{X}_i$ with these kernels, starting from μ_0. Its marginal is μ_{t_i} at every step, so all almost-everywhere bounds hold at the states it visits. It has the recursion

$$\widetilde{X}_{i+1} = \widetilde{X}_i + \delta\bar{v}(t_i, \widetilde{X}_i) + r_i(\widetilde{X}_i) + \xi_{i+1},$$

where ξ_{i+1} is a martingale difference with conditional second moment at most $M^2\delta^2$.

For $x_i = \psi_{0,t_i}(\widetilde{X}_0)$, the deterministic flow has the one-step remainder bound

$$|x_{i+1} - x_i - \delta\bar{v}(t_i, x_i)| \le \tfrac{1}{2}(H + LM)\delta^2.$$

Sum the two recursions and apply discrete Gronwall to the running maximum. The result is

$$\max_{i \le S} |\widetilde{X}_i - x_i| \le e^{LT} \left[\max_{i \le S} \left| \sum_{j=1}^{i} \xi_j \right| + B_0 T\delta \right].$$

Doob's L^2 inequality and orthogonality of the increments give

$$\left\| \max_{i \le S} \left| \sum_{j=1}^{i} \xi_j \right| \right\|_{L^2} \le 2M\sqrt{S}\,\delta = 2M\sqrt{T\delta}.$$

Distance to a point mass is the second moment of displacement from that point. The coupling uses the same initial state for the chain and the deterministic flow, so it proves (67). The chain endpoint has law μ_T for every S; passing to the limit proves $(\psi_{0,T})_{\#}\mu_0 = \mu_T$. Apply the same argument on $[0, t]$ for each earlier time.

The reverse triangle inequality gives (68). Also

$$\Gamma_T^2 = \mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(X_T \mid X_0) + \mathbb{E}|m_T(X_0) - \psi_{0,T}(X_0)|^2. \tag{71}$$

Finally $X_T = X_0 + TPV(Z_0) + \mathcal{E}_T$, with $\|\mathcal{E}_T\|_{L^2} \le A_0 T^2$. Conditional centering is an orthogonal projection in L^2, so

$$\sqrt{\mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(X_T \mid X_0)} \ge \big(T\sqrt{v_0} - A_0 T^2\big)_+.$$

Combine with (71) and (68) to obtain (69). □

Retaining the initial observation. The theorem also identifies a way to prevent this particular loss of information. Instead of conditioning each short transition only on the current observed state, one can retain the initial observation throughout the cascade.

For a fixed initial observation a, define

$$K_i^a(x,\cdot) = \mathrm{Law}\left(X_{t_{i+1}} \mid X_{t_i} = x,\, X_0 = a\right)$$

and

$$\mu_i^a = \mathrm{Law}(X_{t_i} \mid X_0 = a).$$

The kernel K_i^a therefore receives both the current observed state x and the initial observation a.

Proposition 12. *For any Euclidean-valued process and any finite partition, the exact kernels defined above satisfy*

$$\delta_a K_0^a K_1^a \cdots K_{S-1}^a = \mathrm{Law}(X_T \mid X_0 = a) \tag{72}$$

for μ_0-almost every a.

Now let $\widehat{K}_i^a$ be approximations of K_i^a. Suppose that

$$W_2\Big(\widehat{K}_i^a(x,\cdot), \widehat{K}_i^a(x',\cdot)\Big) \le L_i|x - x'|$$

for every a, and define

$$\varepsilon_i^2 = \mathbb{E}\Big[W_2^2\Big(K_i^{X_0}(X_{t_i},\cdot), \widehat{K}_i^{X_0}(X_{t_i},\cdot)\Big)\Big].$$

If the required second moments and the quantity on the right-hand side below are finite, then

$$\Big[\mathbb{E}_{a\sim\mu_0} W_2^2\Big(\mathrm{Law}(X_T \mid X_0 = a), \delta_a \widehat{K}_0^a \widehat{K}_1^a \cdots \widehat{K}_{S-1}^a\Big)\Big]^{1/2} \le \sum_{i=0}^{S-1} \varepsilon_i \prod_{j>i} L_j. \tag{73}$$

Proof of Proposition 12. For μ_0-almost every a, the tower property gives

$$\mu_i^a K_i^a = \mu_{i+1}^a.$$

Starting from $\mu_0^a = \delta_a$ proves (72) by induction.

For the learned bound, compare two compositions at a time, keeping the first i kernels exact and all subsequent kernels learned. The input to the replacement at step i has law μ_i^a. Its conditional W_2 distance changes by at most

$$e_i(a) = \left[\int W_2^2\big(K_i^a(x,\cdot), \widehat{K}_i^a(x,\cdot)\big)\mu_i^a(\,\mathrm{d}x)\right]^{1/2}.$$

Each later learned kernel enlarges this distance by at most its Wasserstein Lipschitz constant. Sum over the replacement steps, then apply Minkowski's inequality over a. Since $\mathbb{E}_a[e_i(a)^2] = \varepsilon_i^2$, this gives (73). □

The first part of Proposition 12 shows that retaining X_0 removes this endpoint memory error when the short-transition kernels are exact. This does not mean that the resulting process reproduces the full joint law of the physical trajectory. It guarantees only the correct conditional law at the final time.

This anchored construction is a separate variant of the cascade. The Cascaded SI evaluated in the main benchmark conditions each stage on the generated current state, as described in Section 3.1; the result above shows what changes theoretically if the initial observation is retained as an additional conditioning variable.

Adding learning error. We now consider the additional error caused by learning the short transitions. There are three conditional endpoint distributions to distinguish:

$$K_{0,T}, \qquad K^{(S)} = K_0 K_1 \cdots K_{S-1}, \qquad \widehat{K}_0 \widehat{K}_1 \cdots \widehat{K}_{S-1}.$$

The first is the true transition from time 0 to time T. The second is the cascade of exact short transitions, each conditioned only on the current observed state. The third is the learned cascade, in which each exact transition K_i is replaced by an approximation $\widehat{K}_i$.

The distance between the first two distributions is the memory error C_S studied above. We now bound the distance between the second and third distributions.

For each segment, define

$$\epsilon_i^2 = \int W_2^2\Big(K_i(x,\cdot), \widehat{K}_i(x,\cdot)\Big)\, \mu_{t_i}(\,\mathrm{d}x).$$

Thus, ϵ_i measures how well the learned transition $\widehat{K}_i$ approximates the exact transition K_i, averaged over the states that occur at time t_i.

To control how this error affects later stages, assume that

$$W_2\Big(\widehat{K}_i(x,\cdot), \widehat{K}_i(x',\cdot)\Big) \le L_i |x - x'|.$$

The constant L_i measures how much the ith learned transition can enlarge a difference between two input states. An error introduced at step i passes through all later learned transitions and can therefore be enlarged by the factor

$$\prod_{j>i} L_j.$$

This leads to the accumulated learning-error bound

$$B_S = \sum_{i=0}^{S-1} \epsilon_i \prod_{j>i} L_j,$$

where the product is equal to one when $i = S-1$. In particular,

$$\mathfrak{D}_{\mu_0}\left(K^{(S)}, \widehat{K}_0 \widehat{K}_1 \cdots \widehat{K}_{S-1}\right) \le B_S.$$

Thus, B_S bounds how much replacing the exact short transitions with their learned approximations changes the final conditional distribution.

The total endpoint distribution error of the learned cascade is

$$E_S = \mathfrak{D}_{\mu_0}\left(K_{0,T}, \widehat{K}_0 \widehat{K}_1 \cdots \widehat{K}_{S-1}\right).$$

Recall that

$$C_S = \mathfrak{D}_{\mu_0}\left(K_{0,T}, K^{(S)}\right)$$

is the error caused by repeatedly conditioning only on the current observed state, even when all short transitions are exact. The triangle inequality and the reverse triangle inequality give

$$|E_S - C_S| \le B_S.$$

Equivalently,

$$(C_S - B_S)_+ \le E_S \le C_S + B_S. \tag{74}$$

These inequalities do not say that the total error is the sum of a memory error and a learning error. The learned approximations may move the endpoint distribution either closer to or farther from the true distribution. The result says only that learning can change the exact-kernel cascade error C_S by at most B_S.

The bounds also do not by themselves show that Cascaded SI is better than Direct SI. The quantity $e_2(T)$ is a lower bound on the endpoint distribution error of any direct forecast whose conditional mean is quadratic, whereas

$$E_S \le C_S + B_S$$

is an upper bound on the error of a learned cascade. Therefore, the condition

$$C_S + B_S < e_2(T)$$

would be sufficient to prove that the cascade is better than every direct forecast with a quadratic conditional mean. Whether this condition holds depends on the system and on the cascade depth S. In particular, these bounds alone do not show that the endpoint error is minimized at a finite depth $S > 1$.

The results above therefore isolate the two mechanisms in the cascading trade-off. Shorter transitions can reduce approximation error, but repeated conditioning on a partial observation can discard information carried by the unresolved state. The bounds keep these effects separate; they do not determine where their balance is optimal. We next turn to the rotation–OU model, where the endpoint error of a specified learner can be computed explicitly and a finite optimal cascade depth can be proved.

Proof of the learned-kernel bound. A kernel satisfying the W_2 Lipschitz bound with constant L_i sends input laws at distance d to output laws at distance at most $L_i d$: couple the inputs, couple the conditional outputs, and take infima. Its compositions have the product Lipschitz bound.

Compare the exact and learned products through hybrid products with an exact prefix and a learned suffix. At replacement index i, the common prefix, averaged over X_0, has law μ_{t_i}. Couple that prefix identically, use an optimal coupling of $K_i(x,\cdot)$ and $\widehat{K}_i(x,\cdot)$, and propagate through the learned suffix. The integrated conditional cost of this replacement is at most $\epsilon_i \prod_{j>i} L_j$. Triangle inequalities over the hybrids give

$$\mathfrak{D}_{\mu_0}(K^{(S)}, \widehat{K}_0 \cdots \widehat{K}_{S-1}) \le B_S.$$

The triangle and reverse triangle inequalities with $K_{0,T}$ prove (74). In particular, when Theorem 11 applies, $|E_S - \Gamma_T| \le B_S + r_S$. Neither B_S nor C_S is presumed monotone. A mean-only fitting error cannot replace ϵ_i, which measures an entire conditional distribution.

C.7 Direct and cascaded forecasts in the rotation–OU model: an explicit trade-off

The previous results identify two opposing effects of cascading. Shorter transitions can be easier to approximate, but repeated conditioning on the observed state can discard information carried by unresolved variables. Those general results do not yet show whether these effects can produce an optimal intermediate cascade depth.

We now exhibit this behavior in a model for which the generated endpoint distribution can be calculated explicitly. This is the model introduced in Section 4.1. We proceed in three steps. First, we construct a family of forecasting problems with the same target endpoint distribution but different amounts of hidden memory. Second, we compute the endpoint distribution generated by Direct SI and Cascaded SI for a specified quadratic learner. Finally, we compare direct forecasting, fine cascades, and an intermediate cascade.

A fixed endpoint task with tunable hidden memory. Let

$$c \sim \mathrm{Unif}[-1,1], \qquad Y_0 \sim \mathcal{N}(0, s_0^2 I_2),$$

independently, and let the hidden state $H_t \in \mathbb{R}^2$ and observed state $Y_t \in \mathbb{R}^2$ satisfy

$$\begin{aligned} \mathrm{d}H_t &= -\kappa H_t \,\mathrm{d}t + \sqrt{2\kappa}\,\mathrm{d}B_t, & H_0 &\sim \mathcal{N}(0, I_2), \\ \mathrm{d}Y_t &= \omega c J Y_t \,\mathrm{d}t + \beta_\rho R(\omega c t) H_t \,\mathrm{d}t + \sigma_\rho \,\mathrm{d}W_t, \end{aligned} \tag{75}$$

where

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad R(\theta) = e^{\theta J},$$

and all initial variables and Brownian motions are independent. The conditioning variable c is retained throughout the forecast. By contrast, H_t is unresolved: every stage is conditioned on the current value of Y_t and on c, but not on H_t.

For fixed $T, q, \kappa, s_0 > 0$, define

$$V_H(t) = 2\left[\frac{t}{\kappa} - \frac{1 - e^{-\kappa t}}{\kappa^2}\right], \qquad \beta_\rho^2 = \frac{\rho q}{V_H(T)}, \qquad \sigma_\rho^2 = \frac{(1-\rho)q}{T}, \tag{76}$$

where $\rho \in [0, 1]$. Thus, the integrated OU process contributes $\rho q I_2$ to the terminal covariance, while the white noise contributes $(1 - \rho)qI_2$.

To remove the deterministic rotation, introduce

$$Z_t = R(-\omega c t) Y_t.$$

The dynamics of Z_t contain no rotational drift, and

$$Z_T \mid (Y_0 = y, c) \sim \mathcal{N}(y, qI_2).$$

Transforming back to the original coordinates gives

$$Y_T \mid (Y_0 = y, c) \sim \mathcal{N}(R(\omega c T)y, qI_2) \tag{77}$$

for every $\rho \in [0, 1]$. Consequently, changing ρ changes the amount of information carried through time by the hidden state without changing the target conditional endpoint distribution.

Exact short transitions. Although the endpoint forecasting task is independent of ρ, the short conditional transitions are not. For $0 \le s \le t$, set

$$\begin{aligned} I_\kappa(s,t) &= \frac{2s}{\kappa} - \frac{(1 - e^{-\kappa s})(1 + e^{-\kappa(t-s)})}{\kappa^2}, \\ v_s &= s_0^2 + \beta_\rho^2 V_H(s) + \sigma_\rho^2 s, \\ b_{s,t} &= s_0^2 + \beta_\rho^2 I_\kappa(s,t) + \sigma_\rho^2 s, \qquad r_{s,t} = \frac{b_{s,t}}{v_s}. \end{aligned} \tag{78}$$

Consider the equal partition $t_i = i\delta$, where $\delta = T/S$, and write

$$r_i = r_{t_i, t_{i+1}}, \qquad q_i = v_{t_{i+1}} - r_i^2 v_{t_i}.$$

Gaussian conditioning gives

$$Y_{t_{i+1}} \mid (Y_{t_i} = y, c) \sim \mathcal{N}(r_i R(\omega c \delta) y, q_i I_2)\,. \tag{79}$$

When $\rho > 0$, composing these exact short transitions does not recover Eq. (77). At every reconditioning step, the cascade retains Y_{t_i} and c but forgets the unresolved state H_{t_i}. A cascade can therefore incur endpoint error even before its short transitions are approximated.

A fixed quadratic learner for Direct SI and Cascaded SI. We next introduce approximation error while keeping the comparison fully specified. We take $0 \le \rho < 1$, so that every q_i is positive, and use the Gaussian stochastic-interpolant family of Proposition 9. The exact scalar coefficients r_i and q_i are supplied to the learner. The only fitted quantity is a shared function $p(c, y)$ of total degree at most two, used to approximate the rotation $R(\omega c\delta)y$ in the conditional mean.

Set

$$\nu = \omega\delta, \qquad j_0(\nu) = \frac{\sin\nu}{\nu}, \qquad j_1(\nu) = \frac{\sin\nu - \nu\cos\nu}{\nu^2},$$

with $j_0(0) = 1$ and $j_1(0) = 0$. Minimizing the expected stochastic-interpolant training objective over the quadratic function class gives, at every stage,

$$p_\nu(c, y) = M_\nu(c)y, \qquad M_\nu(c) = j_0(\nu)I_2 + 3j_1(\nu)cJ. \tag{80}$$

Centered Gaussian symmetry eliminates terms that are nonlinear in y. The fitted map in Eq. (80) is therefore the L^2 projection of the rotation onto functions that are linear in y and affine in c.

The role of the cascade is now explicit. Direct SI uses $\nu = \omega T$ and must approximate the full rotation in one transition. Cascaded SI uses $\nu = \omega T/S$ and composes S simpler fitted rotations.

Exact endpoint distribution error. Define

$$d_\nu(c) = j_0(\nu)^2 + 9j_1(\nu)^2c^2, \qquad A_S = \prod_{i=0}^{S-1} r_i,$$

and let

$$\widehat{q}_0(c) = 0, \qquad \widehat{q}_{i+1}(c) = r_i^2 d_\nu(c)\widehat{q}_i(c) + q_i. \tag{81}$$

After S learned transitions, the generated conditional distribution is Gaussian, with mean $A_S M_\nu(c)^S y$ and covariance $\widehat{q}_S(c)I_2$.

We measure its discrepancy from the target through the average squared conditional Wasserstein error

$$\mathcal{E}_S(\rho) = \frac{1}{2}\int_{-1}^{1} \left[s_0^2 \left\| R(\omega T c) - A_S M_\nu(c)^S \right\|_F^2 + 2\left(\sqrt{q} - \sqrt{\widehat{q}_S(c)}\right)^2 \right] \mathrm{d}c. \tag{82}$$

The first term measures the error in the conditional mean, while the second measures the error in the conditional covariance. Equation (82) gives the complete endpoint error of the specified learner. It is not obtained by adding a separately defined approximation error and memory error.

Direct forecasting and the infinitely deep cascade. Before comparing all values of S, it is useful to identify the two extremes. For Direct SI, $S = 1$. Writing $\Omega = \omega T$, we obtain

$$\mathcal{E}_1(\rho) = 2s_0^2\left[1 - j_0(\Omega)^2 - 3j_1(\Omega)^2\right] > 0 \tag{83}$$

whenever $\omega \neq 0$. The generated covariance is exact, but one quadratic map cannot reproduce the full dependence of $R(\omega T c)y$ on c. Since the endpoint task is independent of ρ, this direct approximation error is also independent of ρ.

At the opposite extreme, let $S \to \infty$ and define

$$a_\rho(t) = \frac{\beta_\rho^2(1 - e^{-\kappa t})}{\kappa v_t}, \qquad A_\infty = \exp\left(\int_0^T a_\rho(t)\,\mathrm{d}t\right),$$
$$q_\infty = v_T - s_0^2 A_\infty^2, \qquad \mathcal{C}_{\rho,\infty} = 2s_0^2(A_\infty - 1)^2 + 2\left(\sqrt{q} - \sqrt{q_\infty}\right)^2. \tag{84}$$

If $\rho = 0$, there is no persistent hidden input and arbitrarily short transitions recover the endpoint distribution. If $0 < \rho < 1$, however, $\mathcal{C}_{\rho,\infty} > 0$: an increasingly deep cascade repeatedly forgets information carried by the hidden OU state.

The comparison therefore contains two competing effects. Direct SI has a fixed approximation error from fitting the full rotation in one transition. Increasing the cascade depth reduces this error by replacing that transition with simpler rotations. But in the presence of persistent hidden memory, a sufficiently deep cascade approaches a nonzero error caused by repeated reconditioning on the observed state. The following theorem shows that, when the hidden-memory contribution is sufficiently weak but nonzero, neither extreme is optimal.

Theorem 13 (An intermediate cascade can be optimal)**.** *Fix $T, \kappa, q, s_0 > 0$ and $\omega \neq 0$. For the centered rotation–OU model and the quadratic stochastic-interpolant learner defined above, Eq.* (82) *has the following properties.*

1. No persistent hidden memory. *If $\rho = 0$, then*

$$\mathcal{E}_S(0) = O(S^{-2}) \qquad \text{as } S \to \infty.$$

Thus, in the absence of the persistent hidden OU contribution, progressively finer cascades recover the target endpoint distribution.

2. Persistent hidden memory. *For every fixed $0 < \rho < 1$, there is a constant $\Lambda_\rho > 0$ such that*

$$\mathcal{E}_S(\rho) = \mathcal{C}_{\rho,\infty} - \Lambda_\rho \frac{T}{S} + O(S^{-2}). \tag{85}$$

Thus, once the cascade becomes sufficiently deep, increasing S moves the endpoint error upward toward the nonzero limit $\mathcal{C}_{\rho,\infty}$.

3. Intermediate optimum. *There exists $\rho_* > 0$ such that, for every $0 < \rho < \rho_*$, the minimum over positive integers S is attained at a finite $S^\star \geq 2$, and*

$$\mathcal{E}_{S^\star}(\rho) < \min\left\{\mathcal{E}_1(\rho), \mathcal{C}_{\rho,\infty}\right\}. \tag{86}$$

Hence the best finite cascade is strictly better than both Direct SI and the infinitely deep cascade limit.

The theorem makes the cascading trade-off explicit. Starting from $S = 1$, increasing the depth reduces the error from approximating the rotation. Once the transitions become too short, however, the benefit of fitting simpler rotations is outweighed by the information lost through repeated conditioning without H_t. The minimum can therefore occur at an intermediate cascade depth.

How the optimal cascade depth depends on memory. The previous theorem establishes the existence of an intermediate optimum. We now determine how its location changes as the hidden-memory contribution becomes small. Define

$$u(t) = \frac{q(1-e^{-\kappa t})}{\kappa V_H(T)(s_0^2 + qt/T)}, \qquad U = \int_0^T u(t)\,\mathrm{d}t, \qquad D = \frac{1}{2}\left[u(T) + \kappa U\right],$$
$$\lambda_2 = 4s_0^2\left(1 + \frac{s_0^2}{q}\right)UD, \qquad B_{\mathrm{fit}} = \frac{\omega^4 T^2(4s_0^2 + q)}{90}. \tag{87}$$

Both λ_2 and B_{fit} are strictly positive.

The origin of the scaling can be seen by writing $\delta = T/S$. For small ρ and small δ, the endpoint error satisfies

$$\mathcal{E}_S(\rho) = \mathcal{C}_{\rho,\infty} - \lambda_2\rho^2\delta + B_{\mathrm{fit}}\delta^2 + O\big(\rho^3\delta + \rho\delta^2 + \delta^3\big). \tag{88}$$

The positive term $B_{\mathrm{fit}}\delta^2$ is the remaining error from approximating the short rotation. The negative term $-\lambda_2\rho^2\delta$ describes the improvement obtained by keeping a finite transition length instead of taking the infinitely deep reconditioning limit. Balancing these two leading terms gives $\delta^\star \asymp \rho^2$, or equivalently $S^\star \asymp \rho^{-2}$. The following theorem makes this statement precise.

Theorem 14 (The optimal cascade depth grows as memory weakens)**.** *Under the assumptions of Theorem 13, every global minimizer over positive integers satisfies*

$$\rho^2 S^\star(\rho) \longrightarrow \frac{2B_{\mathrm{fit}}T}{\lambda_2} \qquad \text{as } \rho \downarrow 0. \tag{89}$$

Moreover,

$$\min_{S\geq 1}\mathcal{E}_S(\rho) = \mathcal{C}_{\rho,\infty} - \frac{\lambda_2^2}{4B_{\mathrm{fit}}}\rho^4 + o(\rho^4). \tag{90}$$

Thus, weaker hidden memory permits a deeper useful cascade:

$$S^\star(\rho) \asymp \rho^{-2}.$$

The rotation–OU model therefore exhibits the same qualitative trade-off that motivates the empirical depth study in the main paper: the direct transition can have substantial approximation error, a moderate cascade can reduce it, and a sufficiently deep cascade can lose part of that improvement through repeated conditioning on an incomplete state. The theorem is a statement about this specified Gaussian process and quadratic SI learner. It does not claim that hidden OU-like memory is the mechanism responsible for the depth curves of the PDE benchmarks, nor does it predict the value of $S^\star$ selected in those experiments.

The conditioning calculations, the uniform expansion in Eq. (88), and the proofs of Theorems 13–14 are given below.

C.8 Proofs for the rotation–OU model

Exact OU conditioning. Conditionally on c, rotate the process in (75) by $Z_t = R(-\omega ct)Y_t$. The visible rotation cancels and

$$Z_t = Y_0 + \beta_\rho\int_0^t H_u\,\mathrm{d}u + \sigma_\rho\widetilde{W}_t. \tag{91}$$

The rotated Brownian motion has the standard law and is independent of H, Y_0 conditionally on c. This entire conditional law is independent of c and ω.

Since H is stationary with covariance $\mathbb{E}[H_s H_t^\top] = e^{-\kappa|t-s|} I_2$, for $0 \le s \le t$ define

$$\begin{aligned} I_\kappa(s,t) &= \int_0^s \int_0^t e^{-\kappa|r-u|}\,\mathrm{d}u\,\mathrm{d}r = \frac{2s}{\kappa} - \frac{(1-e^{-\kappa s})(1+e^{-\kappa(t-s)})}{\kappa^2}, \\ v_s &= s_0^2 + \beta_\rho^2 V_H(s) + \sigma_\rho^2 s, \\ b_{s,t} &= s_0^2 + \beta_\rho^2 I_\kappa(s,t) + \sigma_\rho^2 s, \qquad r_{s,t} = b_{s,t}/v_s. \end{aligned} \tag{92}$$

Then $\mathrm{Cov}(Z_s) = v_s I_2$ and $\mathrm{Cov}(Z_s, Z_t) = b_{s,t} I_2$. Gaussian conditioning gives

$$Z_t \mid Z_s = z \sim \mathcal{N}\big(r_{s,t} z, [v_t - r_{s,t}^2 v_s] I_2\big). \tag{93}$$

At $s = 0, t = T$, $r_{0,T} = 1$ and $v_T - s_0^2 = \beta_\rho^2 V_H(T) + \sigma_\rho^2 T = q$. This proves the endpoint law (77).

For a partition write

$$r_i = r_{t_i,t_{i+1}}, \qquad q_i = v_{t_{i+1}} - r_i^2 v_{t_i}, \qquad A_S = \prod_{i=0}^{S-1} r_i, \qquad \widetilde{q}_S = v_T - s_0^2 A_S^2.$$

Composition of (93) gives

$$\widetilde{Z}_T \mid Y_0 = y \sim \mathcal{N}(A_S y, \widetilde{q}_S I_2). \tag{94}$$

To verify the covariance, start $w_0 = 0$ and use $w_{i+1} = r_i^2 w_i + q_i$. Induction yields $w_i = v_{t_i} - s_0^2 \prod_{j<i} r_j^2$, so $\widetilde{q}_S \ge 0$.

The average squared conditional Wasserstein cost is

$$\mathcal{C}_{\rho,S} = 2s_0^2 (A_S - 1)^2 + 2(\sqrt{q} - \sqrt{\widetilde{q}_S})^2. \tag{95}$$

This is a squared distance, unlike the root distance C_S in (65). To derive it, the conditional mean discrepancy in the co-rotating frame is $(A_S - 1)y$, whose squared norm has expectation $2s_0^2(A_S-1)^2$. Isotropic Gaussian covariances contribute $2(\sqrt{q} - \sqrt{\widetilde{q}_S})^2$. Terminal rotation is an isometry, so the result is independent of ω.

For strict monotonicity in ρ, compute

$$D_{s,t} := I_\kappa(s,t) - V_H(s) = \frac{(1-e^{-\kappa s})(1-e^{-\kappa(t-s)})}{\kappa^2}.$$

Then

$$r_{s,t} = 1 + \frac{\rho q D_{s,t}/V_H(T)}{s_0^2 + qs/T + \rho q[V_H(s)/V_H(T) - s/T]}. \tag{96}$$

Its derivative in ρ is $(qD_{s,t}/V_H(T))(s_0^2 + qs/T)$ divided by the squared denominator, strictly positive for $0 < s < t$. The first partition interval has $r_0 = 1$, and each later one has $r_i > 1$ at $\rho > 0$. Hence A_S increases strictly from one and $\widetilde{q}_S$ decreases from q for $S \ge 2$. Both terms of (95) increase strictly as a function of ρ. At $S = 1$ or $\rho = 0$ the cost is zero. Continuity covers the endpoint $\rho = 1$. This proves monotonicity in the memory parameter for a fixed partition; it does not assert monotonicity under every change of cascade depth.

Optimal quadratic fit and endpoint law. Keep the centered initial law, take $0 \le \rho < 1$, and use equal intervals of length $\delta = T/S$. Set $\Omega = \omega T$ and $\nu = \omega\delta$. The marginal $Y_{t_i} \sim \mathcal{N}(0, v_{t_i} I_2)$ is independent of c, and

$$Y_{t_{i+1}} \mid (Y_{t_i} = y, c) \sim \mathcal{N}(r_i R(\nu c) y, q_i I_2).$$

Here $q_i > 0$ because the physical white-noise increment has positive variance for $\rho < 1$. The parameter c is retained as a conditioning input; it is not interpolated.

Supply the known r_i, q_i and fit a shared polynomial $p(c, y)$ of total degree at most two, with candidate local mean $r_i p(c, y)$. The Gaussian SI reduction multiplies each stage's mean-squared loss by a positive scalar depending on q_i. Centered Gaussian symmetry in y shows that only terms linear in

y can correlate with the target. Their coefficients can have degree at most one in c. Therefore the unique function that minimizes the expected squared error fitted at every stage is

$$p_\nu(c, y) = M_\nu(c)y, \qquad M_\nu(c) = j_0(\nu)I_2 + 3j_1(\nu)cJ. \tag{97}$$

Indeed, $\mathbb{E}_c \cos(\nu c) = j_0(\nu)$, while orthogonal projection of $\sin(\nu c)$ onto the span of c has coefficient $3\mathbb{E}_c[c\sin(\nu c)] = 3j_1(\nu)$. Every positive weighted average of the stage losses has this same minimizer: after restricting to the linear-in-y part, the only additional stage factors are scalar $r_i^2 v_{t_i}$. Thus one shared fit suffices.

Let $d_\nu(c) = j_0(\nu)^2 + 9j_1(\nu)^2c^2$, so $M_\nu M_\nu^\top = d_\nu I_2$. The exact learned Gaussian kernels compose to a Gaussian endpoint conditional on (y, c), with mean $A_S M_\nu(c)^S y$ and scalar covariance $\widehat{q}_S(c)I_2$, where

$$\widehat{q}_0(c) = 0, \qquad \widehat{q}_{i+1}(c) = r_i^2 d_\nu(c)\widehat{q}_i(c) + q_i. \tag{98}$$

Define

$$d_\rho = \frac{1}{2}\left[a_\rho(T) + \int_0^T (\kappa a_\rho(t) + a_\rho(t)^2)\,\mathrm{d}t\right],$$

$$\mathcal{C}(A) = 2s_0^2(A-1)^2 + 2\left(\sqrt{q} - \sqrt{v_T - s_0^2A^2}\right)^2.$$

Then $\mathcal{C}(A_\infty) = \mathcal{C}_{\rho,\infty}$.

Proof of Theorem 13. At $\rho = 0$, $r_i = 1$ and $q_i = q/S$. Uniformly in $c \in [-1, 1]$,

$$M_{\omega\delta}(c) - R(\omega\delta c) = O(\delta^2), \qquad \|M_{\omega\delta}(c)\|_{\mathrm{op}} \le 1 + C\delta^2.$$

A telescoping product gives $M_{\omega\delta}^S - R(\Omega c) = O(\delta)$. Also $d_{\omega\delta} = 1 + O(\delta^2)$, so

$$\widehat{q}_S(c) = \frac{q}{S}\sum_{j=0}^{S-1} d_{\omega\delta}(c)^j = q + O(\delta).$$

Since $q > 0$, both squared errors in (82) are $O(\delta^2)$. Together with (83), this proves part (i).

Next fix $0 < \rho < 1$. Equation (96) is equivalently

$$r_i = 1 + a_\rho(t_i)\frac{1 - e^{-\kappa\delta}}{\kappa}.$$

All following Taylor estimates are uniform in $t_i \in [0, T]$ and $c \in [-1, 1]$, with constants for the fixed physical parameters. Expanding the logarithm gives

$$\log r_i = a_\rho(t_i)\delta - \tfrac{1}{2}[\kappa a_\rho(t_i) + a_\rho(t_i)^2]\delta^2 + O(\delta^3).$$

The left Riemann sum for the smooth function a_ρ, with $a_\rho(0) = 0$, yields

$$\log A_S = \log A_\infty - d_\rho\delta + O(\delta^2), \qquad A_S = A_\infty(1 - d_\rho\delta) + O(\delta^2). \tag{99}$$

Here $d_\rho > 0$ and $A_\infty > 1$.

To justify differentiating the covariance cost, note that $v_t' = 2a_\rho(t)v_t + \sigma_\rho^2$. An integrating factor gives

$$q_\infty = \sigma_\rho^2 A_\infty^2 \int_0^T \exp\left(-2\int_0^t a_\rho(s)\,\mathrm{d}s\right)\mathrm{d}t > 0. \tag{100}$$

Therefore $\mathcal{C}$ is smooth near A_∞, and

$$\mathcal{C}'(A) = 4s_0^2(A-1) + 4s_0^2 A\left(\frac{\sqrt{q}}{\sqrt{v_T - s_0^2A^2}} - 1\right) > 0 \tag{101}$$

for $1 < A \le A_\infty$. It follows that the exact-kernel squared cost is

$$\mathcal{C}_{\rho,S} = \mathcal{C}_{\rho,\infty} - \Lambda_\rho\delta + O(\delta^2), \qquad \Lambda_\rho = A_\infty d_\rho \mathcal{C}'(A_\infty) > 0. \tag{102}$$

It remains to show that fitting the polynomial rotation does not change this first-order coefficient. Identifying $uI + vJ$ with $u + iv$, Taylor expansion near one gives

$$\log\big(j_0(\omega\delta) + 3ij_1(\omega\delta)c\big) = i\omega\delta c + \tfrac{1}{2}\omega^2\delta^2(c^2 - 1/3) + O(\delta^3).$$

Multiplying by $S = T/\delta$ and exponentiating gives

$$\begin{aligned} M_{\omega\delta}(c)^S &= R(\Omega c)[I + \delta\alpha(c)I + O(\delta^2)], & \alpha(c) &= \tfrac{1}{2}\omega^2 T(c^2 - 1/3), \\ d_{\omega\delta}(c) &= 1 + \omega^2\delta^2(c^2 - 1/3) + O(\delta^4), & \mathbb{E}_c\alpha(c) &= 0. \end{aligned} \tag{103}$$

The contribution to the mean risk linear in δ is a scalar multiple of $\mathbb{E}_c\alpha(c)$, so it vanishes.

For the covariance, define nonnegative weights $w_i = q_i \prod_{j=i+1}^{S-1} r_j^2$. They satisfy $\sum_i w_i = \widetilde{q}_S \le q$, and (81) gives

$$\widehat{q}_S(c) = \sum_i w_i d_{\omega\delta}(c)^{S-1-i} = \widetilde{q}_S + \omega^2\delta^2(c^2 - 1/3)\sum_i (S - 1 - i)w_i + O(\delta^2).$$

The displayed correction is $O(\delta)$ and has mean zero in c. The remainder is uniform, since $S\delta = T$ and $\sum_i w_i \le q$. As $\widetilde{q}_S \to q_\infty > 0$, the function $w \mapsto 2(\sqrt{q} - \sqrt{w})^2$ has bounded second derivative in a neighborhood of all these covariances for large S. Taylor expansion and the zero-mean correction therefore give an expected change $O(\delta^2)$. The same conclusion for the mean term was just proved. Consequently

$$\mathcal{E}_S(\rho) = \mathcal{C}_{\rho,S} + O(\delta^2). \tag{104}$$

Combining (102) and (104) proves the second claim. This step is a calculation from the fitted endpoint law, not an assumption that learning and memory errors add.

Finally $a_\rho = O(\rho)$ uniformly on $[0, T]$ as $\rho \downarrow 0$, since $v_t \ge s_0^2$. Thus $A_\infty = 1 + O(\rho)$ and $\mathcal{C}_{\rho,\infty} = O(\rho^2) \to 0$. The strictly positive direct risk (83) is independent of ρ. Choose $\rho_* > 0$ so that $\mathcal{C}_{\rho,\infty} < \mathcal{E}_1$ for $0 < \rho < \rho_*$. Part (ii) supplies a sufficiently large finite S_0 with

$$\mathcal{E}_{S_0}(\rho) < \mathcal{C}_{\rho,\infty} < \mathcal{E}_1(\rho).$$

Because $\mathcal{E}_S(\rho) \to \mathcal{C}_{\rho,\infty}$, every sufficiently large index has risk strictly larger than $\mathcal{E}_{S_0}(\rho)$. The minimum is therefore attained in a finite set of depths and cannot occur at one. This proves the final claim. □

Joint expansion and optimal cascade depth. The fixed-ρ remainder above is insufficient to optimize jointly over S and ρ. We first justify a uniform second-order expansion directly from the finite recurrences that define the learned endpoint law.

Lemma 15 (Expansion of a smooth affine recurrence)**.** *Let θ range over a compact parameter set. Suppose the scalar real or complex functions $F(t, h, \theta)$ and $G(t, h, \theta)$ are smooth near $[0, T] \times \{0\}$, uniformly in θ, with $F(t, 0, \theta) = 1$ and $G(t, 0, \theta) = 0$. For $h = T/S$, let*

$$y_{i+1} = F(ih, h, \theta)y_i + G(ih, h, \theta), \qquad y_0 = y_{\mathrm{init}}(\theta),$$

where the initial value is smooth. Then smooth functions $y^{[j]}(t, \theta)$, $j = 0, 1, 2$, exist such that

$$y_S = y^{[0]}(T, \theta) + hy^{[1]}(T, \theta) + h^2 y^{[2]}(T, \theta) + O(h^3),$$

with a remainder uniform in θ.

Proof. Expand $F = 1 + hf_1 + h^2 f_2 + h^3 f_3 + O(h^4)$ and $G = hg_1 + h^2 g_2 + h^3 g_3 + O(h^4)$, uniformly on the compact set. Writing $y_j = y^{[j]}$, choose the functions by the linear equations

$$\begin{aligned} y_0' &= f_1 y_0 + g_1, \\ y_1' &= f_1 y_1 + f_2 y_0 + g_2 - \tfrac{1}{2}y_0'', \\ y_2' &= f_1 y_2 + f_2 y_1 + f_3 y_0 + g_3 - \tfrac{1}{2}y_1'' - \tfrac{1}{6}y_0''', \end{aligned}$$

with $y_0(0) = y_{\mathrm{init}}$ and $y_1(0) = y_2(0) = 0$. Successive solution gives smooth functions with bounded derivatives on the compact set. Taylor expansion shows that $y_0(t) + hy_1(t) + h^2 y_2(t)$ satisfies one step of the recurrence with defect $O(h^4)$. Since $|F| \le 1 + Ch$, the terminal defect is bounded by

$$Ch^4 \sum_{i=0}^{S-1} (1 + Ch)^i = O(h^3),$$

uniformly in θ. □

Lemma 16 (Joint small-memory and fine-cascade expansion)**.** *For some fixed $0 < \rho_0 < 1$ and all sufficiently large S, uniformly for $0 \le \rho \le \rho_0$ and $h = T/S$,*

$$\mathcal{E}_S(\rho) = \mathcal{C}_{\rho,\infty} - \lambda_2 \rho^2 h + B_{\mathrm{fit}} h^2 + O(\rho^3 h + \rho h^2 + h^3), \tag{105}$$

with the strictly positive constants in (87)*.*

Proof. Write $a = a_\rho$ and

$$r(t,h) = 1 + a(t)(1 - e^{-\kappa h})/\kappa.$$

The complex mean multiplier obeys the recurrence in Lemma 15 with

$$F = r(t,h)[j_0(\omega h) + 3i j_1(\omega h) c], \quad G = 0, \quad y_{\mathrm{init}} = 1.$$

The covariance obeys it with

$$F = r(t,h)^2 d_{\omega h}(c), \quad G = v_{t+h} - r(t,h)^2 v_t, \quad y_{\mathrm{init}} = 0.$$

These coefficients are smooth jointly in (t, h, ρ, c), with removable singularities of j_0, j_1 filled in at zero. Their leading endpoint values are $A_\infty e^{i\Omega c}$ and q_∞, respectively, because the leading differential equations are $y' = (a + i\omega c)y$ and $w' = 2aw + \sigma_\rho^2$, with the stated initial values.

On the compact set $\rho \in [0, \rho_0]$, q_∞ is bounded below by a positive constant by (100). Thus the square-root cost is smooth uniformly near the limiting covariances. Applying the lemma to both recurrences, substituting in (82), and integrating over c yields

$$\mathcal{E}_S(\rho) = \mathcal{C}_{\rho,\infty} - \Lambda_\rho h + B(\rho) h^2 + O(h^3), \tag{106}$$

where B is smooth and the remainder is uniform. The already calculated first-order coefficient is $-\Lambda_\rho$ for $\rho > 0$, and continuity extends that identity to zero. This deduction uses the endpoint recurrences, not an assumed decomposition of squared errors.

To calculate $B(0)$, use $r_i = 1$, $q_i = q/S$, and set

$$\alpha(c) = \omega^2 T(c^2 - 1/3)/2.$$

Uniformly in c,

$$M_{\omega h}(c)^S = R(\Omega c)[I + h\alpha(c) I + O(h^2)],$$
$$\widehat{q}_S(c) = \frac{q}{S} \sum_{j=0}^{S-1} d_{\omega h}(c)^j = q[1 + h\alpha(c)] + O(h^2).$$

The conditional mean contributes $2s_0^2 h^2 \alpha(c)^2 + O(h^3)$ to the risk. The covariance contributes $(q/2)h^2\alpha(c)^2 + O(h^3)$. Since $\mathbb{E}_c(c^2 - 1/3)^2 = 4/45$,

$$B(0) = (2s_0^2 + q/2)\mathbb{E}_c \alpha(c)^2 = \frac{\omega^4 T^2 (4s_0^2 + q)}{90} = B_{\mathrm{fit}}.$$

Smoothness gives $B(\rho) = B_{\mathrm{fit}} + O(\rho)$.

Finally, uniformly for $t \in [0, T]$, $a_\rho(t) = \rho u(t) + O(\rho^2)$, so

$$A_\infty = 1 + \rho U + O(\rho^2), \qquad d_\rho = \rho D + O(\rho^2).$$

Writing $A = 1 + \epsilon$ in the definition of $\mathcal{C}(A)$ above gives

$$\mathcal{C}(1 + \epsilon) = 2s_0^2(1 + s_0^2/q)\epsilon^2 + O(\epsilon^3).$$

Therefore

$$\Lambda_\rho = A_\infty d_\rho \mathcal{C}'(A_\infty) = 4s_0^2(1 + s_0^2/q) U D\, \rho^2 + O(\rho^3) = \lambda_2 \rho^2 + O(\rho^3).$$

Substitution in (106) proves (88). Positivity follows from $u(t) > 0$ for $t > 0$ and $\omega \neq 0$. □

Proof of Theorem 14. Theorem 13 gives a finite global minimizer for every sufficiently small positive ρ. Let

$$x_\star = \lambda_2/(2B_{\text{fit}}) > 0$$

and choose an integer S_ρ nearest to $T/(x_\star\rho^2)$. Then $T/S_\rho = x_\star\rho^2 + O(\rho^4)$, and (88) gives

$$\mathcal{E}_{S_\rho}(\rho) - \mathcal{C}_{\rho,\infty} = -\frac{\lambda_2^2}{4B_{\text{fit}}}\rho^4 + o(\rho^4). \tag{107}$$

In particular this candidate has strictly smaller risk than the infinitely deep cascade limit.

Every sequence of global minimizers satisfies $S^\star(\rho) \to \infty$. Otherwise a subsequence has one fixed finite index S, while the candidate risk tends to zero and $\mathcal{E}_S(\rho) \to \mathcal{E}_S(0) > 0$. The last strict inequality follows already from the mean term:

$$(j_0(\omega T/S) + 3ij_1(\omega T/S)c)^S$$

is a finite polynomial in c and cannot equal $e^{i\Omega c}$ on an interval for $\Omega \neq 0$.

Put $h^\star = T/S^\star$. The uniform expansion and $h^\star \to 0$ imply, for sufficiently small ρ,

$$\mathcal{E}_{S^\star}(\rho) - \mathcal{C}_{\rho,\infty} \geq -C\rho^2 h^\star + \tfrac{1}{2}B_{\text{fit}}(h^\star)^2.$$

The left side is negative by (107), so $h^\star \leq C'\rho^2$. For bounded $x = h/\rho^2$, the same expansion is uniformly

$$\frac{\mathcal{E}_S(\rho) - \mathcal{C}_{\rho,\infty}}{\rho^4} = -\lambda_2 x + B_{\text{fit}}x^2 + o(1).$$

The quadratic has the unique minimum at $x_\star$. Comparison with (107) forces every minimizing sequence to satisfy

$$\frac{h^\star}{\rho^2} \to x_\star,$$

proving the cascade-depth limit (89). Evaluation of that minimum gives (90). □