FLOW MAPS THAT COMPOSE EXACTLY: WHEN CONSISTENCY COSTS ACCURACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned dynamics are often required to compose across time: a direct forecast should agree with a sequence of shorter forecasts. Deterministic dynamics on a sufficient state and Markov transition laws obey this identity, but conditional means, the targets of squared-error training, generally do not. We quantify the price of this mismatch. Under coverage and regularity assumptions, an exactly composing predictor incurs joint excess risk at least proportional to the squared composition defect of the true means. Matching examples attain the smoothness rate and small-noise risk order. A generator criterion identifies diffusions whose means compose, including nonlinear, noisy cases. Transformed linear diffusions provide compatible kernels with explicit representation limits. Fixed-model experiments isolate the forecasting consequences. On noisy Lorenz dynamics, propagating a residual distribution through the same fitted one-step map removes most of the long-horizon excess risk of its point rollout; a single intermediate collapse to the mean restores it. Without noise, autoregression is highly accurate. On motion capture, discarding a learned kernel's intermediate uncertainty degrades its forecasts, although a deterministic map remains the best point predictor. The guiding distinction is what should compose: under uncertainty, transition laws compose even when their means do not.

1 INTRODUCTION

Learning an evolution map makes it possible to predict across a time interval without evaluating every intermediate step. A natural requirement is that a direct prediction agree with the composition of shorter predictions. If $F_{s,t}$ maps the state at time s to time t, this requirement is

$$F_{s,t} = F_{u,t} \circ F_{s,u}, \qquad s \le u \le t. \tag{1}$$

Invertible conjugate constructions can enforce this identity at every parameter value (Bizzi et al., 2025; Mehouachi & Jabari, 2025). Learned flow maps, shortcut models, and consistency models use objectives tied to the underlying evolution or transport (Song et al., 2023; Frans et al., 2025; Boffi et al., 2025b; Huang et al., 2026). An autoregressive map also composes on integer multiples of its time step. These approaches raise a more basic question: *does the quantity being predicted obey this law?*

For squared-error prediction, the population target is the conditional mean $m_{s,t}(x) = \mathbb{E}[X_t \mid X_s = x]$. A Markov process obeys

$$m_{s,t}(x) = \mathbb{E}[m_{u,t}(X_u) \mid X_s = x], \tag{2}$$

which generally differs from $m_{u,t}(m_{s,u}(x))$. Replacing an intermediate distribution by its mean discards information used by a nonlinear continuation (Figure 1). This underlies the classical distinction between recursive and direct nonlinear forecasting (Ben Taieb & Hyndman, 2012; Ben Taieb & Atiya, 2016). For an exactly composing model, a nonzero mean-composition defect creates a structural mismatch: under coverage and regularity assumptions, the model cannot match the optimal point predictions at all three time pairs simultaneously.

The resolution is to ask what should compose. Markov transition laws obey Chapman–Kolmogorov; their means generally do not. Deterministic transport, including probability-flow ODEs, obeys pointwise composition, so our incompatibility bound imposes no positive cost. Under uncertainty,

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

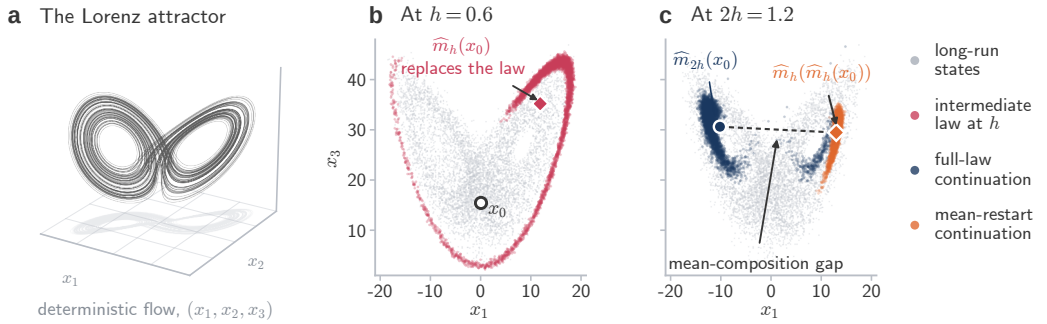

Figure 1: Restarting at the mean changes the forecast. **a,** Deterministic Lorenz geometry for context. **b,** A noisy Lorenz ensemble at $h = 0.6$ and its estimated mean. **c,** Continuing that ensemble to $2h = 1.2$ differs from restarting at its mean; the final means are 34.75 apart in 3D (23.19 in the displayed projection). This selected-state Monte Carlo illustration uses a state with a large gap; Section H gives the selection and simulation protocol.

composing laws avoids the mean-target mismatch, while representation and estimation errors remain. Ensemble forecasting has long used the corresponding inference rule: propagate an ensemble and average at the end (Leith, 1974).

Contributions.

1. *The price of composing means.* A population risk–consistency bound, a sharp smoothness frontier, and matching lower and upper small-noise risk orders (Theorem 1 and Propositions 1 and 5).
2. *When means compose.* A generator criterion and local expansion identify the role of drift curvature and diffusion, including nonlinear noisy exceptions (Section 3).
3. *Composing laws.* A compatible kernel family from classical stochastic transformations, its covariance schedule, and a scalar representation criterion (Sections 4 and 5).
4. *The forecasting consequence.* Fixed-model interventions isolate uncertainty propagation on Lorenz and motion data; separately tuned predictors assess fitted accuracy (Sections 6 and L).

2 The Prediction Cost of Pointwise Composition

Fix three times $s < u < t$. Let (X_s, X_u, X_t) be square-integrable, with marginal laws μ_s, μ_u, μ_t, and fix measurable versions of the conditional means $m_{s,u}, m_{u,t}, m_{s,t}$. All predictors are measurable with $F_{i,j}(X_i) \in L^2$. Define

$$\Delta = \|m_{u,t} \circ m_{s,u} - m_{s,t}\|_{L^2(\mu_s)}, \qquad d_F = \|F_{u,t} \circ F_{s,u} - F_{s,t}\|_{L^2(\mu_s)}. \tag{3}$$

The first quantity is a property of the prediction targets; the second is a property of the model. For each time pair $i < j$, conditional expectation gives the orthogonal decomposition

$$\underbrace{\mathbb{E}\,\|F_{i,j}(X_i) - X_j\|^2 - \mathbb{E}\,\|m_{i,j}(X_i) - X_j\|^2}_{\text{excess prediction risk } \mathcal{E}_{i,j}} = \|F_{i,j} - m_{i,j}\|^2_{L^2(\mu_i)}. \tag{4}$$

Thus the risks below exclude irreducible conditional variance.

Distribution shift at the intermediate time. Let $\nu = (m_{s,u})_{\#}\mu_s$ and $\widetilde{\mathcal{E}}_{u,t} = \|F_{u,t} - m_{u,t}\|^2_{L^2(\nu)}$. The basic bound, before any coverage assumption, is

$$\Delta \le d_F + L\sqrt{\mathcal{E}_{s,u}} + \sqrt{\widetilde{\mathcal{E}}_{u,t}} + \sqrt{\mathcal{E}_{s,t}}. \tag{5}$$

It holds whenever these terms are finite and the two-point Lipschitz inequality below holds. It identifies the relevant distribution shift directly: ν contains intermediate means, while training observes μ_u. Coverage converts the shifted error into an on-distribution risk.

Theorem 1 (Risk–consistency tradeoff). *Suppose* $(m_{s,u})_{\#}\mu_s \le C\mu_u$ *as measures, for* $C < \infty$, *and* $\|F_{u,t}(m_{s,u}(x)) - F_{u,t}(F_{s,u}(x))\| \le L\|m_{s,u}(x) - F_{s,u}(x)\|$ *for* μ_s*-almost every* x. *Then*

$$\Delta \le d_F + L\sqrt{\mathcal{E}_{s,u}} + \sqrt{C\mathcal{E}_{u,t}} + \sqrt{\mathcal{E}_{s,t}}. \tag{6}$$

In particular, with $\mathcal{R} = \mathcal{E}_{s,u} + \mathcal{E}_{u,t} + \mathcal{E}_{s,t}$,

$$\mathcal{R} \;\ge\; \frac{(\Delta - d_F)_+^2}{L^2 + C + 1}. \tag{7}$$

Proof. Insert $F_{u,t}(m_{s,u}(x))$, $F_{u,t}(F_{s,u}(x))$, and $F_{s,t}(x)$ into the difference defining Δ. The resulting four terms are, respectively, a second-leg approximation error at a conditional mean, propagation of the first-leg approximation error, model composition error, and direct approximation error. Minkowski's inequality, the coverage assumption, and the Lipschitz bound give (6). Cauchy–Schwarz gives (7). □

If the architecture composes exactly, $d_F = 0$: a nonzero target defect enforces positive excess risk for every predictor in a class with the stated constants. If the predictor approximates all three conditional means closely, it must instead retain composition error near Δ. For a class constrained by $d_F \le \eta$, the lower bound is $(\Delta - \eta)_+^2/(L^2 + C + 1)$. This is a compatibility result, independent of sample size, optimization method, or the use of a consistency penalty. Relating recursive risk to Δ^2 also requires error control at intermediate means (Equation (22)).

Why coverage is necessary. The second leg is evaluated at $m_{s,u}(X_s)$, while transition training typically sees actual states X_u. These distributions need not coincide. The assumption controls this shift and makes the composition of the chosen conditional-mean versions well-defined up to null sets. It can fail, for example when a fixed initial state yields one intermediate mean inside a continuous distribution. In that case an $L^2(\mu_u)$ error alone does not control the predictor at that point. For $\nu = \mathcal{N}(0, sI_d)$ and $\mu_u = \mathcal{N}(0, vI_d)$, a variance ratio $v/s > 1$ gives $C = (v/s)^{d/2}$, which can grow exponentially with dimension. In high dimension, (5) retains the shifted error explicitly; coverage alone need not give a useful constant.

A separate uniform-error alternative requires no distributional coverage (Proposition 7, proved independently in Section B).

Why smoothness is substantive. An autoregressive predictor iterated at a fixed time step is already exactly compositional on integer multiples of that step. Thus the bound also applies to joint one-step and rollout risks, not only specialized conjugate architectures. The dependence on regularity cannot be removed. Consider the chain on $\{-1, 0, 1\}$ with uniform initial law and transition matrix

$$P = \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}. \tag{8}$$

For the three-state chain in (8), use one map f for each unit step and $f \circ f$ for two steps, and let $\mathcal{R}_*(L)$ be the infimum of the three-pair excess risk over all $f : \mathbb{R} \to \mathbb{R}$ with $\mathrm{Lip}(f) \le L$. Invertibility is not imposed here; this class includes general autoregressive point maps.

Proposition 1 (Sharp smoothness frontier). *For this chain and every* $L \ge 2$,

$$\frac{27}{64L^2 + 240} \;\le\; \mathcal{R}_*(L) \;\le\; \frac{27}{64L^2} + \frac{39}{64L^4}. \tag{9}$$

Consequently $\lim_{L\to\infty} L^2\mathcal{R}_*(L) = 27/64$. *Without a Lipschitz constraint, the infimum over measurable one-step maps is zero, but is not attained.*

The upper bound is an explicit piecewise-linear construction, and the lower bound follows from one necessary Lipschitz constraint; neither uses a numerical optimizer (Section B.3). The general lower bound in Theorem 1 is $3/[16(L^2 + 3)]$, with the same rate and an asymptotic constant smaller by exactly $9/4$. Figure 2 displays the resulting frontier. The construction moves $f(-1)$ a distance $3/(4L)$ from its true mean, then uses slope L near that mean, trading a small one-step displacement for accurate two-step predictions.

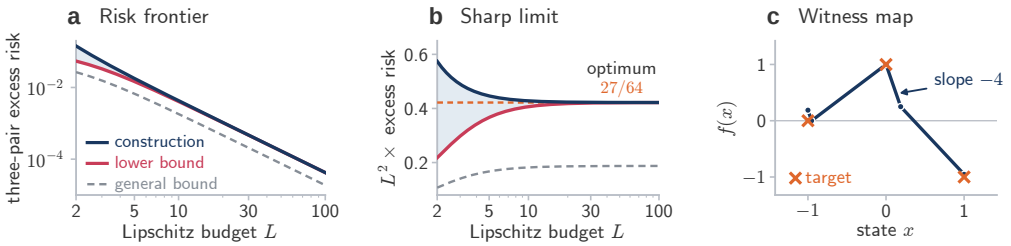

Figure 2: Smoothness sets the price of exact composition. a, Analytic lower and upper bounds enclose the best three-pair excess risk over globally L-Lipschitz one-step maps; shading marks this certified enclosure. **b,** Multiplying by L^2 identifies the limiting constant $27/64$, not the finite-L optimum. **c,** The $L = 4$ witness shifts $f(-1)$ from its true mean and uses a segment of slope -4 to reach the two-step target. Crosses mark one-step targets; the witness need not be invertible.

A mean-preserving consistency target. The Markov identity also suggests a population alternative to feeding a predicted mean into another predictor. For a fixed square-integrable teacher G, regression on the observed intermediate state gives

$$\arg\min_H \mathbb{E}\|H(X_s) - G(X_u)\|^2 = \mathbb{E}[G(X_u) \mid X_s = \cdot]. \tag{10}$$

At $G = m_{u,t}$, the minimizer is $m_{s,t}$ by the Markov identity (2). Thus bootstrapping from observed intermediate states with a fixed or stop-gradient teacher preserves the true means as a population fixed point. Substituting $F_{s,u}(X_s)$ introduces the expectation gap. This temporal-difference-style target (Sutton, 1988; Farebrother et al., 2025) does not guarantee convergence with shared finite networks or a jointly updated teacher.

3 When conditional means compose

For a time-homogeneous Markov process, write $P_h f(x) = \mathbb{E}[f(X_h) \mid X_0 = x]$ and $m_h = P_h \,\mathrm{id}$. Its transition operators satisfy $P_{h+k} = P_h P_k$, hence

$$m_k(m_h(x)) - m_{h+k}(x) = m_k(\mathbb{E}[X_h \mid x]) - \mathbb{E}[m_k(X_h) \mid x]. \tag{11}$$

This vector-valued expectation gap can vanish without the chart or the dynamics being affine. If $\|D^2 m_k\| \le M$ on the relevant convex domain, Taylor's theorem bounds its norm by $(M/2)\mathbb{E}[\|X_h - m_h(x)\|^2 \mid x]$. This is an upper bound; curvature and noise do not by themselves guarantee a nonzero gap.

A generator criterion. Consider the diffusion $dX = b(X)\,dt + \sigma(X)\,dW$, with $a(x) = \sigma(x)\sigma(x)^\top$ and generator

$$\mathcal{L}f = b \cdot \nabla f + \tfrac{1}{2} a : \nabla^2 f. \tag{12}$$

For vector functions, the Hessian contraction is componentwise. Assume the deterministic drift has a unique forward-complete flow Φ_h, continuously differentiable in time and C^2 in space, that the conditional means are classical solutions of the backward equation, and that this equation has a unique solution in a growth class containing the functions under consideration; precise uses are in Section C. A differentiable composition family has $m_0 = \mathrm{id}$, $m_{h+k} = m_k \circ m_h$, and trajectories $h \mapsto m_h(x)$ continuously differentiable, with right derivative $b(x)$ at zero.

Proposition 2 (Mean-flow criterion)**.** *Under these assumptions, the means form a differentiable composition family if and only if*

$$a(x) : \nabla^2 \Phi_h(x) = 0 \qquad \textit{for all } x \in \mathbb{R}^d,\ h \ge 0. \tag{13}$$

When this holds, $m_h = \Phi_h$. Affine drift is sufficient, regardless of the state dependence of the diffusion, provided the stated moment and uniqueness assumptions hold.

Indeed a differentiable mean semigroup must have infinitesimal generator b on states and hence must equal its ODE flow. Comparing the ODE flow's transport equation with the diffusion backward equation gives (13). This criterion concerns all lags; cancellation of one Taylor coefficient is only a necessary local condition.

Proposition 3 (Local mean-composition defect)**.** *Suppose the conditional means admit the second-order expansion with a locally uniform third-order remainder and coefficient regularity stated in Section C.2. As $h \downarrow 0$,*

$$m_h(m_h(x)) - m_{2h}(x) = -\frac{h^2}{2}\, a(x) : \nabla^2 b(x) + O(h^3). \tag{14}$$

Here $\mathcal{L}b = Db\, b + \frac{1}{2}a : \nabla^2 b$: the drift-only contribution cancels in the composition, leaving the diffusion-weighted curvature term. This is the continuous-time counterpart of recursive-forecast curvature bias (Ben Taieb & Atiya, 2016). A nonzero coefficient establishes a small-time obstruction to composing conditional means.

A nonlinear stochastic exception. Let Z be standard Brownian motion and $X = \sinh Z$. Its chart is nonlinear and its noise is nonzero, yet

$$m_h(x) = \mathbb{E}[\sinh(\operatorname{arsinh} x + W_h)] = e^{h/2}x, \qquad m_k \circ m_h = m_{h+k}. \tag{15}$$

Itô's formula gives $dX = \frac{1}{2}X\, dt + \sqrt{1+X^2}\, dW$, consistent with the affine-drift case of Proposition 2. Neither nonlinearity nor noise alone forces a defect.

Non-affine drift can also compose in higher dimension. With independent Brownian coordinates, consider $dX_1 = dW_1$, $dX_2 = dW_2$, and $dX_3 = (X_1^2 - X_2^2)\, dt + dW_3$. The drift is non-affine and the noise is nondegenerate, yet $m_h(x) = (x_1, x_2, x_3 + h(x_1^2 - x_2^2))$ composes exactly: the two equal Brownian variance contributions cancel. Equivalently, each coordinate of the drift flow is harmonic, so (13) holds. This illustrates why drift curvature must be contracted with the covariance, rather than merely declared nonzero.

4 From conjugate maps to compatible transition kernels

Deterministic starting point. With invertible charts ϕ_t and invertible matrices $\Lambda(t)$, the maps

$$F_{s,t} = \phi_t^{-1} \circ \Lambda(t)\Lambda(s)^{-1} \circ \phi_s \tag{16}$$

compose by cancellation. Every invertible evolution family on a finite interval admits this algebraic form: at terminal time r, set $\phi_t = F_{t,r}$ and $\Lambda = I$ (Section D). This does not supply finite-network efficiency. A fixed chart with $F_h = \phi^{-1} \circ e^{hA} \circ \phi$ is more restrictive.

A compatible stochastic extension. Keep a fixed diffeomorphism ϕ, write $\psi = \phi^{-1}$, and use the affine latent transition

$$Z_{t+h} = R_h Z_t + r_h + \eta_h, \quad R_h = e^{hA}, \quad r_h = \int_0^h e^{vA} c\, dv, \quad \eta_h \sim \mathcal{N}(0, \Sigma_h), \tag{17}$$

with independent innovations on successive intervals. The corresponding kernel is $P_h(x, B) = \Pr\{\psi(R_h\phi(x) + r_h + \eta_h) \in B\}$. This is a transformed linear diffusion, a special case of the stochastic-process flow framework (Deng et al., 2020; 2021).

Proposition 4 (The Gaussian covariance schedule)**.** *Let $(\Sigma_h)_{h\geq 0}$ be a continuous positive-semidefinite schedule with $\Sigma_0 = 0$. For (17), Chapman–Kolmogorov holds at every nonnegative pair of lags if and only if, for some $Q \succeq 0$,*

$$\Sigma_{h+k} = R_k \Sigma_h R_k^\top + \Sigma_k, \quad \textit{equivalently} \quad \Sigma_h = \int_0^h e^{vA} Q e^{vA^\top}\, dv. \tag{18}$$

Invertible coordinate changes preserve this identity. At $Q = 0$ the transition kernels reduce to Dirac measures at the deterministic affine conjugate maps.

This is the classical linear-diffusion covariance law (Särkkä & Solin, 2019). Gaussian convolution proves the covariance identity; continuity of the cocycle already implies the right derivative Q exists, yielding the integral formula (Section D.2). The mean law already holds because $R_{h+k} = R_k R_h$ and $r_{h+k} = R_k r_h + r_k$. Thus, *within this latent Gaussian family*, choosing infinitesimal noise Q determines all horizon covariances. Arbitrary per-horizon noise calibration does not supply the same guarantee.

For $h > 0$ with $\Sigma_h \succ 0$, the transition density is

$$p_h(y \mid x) = \mathcal{N}\big(\phi(y); R_h\phi(x) + r_h, \Sigma_h\big)\,|\det D\phi(y)|\,. \tag{19}$$

It supports multi-lag maximum likelihood using a shared chart, drift, and noise generator. Singular covariances still define kernels but need not define Lebesgue densities. Positive noise also removes the deterministic interpretation of a negative lag as an inverse transition. Equality of direct and composed *laws* is exact; samplewise equality requires coupling the direct innovation as $\eta_{h+k} = R_k\eta_h + \eta_k$.

The conditional means still need not compose. For $c = 0$, let $Q = \epsilon Q_0$, $z_1 = R_h\phi(x)$, $\bar{z} = R_k z_1$, and $S_h = \Sigma_h/\epsilon$. Under the smoothness and tail conditions in Section D.3,

$$\begin{aligned} m_k(m_h(x)) - m_{h+k}(x) &= \epsilon\mathcal{D}_{h,k}(x) + O(\epsilon^2), \\ \mathcal{D}_{h,k}(x) &= \tfrac{1}{2}\left\{D\psi(\bar{z})R_k[D\psi(z_1)]^{-1}H\psi(z_1)[S_h] - H\psi(\bar{z})[R_k S_h R_k^\top]\right\}, \end{aligned} \tag{20}$$

where $H\psi(z)[S] = \sum_{ij}\partial_{ij}\psi(z)S_{ij}$. The coefficient compares the propagation of the first-leg mean correction with the effect of the propagated covariance. It can cancel for nonlinear charts, as (15) shows. The extension resolves a mismatch of *prediction objects*; it does not alter the true conditional means or force their composition defect to vanish.

Proposition 5 (Matching small-noise order)**.** *Let $dZ = aZ\,dt + \sqrt{\epsilon}\,dW$ with $a < 0$, $Z_0 \sim \mathcal{N}(0, \tau^2)$, $\tau > 0$, and $X = \sinh Z$. Fix $h, k > 0$ and $L \geq e^{ak}$. Let $\mathcal{A}_L$ be the class of measurable triples (F_h, F_k, F_{h+k}) with $F_{h+k} = F_k \circ F_h$ and $\mathrm{Lip}(F_k) \leq L$, with finite risks under the actual laws at times $0, h$. Then, as $\epsilon \downarrow 0$,*

$$\inf_{F\in\mathcal{A}_L}\left(\mathcal{E}_{0,h} + \mathcal{E}_{h,h+k} + \mathcal{E}_{0,h+k}\right) = \Theta(\epsilon^2).$$

The upper bound is attained in order by the deterministic zero-noise conjugate family.

The fixed budget $L \geq 1$ is a special case. The proof verifies coverage and a matching construction (Section B.4).

5 Representation Limits and State Information

The chart cannot linearize every diffusion. Consider $dX = b(X)\,dt + \sigma(X)\,dW$ on an open interval I, with $\sigma > 0$, $b \in C^1$, and $\sigma \in C^2$. Suppose the Lamperti coordinate $\ell(x) = \int_{x_*}^{x}\sigma(v)^{-1}\,dv$ maps I onto $\mathbb{R}$, and assume nonexplosion and the usual uniqueness conditions. The following specializes the classical Lamperti transformation (Kloeden & Platen, 1992; Møller & Madsen, 2010) to affine latent dynamics.

Proposition 6 (Scalar affine-latent representation)**.** *A time-independent increasing C^2 chart transforms X into $dZ = (\alpha Z + \gamma)\,dt + \rho\,dW$, with $\rho > 0$, if and only if*

$$\sigma(x)\frac{d}{dx}\left(\frac{b(x)}{\sigma(x)} - \frac{\sigma'(x)}{2}\right) = \alpha \quad \textit{is constant on } I. \tag{21}$$

Every such chart has the form $\phi = \rho\ell + \beta$.

The proof matches the drift and diffusion coefficients in Itô's formula (Section E). If σ is a nonzero constant, the chart must be affine and b must be affine. In particular $b(x) = x - x^3$ with constant noise fails (21), since $b'(x) = 1 - 3x^2$ is not constant. A flexible invertible chart may approximate selected double-well transition laws, but cannot make this fixed-chart, same-dimensional affine diffusion exact at all horizons. Matching a stationary density alone does not establish a dynamical representation.

Table 1: Oracle supervision. Three-pair excess risk, mean ± SD over three initialization seeds. The last column averages per-fit theorem bounds at $q = 1$, using an analytic coverage upper bound and global Lipschitz bounds. A dash indicates that exact point composition is not imposed. All entries are quadrature evaluations.

Model	Params.	$q = 0$	$q = 1$	Bound, $q = 1$
Cubic composing map	2	$(2.97 \pm 2.24) \times 10^{-15}$	$(3.82 \pm 0.00) \times 10^{-2}$	0.00332
Compatible cubic mean	3	$(1.76 \pm 0.20) \times 10^{-12}$	$(4.89 \pm 0.82) \times 10^{-15}$	—
Neural autoregression	4417	$(2.68 \pm 1.19) \times 10^{-6}$	$(4.96 \pm 0.01) \times 10^{-3}$	2.11×10^{-5}
Direct MLP	4417	$(1.50 \pm 0.41) \times 10^{-4}$	$(9.03 \pm 7.84) \times 10^{-5}$	—

Noise does not replace missing state. Pairwise observed conditionals can themselves violate CK when unresolved variables retain memory. For $Y_t = U + tV$ with independent standard Gaussian U, V, the observed $0 \to 2$ kernel differs from the composition through time 1; retaining V restores deterministic composition (Section E.2). Thus a compatible learned kernel requires a sufficient state for an exact physical interpretation. Our motion model uses a finite delay state as an approximation.

6 Controlled Experiments

Oracle regression tests the approximation cost with exact targets. Lorenz and motion-capture experiments then compare fitted predictors and test the effect of intermediate mean collapse. Learned studies report means and sample standard deviations over three seeds; the appendix gives the protocols and results at every evaluated horizon.

6.1 Oracle supervision isolates target incompatibility

For $dZ = -Z\,dt + \sqrt{q}\,dW$ and $X = \psi_c(Z) = Z + cZ^3$ ($c = 3/5$), all models minimize the same three-pair excess risk at $0 \to 0.5$, $0.5 \to 1$, and $0 \to 1$ using population quadrature. The cubic map and neural autoregression compose exactly; the compatible mean need not. At $q = 1$, their fitted risks are 0.0382, 0.00496, and below 6×10^{-15}, respectively (Table 1). The mean-composition gap holds for every $c > 0$ with positive noise, away from $x = 0$ (Section I.1). To apply Theorem 1, we verify a conservative analytic coverage bound $C \leq 1.995$. At the cubic map's fitted Lipschitz budget, the resulting lower bound is about 0.00332. The looser neural bound illustrates sensitivity to regularity, not a failure of the inequality. At $q = 0, 0.1$ the direct MLP is worse than neural autoregression, so finite fitting does not yield a universal ranking (Sections I and J.5).

6.2 Lorenz: what noise changes in a point rollout

We compare horizon-specific predictors (heads) with autoregression (AR), which iterates one map. Direct heads regress future observations; tower heads regress a shorter-horizon head evaluated at an observed intermediate state; point heads regress targets obtained by evaluating the shorter-horizon head at a predicted mean. We separately tune ridge penalties, train AR rollouts, and test two feature budgets. Methods share data and features; AR stores one map, the head-based methods six.

At unit noise and $h = 1.6$, the larger family gives excess risks 48.04 ± 3.58 for rollout-trained AR, 5.72 ± 0.10 for direct heads and 4.47 ± 0.31 for tower heads. Tower's 90.7% excess-risk gain over AR becomes 27.3% in total risk after adding the estimated irreducible variance 111.60. All seeds and both budgets favor tower over rollout AR at this horizon. Without noise, tuned AR gives 0.006 versus tower's 41.99; full results and refinement checks are in Section J.4.

Stochastic propagation with a fixed map. For the shared one-step fit in Section I, add training-residual Gaussian noise to the map f: $K(x, \cdot) = \mathcal{N}(f(x), \widehat{\Sigma})$. Iterating this discrete-time kernel with 512 particles reduces lag-1.6 excess risk from 61.69 ± 9.26 to 0.60 ± 0.22. With 256 particles and matched random numbers, one mean collapse at 0.8 raises terminal risk from 0.80 ± 0.17 to 65.43 ± 8.98. Nothing is refitted, and the scores include finite-ensemble Monte Carlo error (Section K). The separately tuned predictors above address a different comparison. The statewise association in Figure 3c is empirical; the fitted-model risk increase need not equal the physical mean defect.

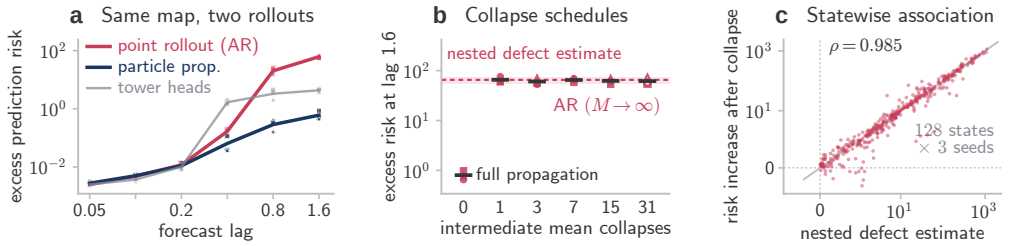

Figure 3: One fitted map, different propagation rules. a, Point rollout and 512-particle propagation of the same one-step map; tower heads are a separately fitted reference. Symbols show three data seeds. **b,** Terminal risk with 256 particles and intermediate mean collapses; open AR symbols denote the infinite-particle every-step-collapse limit. Only the zero/one-collapse comparison uses matched random numbers. Shading is the nested defect estimate, mean ± seed SD. **c,** Paired risk increase versus that estimate for 128 states per seed, on signed symmetric-log axes; $\rho = 0.985$ is Pearson correlation on raw values. The estimator retains plug-in and discretization error; its references are coupled to those used for the risks (Section K).

Table 2: CMU forecasting. Mean ± SD across three model seeds; lower is better. MSE averages squared error in the original 50-channel coordinates. ES is the multivariate energy score, also defined for deterministic predictions. Capacities and objectives differ.

	Setup 1		Setup 2	
Model	MSE	ES	MSE	ES
Direct point MLP	22.62 ± 1.03	30.66 ± 0.55	17.95 ± 2.11	25.58 ± 0.27
Composing point map	8.91 ± 1.22	18.71 ± 0.94	4.68 ± 0.36	14.45 ± 0.58
Compatible kernel	14.21 ± 3.00	17.80 ± 2.09	8.87 ± 0.67	13.93 ± 0.60
Loose density	19.76 ± 1.71	21.82 ± 1.02	8.47 ± 0.55	13.84 ± 0.55

6.3 Real motion data: fitting versus propagating uncertainty

We use the public 50-channel CMU motion sequences and original 16/3/4 train/validation/test split from Yıldız et al. (2019). All models consume the last three poses (a 150-dimensional delay state). Setup 1 trains on 300-frame sequences and predicts 297 frames from three context frames; setup 2 trains on 200-frame prefixes and predicts frames 100–299. Normalization uses training data only. Point models minimize MSE, density models minimize NLL, and validation selects checkpoints within the fixed training budget (Section J.1).

Forecast accuracy. The exactly composing deterministic map wins MSE in both setups and all seeds (Table 2). The compatible kernel improves over the direct MLP and gives better 90% interval coverage than the loose density model, whose covariance need not obey CK; the latter slightly wins mean MSE and energy score in setup 2. These comparisons use one shared evaluation protocol.

Intermediate mean collapse. Holding the kernel parameters and initial state fixed removes retraining and model-capacity differences. At eight steps, collapsing seven intermediate ensembles raises MSE from 14.214 to 15.405 (8.4%) and energy score from 17.792 to 20.244 (13.8%). Coverage falls from 80.5% to 61.9%, and predictive variance falls by 56.9%. Both score increases hold in all three model seeds and persist with 256 instead of 512 particles. The effect is not uniform across individual test sequences, and the baseline itself under-covers (Section J.2). Thus the intervention tests the value of retaining a distribution within the same fitted forecasting system.

On CMU, stronger point consistency regularization improves both mean MSE and measured validation consistency, revealing no empirical frontier. The true composition defect is unmeasured there; fitted rankings alone cannot identify the regime (Section J.3). Warped-OU and double-well fits likewise show that exact CK does not guarantee best fitted accuracy (Section G).

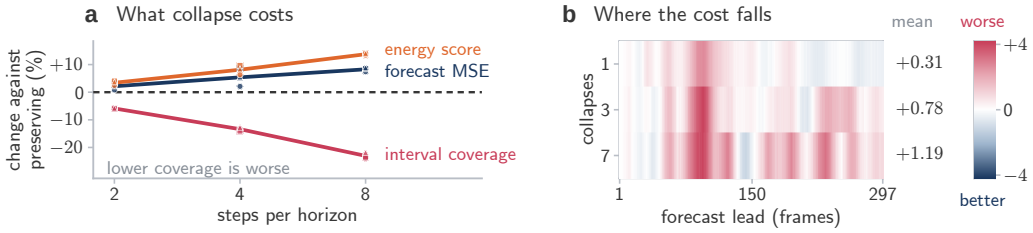

Figure 4: The cost of intermediate mean collapse in a fixed CMU kernel. a, Within-seed changes from preserving the ensemble, averaged over two 512-particle repeats. Symbols mark three seeds; lines show their mean. Coverage changes are relative percent, not percentage points. **b,** Collapse-minus-preserve MSE by lead, averaged over seeds and repeats; right-hand values average each row. Each horizon uses 2, 4, or 8 steps (1, 3, or 7 collapses), with a stochastic final step. Blue cells indicate improved MSE.

7 Related Work

Forecasting, operators, and state. Recursive forecasting bias and ensemble propagation are classical (Ben Taieb & Hyndman, 2012; Ben Taieb & Atiya, 2016; Leith, 1974). Stochastic Koopman operators compose on observables (Črnjarić-Žic et al., 2020); their action on the identity need not yield composing means. Related studies address MSE versus marginal realism (Green et al., 2026), conditional-mean and estimation barriers (Chen, 2026; Green et al., 2025), and memory introduced by projection (Chorin et al., 2000).

Flow maps and stochastic representations. Flow Matching, Rectified Flow, and Neural Flows learn transport or evolution (Lipman et al., 2023; Liu et al., 2023; Biloš et al., 2021). Consistency, shortcut, and distilled flow-map methods concern finite-time transport (Song et al., 2023; Frans et al., 2025; Boffi et al., 2025b;a). Recursive Flow Matching enforces consistency across discretization scales of a generative forecaster (Huang et al., 2026). Semigroup learning and diagnostics (Chen & Wu, 2023; Shikhman, 2026; Liu et al., 2026) depend on the chosen state and target. Recent stochastic flow maps condition on driving-noise information to approximate solution paths (Pan et al., 2026; McCallum et al., 2026); those inputs fall outside our state-only mean class.

Conjugate dynamics (Bizzi et al., 2025; Mehouachi & Jabari, 2025), warped Gaussian processes (Snelson et al., 2003; Maroñas et al., 2021), and transformed stochastic processes (Deng et al., 2020; 2021; El Bekri et al., 2024) establish the ingredients of our kernel family. Alternative inference methods include latent SDEs, continuous-discrete state-space models, SDE Matching, and Neural Stochastic Flows (Li et al., 2020; Ansari et al., 2023; Bartosh et al., 2025; Kiyohara et al., 2025).

Limitations. Large coverage or Lipschitz constants weaken the population bound, which does not establish finite-sample superiority. The frontier uses finite support and the noise order is asymptotic. Lorenz head comparisons have unequal capacity. CMU uses one split, four test sequences and three seeds, with different training objectives; its delay state need not be Markov or its pose projection CK-consistent. Exact composition alone guarantees neither stability nor calibration.

8 Conclusion

Composition is a property of the prediction target. Markov transition laws compose; their conditional means generally do not. Our bound and sharp examples quantify the resulting prediction cost under coverage and regularity conditions. Point forecasting can benefit from preserving uncertainty through nonlinear continuations: ensemble propagation through one fixed map removed most of the noisy Lorenz excess risk, while mean collapse degraded a learned motion forecaster. This mechanism does not determine fitted-model rankings; a deterministic map won point accuracy on motion capture. Nor can noise replace a sufficient state when unresolved variables carry memory. The prediction target and available state should determine the consistency constraint.

AI USE STATEMENT

Claude and Codex were used throughout this work: developing and checking the mathematics, literature search, code and experiments, and drafting and editing the text. Grammarly was used for spelling and grammar. The authors take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

The supplementary material contains notebooks that reproduce the main results of the paper. The full code will be released on GitHub upon acceptance.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

Abdul Fatir Ansari, Alvin Heng, Andre Lim, and Harold Soh. Neural continuous-discrete state space models for irregularly-sampled time series. In *International Conference on Machine Learning*, volume 202, pp. 926–951, 2023.

Grigory Bartosh, Dmitry Vetrov, and Christian A. Naesseth. SDE Matching: Scalable and simulation-free training of latent stochastic differential equations. In *International Conference on Machine Learning*, volume 267, pp. 3054–3070, 2025.

Souhaib Ben Taieb and Amir F. Atiya. A bias and variance analysis for multistep-ahead time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 27(1):62–76, 2016.

Souhaib Ben Taieb and Rob J. Hyndman. Recursive and direct multi-step forecasting: the best of both worlds. Working Paper 19/12, Monash University, 2012.

Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural ODEs. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Arthur Bizzi, Lucas Nissenbaum, and João M. Pereira. Neural Conjugate Flows: a physics-informed architecture with flow structure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15576–15586, 2025.

Nicholas Boffi, Michael Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: a mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025b.

Junfeng Chen. Diagnosing the conditional-mean barrier in scientific machine-learning surrogates. *Journal of Machine Learning for Modeling and Computing*, 7(3):49–74, 2026. doi: 10.1615/JMachLearnModelComput.2026065274.

Junfeng Chen and Kailiang Wu. Deep-OSG: Deep learning of operators in semigroup. *Journal of Computational Physics*, 493:112498, 2023.

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. doi: 10.1073/pnas.97.7.2968.

Nelida Črnjarić-Žic, Senka Maćešić, and Igor Mezić. Koopman operator spectrum for random dynamical systems. *Journal of Nonlinear Science*, 30:2007–2056, 2020.

Ruizhi Deng, Bo Chang, Marcus A. Brubaker, Greg Mori, and Andreas M. Lehrmann. Modeling continuous stochastic processes with dynamic normalizing flows. In *Advances in Neural Information Processing Systems*, 2020.

Ruizhi Deng, Marcus A. Brubaker, Greg Mori, and Andreas M. Lehrmann. Continuous latent process flows. In *Advances in Neural Information Processing Systems*, 2021.

Naoufal El Bekri, Lucas Drumetz, and Franck Vermet. Time-changed normalizing flows for accurate SDE modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6395–6399, 2024.

Jesse Farebrother, Matteo Pirotta, Andrea Tirinzoni, Rémi Munos, Alessandro Lazaric, and Ahmed Touati. Temporal difference flows. In *International Conference on Machine Learning*, volume 267, pp. 16102–16143, 2025.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025.

Riku Green, Huw Day, Zahraa S. Abdallah, and Telmo M. Silva Filho. Epistemic error decomposition for multi-step time series forecasting: Rethinking bias–variance in recursive and direct strategies. *arXiv preprint arXiv:2511.11461*, 2025.

Riku Green, Zahraa S. Abdallah, and Telmo M. Silva Filho. Expectations vs. realities: The cost of MSE-optimal forecasting under conditional uncertainty. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026.

Jiahe Huang, Sihan Xu, Sharvaree Vadgama, and Rose Yu. Recursive flow matching. *arXiv preprint arXiv:2605.26535*, 2026.

Naoki Kiyohara, Edward Johns, and Yingzhen Li. Neural stochastic flows: Solver-free modelling and inference for SDE solutions. In *Advances in Neural Information Processing Systems*, 2025.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

C. E. Leith. Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102(6):409–418, 1974.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 3870–3882, 2020.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Andy Zeyi Liu, Haoran Sun, Lucas Baker, Randall Balestriero, and John Sous. Semigroup-JEPA: Latent dynamics consistency for zero-shot physics generalization. *arXiv preprint arXiv:2609.10464*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Juan Maroñas, Oliver Hamelijnck, Jeremias Knoblauch, and Theodoros Damoulas. Transforming Gaussian processes with normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 1081–1089, 2021.

Sam McCallum, Zander W. Blasingame, Timothy Herschell, Niklas Rindtorff, Alexander Tong, and James Foster. Strong stochastic flow maps. *arXiv preprint arXiv:2606.01086*, 2026.

Fares Mehouachi and Saif Eddin Jabari. FlowMixer: a depth-agnostic neural architecture for interpretable spatiotemporal forecasting. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jan Kloppenborg Møller and Henrik Madsen. From state dependent diffusion to constant diffusion in stochastic differential equations by the Lamperti transform. IMM-Technical Report 2010-16, Technical University of Denmark, 2010.

Zhengkai Pan, Peter Potaptchik, Wenxi Yao, Michael S. Albergo, and Jakiw Pidstrigach. Itô maps for any-step SDEs. *arXiv preprint arXiv:2606.11156*, 2026.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019. doi: 10.1017/9781108186735.

Lennon J. Shikhman. Semigroup consistency as a diagnostic for learned physics simulators. *arXiv preprint arXiv:2605.26324*, 2026.

Edward Snelson, Carl Edward Rasmussen, and Zoubin Ghahramani. Warped Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 16, 2003.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, volume 202, pp. 32211–32252, 2023.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Çağatay Yıldız, Markus Heinonen, and Harri Lähdesmäki. ODE2VAE: Deep generative second order ODEs with Bayesian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A NOTATION AND THE SCOPE OF COMPOSITION

Appendix guide. Sections B–E give the proofs. Sections F–G cover analytic and scalar experiments; Section H explains the opening Lorenz illustration. Sections I–K give the oracle, motion-capture, and Lorenz protocols and results, including the fixed-map interventions. Section L closes with an exploratory comparison across three nonlinear systems.

Symbol	Meaning
$F_{s,t}$	Learned deterministic point map
$m_{s,t}$	A version of $\mathbb{E}[X_t \mid X_s = \cdot]$
μ_s	Distribution of the actual state X_s
Δ, d_F	True-mean and learned-map composition defects in $L^2(\mu_s)$
$\mathcal{E}_{i,j}$	Squared-loss excess risk over the conditional mean
P_h	Markov transition operator or its corresponding kernel
ϕ, ψ	Invertible chart and inverse chart
A, c, Q	Latent affine drift matrix, offset, and diffusion covariance
R_h, r_h, Σ_h	Finite-lag latent multiplier, offset, and covariance
ϵ	Covariance intensity in $Q = \epsilon Q_0$

The Markov law composes transition *operators*: $(P_h P_k)f(x) = P_h(P_k f)(x)$. In kernel notation this is $P_{h+k}(x, B) = \int P_k(y, B)P_h(x, dy)$. Point-map composition instead applies a nonlinear function to a single intermediate state. The two operations agree for deterministic Dirac kernels, but cannot be interchanged merely by taking the first moment. We assume sufficient moment integrability wherever expectations appear.

B PREDICTION-RISK PROOFS AND A MATCHING NOISE ORDER

B.1 POPULATION BOUND

For any square-integrable predictor $F_{i,j}(X_i)$, write $F_{i,j}(X_i) - X_j = (F_{i,j} - m_{i,j})(X_i) + (m_{i,j}(X_i) - X_j)$. The cross term has zero expectation after conditioning on X_i, proving (4). Neither this decomposition nor the algebraic inequality below requires the joint process to be Markov. The Markov assumption is needed for the interpretation (2) and the kernel construction, not for the general risk bound.

For brevity set $m_1 = m_{s,u}$, $m_2 = m_{u,t}$, $m_{12} = m_{s,t}$, and similarly for F. The exact identity used in Theorem 1 is

$$\begin{aligned} m_2(m_1(x)) - m_{12}(x) = {} & [m_2(m_1(x)) - F_2(m_1(x))] \\ & + [F_2(m_1(x)) - F_2(F_1(x))] \\ & + [F_2(F_1(x)) - F_{12}(x)] + [F_{12}(x) - m_{12}(x)]. \end{aligned}$$

Coverage bounds the squared $L^2(\mu_s)$ norm of the first term by

$$\int \|m_2(y) - F_2(y)\|^2 \, d[(m_1)_{\#}\mu_s](y) \le C \int \|m_2(y) - F_2(y)\|^2 \, d\mu_u(y).$$

The second term has norm at most $L\sqrt{\mathcal{E}_{s,u}}$. Coverage also places $F_2(m_1(X_s))$ and $m_2(m_1(X_s))$ in L^2; the two-point Lipschitz bound then places $F_2(F_1(X_s))$ in L^2. Thus both defects are finite under the stated predictor integrability. The last two norms are d_F and $\sqrt{\mathcal{E}_{s,t}}$. This proves the claimed bound and also clarifies why the Lipschitz bound concerns the learned continuation F_2. Squaring the nonnegative part of $\Delta - d_F$ and applying Cauchy–Schwarz to coefficients $(L, \sqrt{C}, 1)$ yields (7). More generally, for positive weights w_1, w_2, w_{12},

$$w_1\mathcal{E}_{s,u} + w_2\mathcal{E}_{u,t} + w_{12}\mathcal{E}_{s,t} \ge \frac{(\Delta - d_F)_+^2}{L^2/w_1 + C/w_2 + 1/w_{12}}.$$

When recursive risk tracks the true mean defect. Let $r_{\mathrm{rec}} = \|F_2 \circ F_1 - m_{12}\|_{L^2(\mu_s)}$. Inserting $F_2(m_1(x))$ and applying the same two-point Lipschitz condition gives

$$\|F_2 \circ F_1 - m_2 \circ m_1\|_{L^2(\mu_s)} \le L\sqrt{\mathcal{E}_{s,u}} + \sqrt{\widetilde{\mathcal{E}}_{u,t}}.$$

The reverse triangle inequality therefore implies

$$|r_{\mathrm{rec}} - \Delta| \le L\sqrt{\mathcal{E}_{s,u}} + \sqrt{\widetilde{\mathcal{E}}_{u,t}}. \tag{22}$$

If the right side is B, then $|r_{\mathrm{rec}}^2 - \Delta^2| \le B(2\Delta + B)$. This elementary consequence of the same decomposition clarifies empirical attribution: small one-step error on observed states alone does not establish that recursive excess risk is close to Δ^2. It also requires control of the continuation at intermediate means and of error amplification. Coverage gives the sufficient replacement $\widetilde{\mathcal{E}}_{u,t} \le C\mathcal{E}_{u,t}$.

B.2 Uniform bound

Proposition 7 (Uniform alternative)**.** *Suppose $m_{u,t}$ is L_m-Lipschitz and uniform errors $U_{s,u}, U_{u,t}, U_{s,t}$ control $F - m$ on domains containing all required arguments. Then, pointwise on the initial domain,*

$$\|m_{u,t}(m_{s,u}(x)) - m_{s,t}(x)\| \le \|F_{u,t}(F_{s,u}(x)) - F_{s,t}(x)\| + L_m U_{s,u} + U_{u,t} + U_{s,t}. \tag{23}$$

For Proposition 7, insert $m_2(F_1(x))$, $F_2(F_1(x))$, and $F_{12}(x)$. The resulting terms are bounded by $L_m U_{s,u}$, $U_{u,t}$, the pointwise composition defect, and $U_{s,t}$. The supremum defining $U_{u,t}$ must cover $F_1(x)$; controlling only the original input domain would be insufficient. For exact composition and a common uniform error U, any nonzero target defect δ implies $U \ge \delta/(L_m + 2)$.

B.3 An asymptotically sharp smoothness frontier

For the chain in (8), let $a = f(-1)$, $b = f(0)$, $c = f(1)$ and $g = f \circ f$. The one-step losses at times zero and one have combined weights $(3/4, 1/2, 3/4)$, so

$$\mathcal{R}(f) = \tfrac{3}{4}a^2 + \tfrac{1}{2}(b-1)^2 + \tfrac{3}{4}(c+1)^2 + \tfrac{1}{3}\{(g(-1) - \tfrac{1}{4})^2 + (g(0)+1)^2 + g(1)^2\}.$$

For the lower bound, put $v = b - 1$ and $w = g(-1) - 1/4$. The Lipschitz constraint $|f(0) - f(a)| \le L|a|$ gives $3/4 \le L|a| + |v| + |w|$. Weighted Cauchy–Schwarz then yields

$$\mathcal{R}(f) \ge \tfrac{3}{4}a^2 + \tfrac{1}{2}v^2 + \tfrac{1}{3}w^2 \ge \frac{(3/4)^2}{(4/3)L^2 + 2 + 3} = \frac{27}{64L^2 + 240}.$$

For the upper bound choose $a = 3/(4L)$ and $c = -1 + a/L$. Define f by linear interpolation through the ordered knots

x	-1	c	0	a	1
$f(x)$	a	0	1	$1/4$	c

and extend it constantly beyond the outer knots. For $L \ge 2$ all knots are distinct and ordered. The slopes are $-L$, $1/(1 - a/L)$, $-L$, and $(c - 1/4)/(1 - a)$. The second is at most L, since $L - 3/(4L) \ge 1$. For the last, $|(c-1/4)/(1-a)| \le L$ is equivalent to $L + 3/(4L^2) \ge 2$, which also holds. Hence $\mathrm{Lip}(f) \le L$ on $\mathbb{R}$. The composed values are $g(-1) = 1/4$, $g(0) = c$, and $g(1) = 0$, giving

$$\mathcal{R}(f) = \tfrac{3}{4}a^2 + \left(\tfrac{3}{4} + \tfrac{1}{3}\right)(a/L)^2 = \frac{27}{64L^2} + \frac{39}{64L^4}.$$

Multiplication of both bounds by L^2 and squeezing proves the limit. Letting $L \to \infty$ proves the zero infimum without a uniform regularity constraint. It cannot be attained: zero risk would require $f(-1) = 0$, $f(0) = 1$, and $g(-1) = 1/4$, whereas composition would give $g(-1) = f(0) = 1$. This result concerns repeated application of one map on this chain; it is not a characterization of the finite-L optimum or of every time-inhomogeneous compositional family.

B.4 The noise order can be attained

The risk lower bound has the correct order in covariance intensity on a concrete nonlinear diffusion. This statement is about an approximation class, not the finite-sample accuracy of an estimator.

We prove Proposition 5.

Proof. Write $r_t = e^{at} \in (0,1)$ and $v_t = (e^{2at}-1)/(2a) > 0$. The conditional means are $m_t^\epsilon(x) = e^{\epsilon v_t/2}\sinh(r_t \operatorname{arsinh} x)$. The covariance identity $v_{h+k} = r_k^2 v_h + v_k$ gives the exact defect

$$e^{\epsilon v_k/2}\sinh\Big(r_k \operatorname{arsinh}(e^{\epsilon v_h/2}\sinh z)\Big) - e^{\epsilon(v_k + r_k^2 v_h)/2}\sinh(r_k z), \quad z = r_h \operatorname{arsinh} x.$$

Differentiating at zero yields the coefficient

$$D(x) = \tfrac{1}{2} v_h r_k \cosh(r_k z)\left[\tanh z - r_k \tanh(r_k z)\right], \qquad z = r_h \operatorname{arsinh} x.$$

For $x > 0$ the bracket is strictly positive; for $x < 0$ it is strictly negative. Thus $\|D\|_{L^2(\mu_0)} > 0$. To justify the norm expansion, put $T_d(z) = \operatorname{arsinh}(e^d \sinh z)$. Then $|T_d(z) - z| \le d$, $\partial_d T_d = \tanh T_d$, and $\partial_d^2 T_d = \tanh T_d \operatorname{sech}^2 T_d$ for $d \ge 0$. Thus the second ϵ derivative of the displayed defect is bounded, uniformly for $0 \le \epsilon \le \epsilon_0$, by $Ce^{r_k|z|}$. Since $z = r_h Z_0$ is Gaussian, this envelope lies in L^2. Taylor's theorem therefore gives a remainder $O(\epsilon^2)$ in $L^2(\mu_0)$, and $\Delta_\epsilon = \epsilon\|D\|_{L^2(\mu_0)} + O(\epsilon^2)$.

We verify coverage rather than assuming it. In the latent coordinate, the law of the conditional mean is the law of $T(U) = \operatorname{arsinh}(e^{\epsilon v_h/2}\sinh U)$ with $U \sim \mathcal{N}(0, S)$, $S = r_h^2\tau^2 > 0$. The actual state has latent law $\mathcal{N}(0, S + \epsilon v_h)$. The odd increasing map T satisfies $T' \ge 1$ and $|T(u) - u| \le d = \epsilon v_h/2$. Comparing Gaussian densities therefore gives

$$\frac{d[(m_h^\epsilon)_\# \mu_0]}{d\mu_h^\epsilon} \le C_\epsilon := (1 + \epsilon v_h/S)^{1/2}\exp(\epsilon v_h/8). \tag{24}$$

For completeness, with $V = S + \epsilon v_h$ the exponent in the density ratio is bounded by $z^2/(2V) - (|z| - d)_+^2/(2S)$, whose supremum is $d^2/(2(V-S)) = \epsilon v_h/8$. Thus C_ϵ is uniformly bounded near zero. Theorem 1 supplies the $\Omega(\epsilon^2)$ lower bound for every $F \in \mathcal{A}_L$.

For the upper bound take $g_t(x) = \sinh(r_t \operatorname{arsinh} x)$. These maps compose exactly and $g_t'(x) = r_t\cosh(r_t \operatorname{arsinh} x)/\sqrt{1+x^2} \le r_t \le 1$. Moreover $m_t^\epsilon - g_t = (e^{\epsilon v_t/2} - 1)g_t$. Gaussian exponential moments bound its squared norm by $O(\epsilon^2)$ under both μ_0 and μ_h^ϵ. The sum of the three risks is therefore $O(\epsilon^2)$. □

C Mean composition and diffusion generators

C.1 Generator criterion

Here are the analytic assumptions used by Proposition 2. The diffusion is nonexplosive with finite first moments. Its conditional means $m_h(x)$ are continuous at zero, continuously differentiable in h, twice differentiable in x, satisfy $\partial_h m_h = \mathcal{L} m_h$ with $m_0 = \mathrm{id}$, and have $\partial_h m_h|_{h=0} = b$. The ODE $\dot{x} = b(x)$ has unique forward solutions throughout the time range considered, with a flow Φ_h that is continuously differentiable in time and C^2 in space. The classical backward equation with initial function id is unique in a function class containing m_h and Φ_h. These conditions can be checked by standard regularity, growth, and moment assumptions in specific diffusion families; they are not inferred from the existence of an arbitrary SDE alone.

If $m_{h+k} = m_k \circ m_h$, differentiating in k at zero gives $\partial_h m_h(x) = b(m_h(x))$. ODE uniqueness implies $m_h = \Phi_h$. The flow also satisfies $\partial_h \Phi_h = D\Phi_h\, b$, by differentiating $\Phi_{h+k} = \Phi_h \circ \Phi_k$ in k at zero. Subtracting this identity from $\partial_h m_h = \mathcal{L} m_h$ yields $a : \nabla^2 \Phi_h = 0$. Conversely, if this contraction vanishes, then Φ_h solves the backward equation with initial data id. Its assumed uniqueness yields $m_h = \Phi_h$, which composes. For affine b, the drift flow is affine in the initial state, so its Hessian vanishes. This proof does not require a to be constant.

C.2 The local defect

A sufficient expansion hypothesis for Proposition 3 is

$$m_h(x) = x + hb(x) + \tfrac{1}{2}h^2 \mathcal{L} b(x) + O(h^3)$$

locally uniformly, with $b \in C^2$ and $\mathcal{L} b \in C^1$ on a neighborhood of the compact set in question and its small-time mean images. It follows, for example, from an iterated Dynkin expansion when the

required generator iterates and their moments are controlled. With this hypothesis, Taylor expansion gives

$$m_h(m_h(x)) = x + 2hb(x) + h^2\{\mathcal{L}b(x) + Db(x)b(x)\} + O(h^3),$$
$$m_{2h}(x) = x + 2hb(x) + 2h^2\mathcal{L}b(x) + O(h^3).$$

Subtracting leaves $h^2(Db\,b - \mathcal{L}b) = -\frac{1}{2}h^2 a : \nabla^2 b$. The corresponding two-length expression has leading term $-(hk/2)a : \nabla^2 b$ with a remainder $O((h+k)^3)$ as both lags vanish. In particular the direct full-step mean minus the composition of two half-step means has leading term $h^2 a : \nabla^2 b/8$. For $b(x) = x - x^3$ and constant scalar noise σ, the equal-step composed-minus-direct coefficient is $3\sigma^2 x$.

C.3 The conditional expectation gap

Under the Markov property, the tower property gives $m_{h+k}(x) = \mathbb{E}[m_k(X_h) \mid X_0 = x]$. Expand m_k about $\bar{x} = m_h(x)$ with an integral second-order remainder. The expected linear term vanishes because $\mathbb{E}[X_h - \bar{x} \mid x] = 0$. If the operator norm of the vector Hessian is bounded by M on every segment from $\bar{x}$ to a possible X_h, the remainder has norm at most $(M/2)\mathbb{E}[\|X_h - \bar{x}\|^2 \mid x]$. This proves the bound following (11). No sign or nonvanishing conclusion follows without further conditions.

D Deterministic representations and the Gaussian extension

D.1 Cancellation and the anchor representation

For the two-time representation in (16),

$$F_{u,t} \circ F_{s,u} = \phi_t^{-1}\Lambda(t)\Lambda(u)^{-1}\underbrace{\phi_u\phi_u^{-1}}_{\mathrm{id}}\Lambda(u)\Lambda(s)^{-1}\phi_s = F_{s,t}.$$

The identity uses no commutativity assumption. Also $F_{s,s} = \mathrm{id}$ and $F_{s,t}^{-1} = F_{t,s}$ when all time orders are included. Conversely, for any invertible evolution family on a common state space over a finite interval with terminal time r, set $\phi_t = F_{t,r}$. The forward composition identity $F_{s,r} = F_{t,r} \circ F_{s,t}$ gives $\phi_t^{-1} \circ \phi_s = F_{s,t}$ using only forward maps and their inverses. The matrix ratio can be absorbed into a chart $\widetilde{\phi}_t = \Lambda(t)^{-1}\phi_t$; without architectural restrictions it introduces no additional representation power.

For globally Lipschitz time-dependent vector fields with integrable Lipschitz and linear-growth bounds on a finite time interval, the ODE flow is invertible and the anchor construction applies. The resulting chart can itself be as hard to compute as the original flow. This representation does not provide finite-network approximation rates, cheap evaluation, or universal fixed-chart linearization. A fixed-chart spectral construction instead uses $G_h = \phi^{-1}\mathcal{F}^{-1}\,\mathrm{diag}_k(e^{hA(k)})\mathcal{F}\phi$ on a finite grid, with the required real-field symmetry; it is a concrete restricted realization.

D.2 Proof of the covariance characterization

Composing two independent affine Gaussian transitions gives mean $R_k R_h z + R_k r_h + r_k$ and covariance $R_k\Sigma_h R_k^\top + \Sigma_k$. The affine means satisfy the semigroup law by their definitions. Equality of Gaussian laws, including degenerate ones, is therefore equivalent to the covariance identity.

Continuity suffices to justify differentiability. Fix a sufficiently small $\delta > 0$ and define the linear map $T_\delta(B) = \int_0^\delta R_v B R_v^\top\,dv$. It is invertible on symmetric matrices because $T_\delta/\delta \to I$ in operator norm. Integrating the covariance cocycle gives

$$T_\delta(\Sigma_h) = \int_h^{h+\delta} \Sigma_v\,dv - \int_0^\delta \Sigma_v\,dv.$$

The right side is continuously differentiable in h (with a right derivative at zero), so Σ_h is also, and $Q = \Sigma_0'$ exists.

To derive the integral schedule, interchange the two lag symbols in this identity, obtaining $\Sigma_{h+\delta} - \Sigma_h = R_h \Sigma_\delta R_h^\top$. Dividing by $\delta \downarrow 0$ gives the right derivative $R_h Q R_h^\top$. Continuity of this expression yields the integral formula; equivalently one can sum the identity over a uniform partition and take the associated Riemann-sum limit using $\Sigma_\delta = \delta Q + o(\delta)$. The derivative Q is positive semidefinite because Σ_δ/δ is. Conversely, splitting the covariance integral at k proves the cocycle identity directly.

For a bimeasurable bijection ϕ and any Markov semigroup K_h, define $P_h(x, B) = K_h(\phi(x), \phi(B))$. Change variables $z = \phi(y)$ in the kernel integral:

$$\int P_k(y, B) P_h(x, dy) = \int K_k(z, \phi(B)) K_h(\phi(x), dz) = K_{h+k}(\phi(x), \phi(B)).$$

The implication is reversible by conjugating with ϕ^{-1}. If ϕ is a diffeomorphism and $\Sigma_h \succ 0$, the ordinary change-of-variables formula gives (19). At zero noise this must instead be read as an equality of Dirac kernels. All stochastic time lags here are nonnegative.

For nonautonomous linear latent dynamics $R_{s,t}$, a time-dependent chart can be used with $\Sigma_{s,t} = \int_s^t R_{v,t} Q_v R_{v,t}^\top \, dv$ and the corresponding affine forcing. The identity becomes $\Sigma_{s,t} = R_{u,t} \Sigma_{s,u} R_{u,t}^\top + \Sigma_{u,t}$. This extension follows the same argument, but is not an expressivity theorem for general state-dependent diffusions.

D.3 The small-noise expansion

We make the tail requirement explicit. It suffices that ψ is a C^4 diffeomorphism, that its derivatives through order four have at most exponential-linear growth, $\left\| D^j \psi(z) \right\| \le C \exp(c \|z\|)$, and that $\phi = \psi^{-1}$ is C^2 locally. For initial states in a compact set and fixed finite lags, these assumptions control all Taylor remainders under the Gaussian innovations. Other growth conditions yielding the same uniform integrability also suffice. Merely bounding derivatives near the noise-free orbit is not enough.

For $\xi \sim \mathcal{N}(0, \epsilon S)$, Taylor expansion through cubic order gives

$$\mathbb{E}\psi(z + \xi) = \psi(z) + \tfrac{\epsilon}{2} H\psi(z)[S] + O(\epsilon^2). \tag{25}$$

Odd Gaussian moments vanish. The fourth-order remainder is controlled by $C\mathbb{E}[\|\xi\|^4 \exp(c\|\xi\|)] = O(\epsilon^2)$, uniformly for bounded z. At lag h, this yields $m_h(x) = \psi(z_1) + \frac{\epsilon}{2} H\psi(z_1)[S_h] + O(\epsilon^2)$, hence

$$\phi(m_h(x)) = z_1 + \tfrac{\epsilon}{2} [D\psi(z_1)]^{-1} H\psi(z_1)[S_h] + O(\epsilon^2).$$

Propagating for lag k and applying (25) again gives

$$m_k(m_h(x)) = \psi(\bar{z}) + \tfrac{\epsilon}{2} D\psi(\bar{z}) R_k [D\psi(z_1)]^{-1} H\psi(z_1)[S_h] + \tfrac{\epsilon}{2} H\psi(\bar{z})[S_k] + O(\epsilon^2).$$

The direct mean replaces the two order-ϵ terms by $\frac{\epsilon}{2} H\psi(\bar{z})[R_k S_h R_k^\top + S_k]$. Subtracting proves (20). A nonzero coefficient establishes a defect for all sufficiently small positive ϵ; a zero coefficient establishes neither exact composition nor its failure at higher orders. Globally affine ψ gives exact mean composition, while the sinh/Brownian example shows that it is not a necessary condition.

E Representational boundaries

E.1 Scalar criterion

Under a time-independent increasing chart $Z = \phi(X)$, Itô's formula gives

$$dZ = \left(\phi' b + \tfrac{1}{2}\phi'' \sigma^2\right)(X)\, dt + \phi'(X)\sigma(X)\, dW.$$

For its diffusion coefficient to be a constant $\rho > 0$, necessarily $\phi' = \rho/\sigma$, hence $\phi = \rho\ell + \beta$ and $\phi'' = -\rho\sigma'/\sigma^2$. The drift becomes $\rho(b/\sigma - \sigma'/2)$. Equating it to $\alpha\phi + \gamma$ and differentiating gives (21). Conversely, if that equation holds, integration shows that $b/\sigma - \sigma'/2 = \alpha\ell + d$ for a constant d. Choosing any $\rho > 0, \beta$ and setting $\gamma = \rho d - \alpha\beta$ yields the required transformed diffusion. The range assumption on ℓ supplies a global chart onto $\mathbb{R}$. Without it the same calculation is local

or applies only on the transformed interval; global boundary behavior requires separate treatment. Decreasing charts give the same criterion after a sign convention.

For constant σ, (21) reduces to $b' = \alpha$. Thus non-affine drift with constant additive noise cannot have an exact fixed-chart affine-diffusion representation. An exact match of all transition kernels would also imply equality of their generators and therefore cannot evade this restriction. This conclusion says nothing about a time-dependent chart, extra latent state, a different latent generator, or approximation at finitely many lags.

E.2 An observed process with memory

For $Y_t = U + tV$, the covariance is $\mathrm{Cov}(Y_s, Y_t) = 1 + st$ and the variance at time s is $1 + s^2$. The Gaussian conditioning formula gives

$$\mathrm{Law}(Y_t \mid Y_s = x) = \mathcal{N}\left(\frac{1+st}{1+s^2}x, \frac{(t-s)^2}{1+s^2}\right), \tag{26}$$

since

$$(1+t^2) - \frac{(1+st)^2}{1+s^2} = \frac{(t-s)^2}{1+s^2}.$$

From 0 to 1 the kernel is $\mathcal{N}(x, 1)$; from 1 to 2 it is $\mathcal{N}(3y/2, 1/2)$. Composing gives variance $9/4 + 1/2 = 11/4$ and mean $3x/2$, distinct from the direct $\mathcal{N}(x, 4)$ law. This verifies a failure of observed-state CK explicitly. The sufficient two-dimensional state evolves by the deterministic linear map $\left(\begin{smallmatrix} 1 & h \\ 0 & 1 \end{smallmatrix}\right)$. The example demonstrates the need for retained information; it does not establish a general procedure or lower bound for minimal state dimension.

F Analytic and scalar numerical experiments

The finite-state, analytic, and scalar likelihood experiments use 64-bit arithmetic. Reported standard deviations describe variability across generated training datasets; quadrature error is assessed separately. Learned warped-OU and double-well models are specified in Section G.

F.1 Finite-state risk calculation and optimization

With the state ordering $(-1, 0, 1)$, the distributions are $\mu_0 = (1/3, 1/3, 1/3)$ and $\mu_1 = (5/12, 1/6, 5/12)$. The conditional-mean image of μ_0 is again uniform, so $C = \max_i \mu_0(i)/\mu_1(i) = 2$. The composition defect is $(3/4, 0, 0)$, with norm $\sqrt{3}/4$. We set $f(x) = c_0 + c_1 x + c_2 x^2$ and fit a quadratic g through its values at the three states. Given f, minimizing the penalized objective over these three values is explicit:

$$g(x_i) = \frac{m_2(x_i) + \lambda f(f(x_i))}{1+\lambda}.$$

The total excess risk is

$$\mathcal{R} = \sum_i [\mu_0(i) + \mu_1(i)]\{f(x_i) - m_1(x_i)\}^2 + \sum_i \mu_0(i)\{g(x_i) - m_2(x_i)\}^2.$$

The consistency penalty is $\lambda \sum_i \mu_0(i)\{f(f(x_i)) - g(x_i)\}^2$. For finite λ, the quadratic interpolant defines g away from the three states. For the exact case we instead set $g = f \circ f$ on all of $\mathbb{R}$; $F_n = f^n$ then gives an exact discrete semigroup. The risk theorem does not require invertibility or continuous-time embeddability.

We minimize over (c_0, c_1, c_2) using a NumPy implementation of Nelder–Mead with initial simplex displacement 0.15, maximum 1,800 iterations, and maximum simplex displacement stopping tolerance 10^{-10}. Initial points are $(1, -1/2, -3/2)$, $(0, 0, 0)$, $(0, 1/2, 0)$, and the previous penalty solution where available; we select the lowest attained objective. The continuation's Lipschitz bound is the maximum of $|c_1 + 2c_2 x|$ at the endpoints of the interval containing all $m_1(x_i)$ and $f(x_i)$. This interval contains every line segment used in the proof. The numerical results are as follows.

Table 3: Complete finite-state study. $\mathcal{R}$ is the sum of three excess risks, d_F is the composition defect, and the last column is the risk lower bound in (7). Risks are evaluated by exact sums over the finite state space. The lower bound is not tight in its numerical constant on this fitted family.

λ	$\mathcal{R}$	d_F	Lower bound
0	0.00000	0.43301	0.00000
0.01	0.00003	0.42574	0.00000
0.1	0.00198	0.37804	0.00020
1	0.04686	0.17633	0.00453
10	0.11118	0.02409	0.01146
100	0.12175	0.00248	0.01269
1000	0.12285	0.00025	0.01282
∞	0.12298	0.00000	0.01284

F.2 Analytic noise study

The parameters are $a = -0.7$, $x = 0.9$, $h = 0.4$, and $k = 0.7$. We use 15 intensities $\epsilon_j = 10^{-4+3.5j/14}$, $j = 0, \dots, 14$. Exact means follow from the Gaussian moment-generating function. The plotted leading term uses (20), with $\psi = \sinh$ and $\phi = \operatorname{arsinh}$. The slope is the least-squares slope of log defect versus log intensity on the first six points. The relative remainder is $|m_k(m_h(x)) - m_{h+k}(x) - \epsilon\mathcal{D}|/|\epsilon\mathcal{D}|$. These deterministic evaluations have no seed or sampling error.

F.3 Transition-likelihood fitting

The data-generating chart is $\phi_{\beta_*}(x) = \sinh(\beta_* x)/\beta_*$, $\beta_* = 1.1$. For each pair independently draw $Z_0 \sim \mathcal{N}(0,1)$ and $\xi \sim \mathcal{N}(0,1)$, choose its lag from the balanced set $\{0.2, 0.5, 1.0\}$, and set

$$X_0 = \frac{\operatorname{arsinh}(\beta_* Z_0)}{\beta_*}, \quad X_h = \frac{\operatorname{arsinh}\big(\beta_*[e^{a_* h} Z_0 + \sqrt{v_h}\xi]\big)}{\beta_*}, \quad v_h = q_* \frac{e^{2a_* h} - 1}{2a_*},$$

where $a_* = -0.7, q_* = 0.8$. There are 9,000 pairs per seed and seeds are $0, 1, 2$. Pairs are independent rather than overlapping samples from one long trajectory.

The compatible fit uses $a = -e^{\theta_1}, q = e^{\theta_2}, \beta = e^{\theta_3}$ and minimizes the exact average negative log-density. Nelder–Mead uses the settings above and starts $(-0.2, -0.1, 0)$, $(-1, 0.2, -0.5)$, and $(0.2, -0.5, 0.5)$. Parameters with any $|\theta_i| > 4$ receive a large finite penalty. We select the best attained objective.

For the free schedule, at each fixed β we encode the pairs and use the zero-intercept Gaussian regression maximum-likelihood coefficients at each training lag j:

$$\widehat{R}_j = \frac{\sum_i z_i w_i}{\sum_i z_i^2}, \qquad \widehat{\Sigma}_j = \frac{1}{n_j} \sum_i (w_i - \widehat{R}_j z_i)^2.$$

The chart parameter is fitted by profiling this likelihood over $\log\beta$, initialized at zero, with $|\log\beta| \le 3$ enforced by a penalty. Nonpositive multipliers are rejected; all selected fits have positive multipliers. Piecewise linear interpolation is applied to $\log R_j$ and Σ_j, adding the identity knot $(h, \log R, \Sigma) = (0, 0, 0)$. The last segment is continued for the extrapolated lag. All evaluated variances are positive.

Evaluation reports

$$\mathbb{E}_{Z_0 \sim \mathcal{N}(0,1)} \operatorname{KL}(p_*(\cdot \mid \psi_*(Z_0), h) \,\|\, p_\theta(\cdot \mid \psi_*(Z_0), h)) .$$

It is computed using independent quadrature, not training examples: 48-point Gauss–Hermite rules on each of Z_0 and the latent innovation, with $\sqrt{2}$ node scaling and $1/\sqrt{\pi}$ weight normalization. The script evaluates the density ratio in observation coordinates, including each model's Jacobian. A separate 96-point-per-axis evaluation checks numerical convergence; the maximum absolute KL change is 2.2×10^{-9}. Expected KL is an average accuracy criterion, not a pointwise guarantee at every initial state.

For structural consistency, use the analytic one-dimensional Gaussian formula $W_2^2 = (\mu_1 - \mu_2)^2 + (\sqrt{v_1} - \sqrt{v_2})^2$. At each split (h, k), compare direct mean/variance (R_{h+k}, Σ_{h+k}) at latent $z = 1$ with composed $(R_k R_h, R_k^2 \Sigma_h + \Sigma_k)$. This removes Monte Carlo uncertainty from the CK diagnostic. Because the diagnostic is in each learned chart, its magnitude is not an observation-space cross-model accuracy comparison.

F.4 A LOW-DIMENSIONAL LIKELIHOOD CONTROL

The likelihood protocol above fits a realizable scalar process with $A = -0.7$, $Q = 0.8$, and $\phi_\beta(x) = \sinh(\beta x)/\beta$, $\beta = 1.1$. The compatible model has three parameters; the free per-lag schedule has seven, including the shared chart parameter. Both use 9,000 independent training pairs per seed. This control separates accuracy from internal consistency.

Table 4: Expected conditional KL to the known transition law, in nats. Mean $\pm$ sample standard deviation over three independently generated training sets. Evaluation uses 48×48 Gauss–Hermite quadrature over the initial latent state and innovation. "Train" denotes an observed lag, not reuse of training samples; extra.: extrapolated lag. Both models learn the same one-parameter chart family.

Lag	Coverage	Compatible schedule	Free per-lag schedule
0.20	train	0.00019 ± 0.00013	0.00053 ± 0.00046
0.35	held out	0.00022 ± 0.00016	0.00064 ± 0.00025
0.50	train	0.00025 ± 0.00018	0.00027 ± 0.00016
0.75	held out	0.00030 ± 0.00022	0.00101 ± 0.00018
1.00	train	0.00033 ± 0.00024	0.00050 ± 0.00019
1.50	extra.	0.00037 ± 0.00027	0.00251 ± 0.00075

The compatible model's mean KL stays below 4×10^{-4}, whereas the free schedule reaches 2.51×10^{-3} at lag 1.5. Compatible latent Gaussians agree under composition to 1.11×10^{-16} at $z = 1$ (Figure 5). The gains reflect a correctly specified OU constraint and parameter sharing; they do not isolate parameter-count and interpolation effects or establish superiority on arbitrary diffusions.

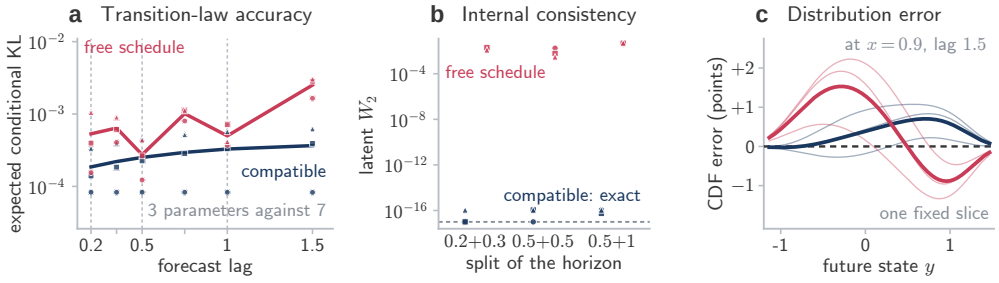

Figure 5: Accuracy, internal consistency, and distribution error. A, Expected conditional KL at all six lags, with individual data seeds and their means; grey guides mark training lags. **B,** Direct-versus-composed latent Gaussian W_2 at $z = 1$ for three time splits, within each model's own coordinates. Raw values below 10^{-17} are displayed at that numerical floor; the compatible schedule obeys CK algebraically. **C,** Signed conditional-CDF error in percentage points at physical input $x = 0.9$ and lag 1.5. Thin curves show three seeds; thick curves their mean, over the true conditional 0.5%–99.5% quantile range. C is one illustrative slice. The models differ in parameter count and interpolation constraints.

F.5 CONSISTENCY CAN INCREASE EXCESS PREDICTION RISK

Use the three-state chain in (8), with uniform initial law. Its one-step means are $(0, 1, -1)$ and its two-step means are $(1/4, -1, 0)$ at the three states. The true mean-composition defect is $\Delta = \sqrt{3}/4$, and the coverage constant in Theorem 1 is $C = 2$. Fit a quadratic first-step predictor f and a direct two-step predictor g by minimizing $\mathcal{R} + \lambda \|f \circ f - g\|^2_{L^2(\mu_0)}$. Both unconstrained conditional means are representable in these classes. At $\lambda = \infty$ we define $g = f \circ f$ exactly, giving a discrete semigroup

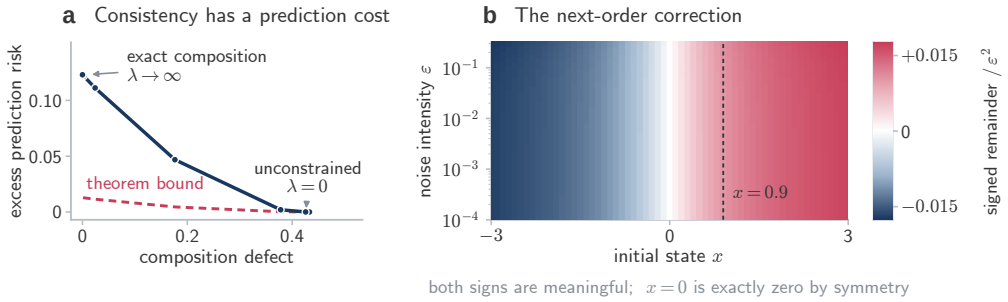

Figure 6: Compatibility mechanisms. A, Exact finite-state risk and composition defect for saved fits as the consistency penalty varies; the dashed curve evaluates the theorem's bound using each fit's regularity. These fits are not certified global optima. **B,** Signed analytic remainder $[\delta_{h,k}(x,\epsilon) - \epsilon\mathcal{D}_{h,k}(x)]/\epsilon^2$ for the sinh-observed OU example, with $(h,k) = (0.4, 0.7)$. The dashed vertical line identifies the original input $x = 0.9$. Colors show the next-order correction, evaluated in high precision.

through iterates of f. Figure 6a shows that reducing composition error increases excess risk in this example. The exact model has excess risk 0.12298, above the theorem's lower bound 0.01284. These are numerically optimized fits within the specified family, not claims of global optimality over all compositional models. Section F gives the optimizer, all penalty values, and exact risk calculations.

F.6 An analytic test of the noise expansion

Let $dZ = -0.7Z\,dt + \sqrt{\epsilon}\,dW$ and $X = \sinh Z$. For $v_h = (e^{-1.4h} - 1)/(-1.4)$, the exact mean is $m_h(x) = e^{\epsilon v_h/2}\sinh(e^{-0.7h}\operatorname{arsinh} x)$. At $x = 0.9$ and lags $(0.4, 0.7)$, (20) predicts coefficient 0.03273256. Across 15 logarithmically spaced noise intensities, the six smallest give log–log slope 1.0002; at $\epsilon = 10^{-4}$ the relative leading-term error is 0.0031% in the saved one-input diagnostic. Figure 6B instead displays the signed second-order remainder across input and noise intensity, making departures from the leading term visible. With latent drift zero, the same chart gives the exactly composing exception in (15). No sampling estimate is used.

G Warped-OU and double-well forecasting

We fit four models on warped OU and two on a double-well diffusion, each with three independently generated training datasets. These learned models complement the scalar likelihood control in Section F.4.

Table 5: Conditional-mean accuracy on warped OU. Mean ± sample standard deviation over three data seeds. The unseen lag 1.5 tests extrapolation; $\mathcal{R}$ sums excess risks for time pairs $0 \to 0.5$, $0.5 \to 1$, and $0 \to 1$ on their actual marginals. These errors exclude irreducible conditional variance.

Model	Params.	Excess MSE, $h = 1.5$	Three-pair $\mathcal{R}$
Direct MLP	4417	0.14751 ± 0.07645	0.09570 ± 0.04012
Composing map	74	0.00811 ± 0.00003	0.07679 ± 0.01731
Compatible kernel	75	0.00105 ± 0.00079	0.08593 ± 0.05213
Mixture	5327	0.14928 ± 0.08311	0.05819 ± 0.02296

G.1 Data and independent evaluation

For warped OU, let $dZ = -Z\,dt + dW$, $Z_0 \sim \mathcal{N}(0,1)$, and $X = \psi(Z) = Z + \frac{3}{5}Z^3$. Writing $r_h = e^{-h}$, $v_h = (1 - e^{-2h})/2$, and $u = r_h\psi^{-1}(x)$ gives

$$m_h(x) = u + \tfrac{3}{5}(u^3 + 3uv_h).$$

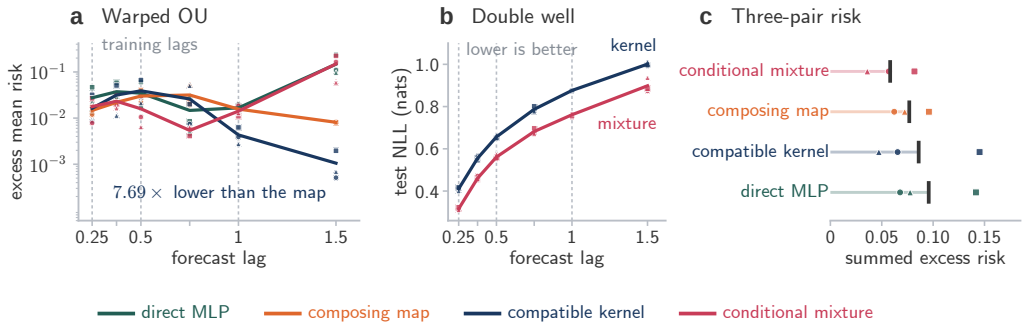

Figure 7: Compatibility and accuracy are distinct properties. A, Warped-OU excess mean-prediction risk over all forecast lags (logarithmic vertical axis). **B,** On the misspecified double-well process, the mixture achieves lower refined-simulation NLL. **C,** Three-pair excess risk for the same warped-OU fits: the kernel's long-horizon improvement does not imply a uniform joint-risk advantage. Dots show all three data seeds, and lines or black ticks show means; C uses circle/square/triangle for seeds 0/1/2. Grey guides in A/B mark training lags. Capacities and training objectives differ across models.

The conditional variance follows from Gaussian moments through order six; the exact transition density follows by the change of variables through the increasing cubic map. For each lag, half the inputs are drawn from the actual time-0 marginal and half from the actual time-0.5 marginal, whose latent variance is $e^{-1} + (1 - e^{-1})/2$. Transitions are independently sampled conditional on these inputs.

Training lags are $0.25, 0.5, 1.0$, with 12,000 pairs per lag (36,000 per seed). Validation uses 2,000 independent pairs per training lag. Test lags are $0.25, 0.375, 0.5, 0.75, 1.0, 1.5$, with 5,000 independent pairs per lag. Seeds $0, 1, 2$ regenerate the data; all models within a seed share the same pairs. The reported excess MSE directly compares a fitted conditional mean with the analytic mean, without estimating or subtracting a noisy Bayes-risk floor. Conditional KL is estimated by the sample mean of $\log p_*(Y \mid X, h) - \log p_\theta(Y \mid X, h)$; finite-sample values are not clipped. Prediction MSE against realized targets, likelihood, PIT values, and central-90% coverage are reported separately.

For the double well, $dX = (X - X^3)dt + 0.8dW$, draw inputs uniformly from $[-1.8, 1.8]$. Training and validation trajectories use the tamed-Euler update

$$X_{j+1} = X_j + \frac{\delta b(X_j)}{1 + \delta|b(X_j)|} + 0.8\sqrt{\delta}\,\xi_j, \qquad \delta = 10^{-3}.$$

Test pairs include a coupled simulation at $\delta/2$, with each coarse Gaussian increment formed from the sum of two fine increments. NLL on this refined simulation is the primary density comparison; no exact continuous-time density is claimed. The largest mean NLL change over all models, seeds, and lags is 4.62×10^{-4}, and the largest paired endpoint RMS change is 3.39×10^{-4}. These checks assess the tested refinement, not a rigorous bound on time-discretization bias.

G.2 Architectures, Objectives, and Optimization

Inputs and outputs are divided by the training-input standard deviation for each seed. The shared chart family, in normalized coordinates, is

$$\phi(x) = \frac{ax + \sum_{j=1}^{24} c_j[\tanh(w_j x + b_j) - \tanh(b_j)]}{a + \sum_{j=1}^{24} c_j w_j[1 - \tanh^2(b_j)]}, \qquad a, c_j, w_j > 0.$$

Thus $\phi(0) = 0$, $\phi'(0) = 1$, and the chart is increasing and onto $\mathbb{R}$. The denominator fixes the scale gauge. Its finite asymptotically affine tails need not match the true inverse-cubic encoder globally. The composing map uses $F_h(x) = \phi^{-1}(e^{-\alpha h}\phi(x))$, $\alpha > 0$ (74 parameters). The compatible kernel adds a positive covariance rate q and latent variance $q(1 - e^{-2\alpha h})/(2\alpha)$ (75 parameters). Inversion uses analytic global brackets and 30 bisections in float32 or 52 in float64, with an implicit-gradient correction for training.

The direct MLP has two width-64 SiLU hidden layers, takes $(x, \log(1+h))$ as input, and outputs a residual $x + hf_\theta(x, h)$ (4,417 parameters). The mixture model uses the same hidden widths and five Gaussian components, with means $x + h\delta_j(x, h)$ and standard deviations $\sqrt{h}[\mathrm{softplus}(s_j) + 0.02]$ (5,327 parameters). It imposes no CK constraint. Point models minimize squared error; density models minimize negative log-likelihood. These are not comparisons at matched capacity or identical objectives.

Adam uses learning rate 10^{-3}, batch size 1,024, gradient-norm clipping at 10, and cosine decay to 5% of the initial learning rate. Point models receive 4,000 updates and density models 6,000. Validation is evaluated every 250 updates; its lowest loss selects the returned checkpoint. Only the observed training lags enter validation. Float32 is used for training with TF32 disabled; structural checks use float64. The compatible warped-OU runs all selected their final update, so the results are not presented as fully converged class optima.

Compatible-kernel means integrate the inverse chart using 48-point Gauss–Hermite quadrature. Comparison with 24 nodes on 512 evenly spaced test indices gives a maximum absolute difference of 1.31×10^{-4} and RMS differences of approximately 2–3 $\times 10^{-5}$ on warped OU. The mixture means are evaluated analytically. The kernel's zero-noise map is never substituted for its conditional mean.

G.3 All horizons and the three-time risk

Table 6: Warped-OU excess MSE to the exact conditional mean, mean $\pm$ sample standard deviation over three data seeds. Lags $0.375, 0.75$ are interpolation tests and 1.5 is extrapolation.

Lag	Direct MLP	Composing map	Compatible kernel	Mixture
0.25	0.0276 ± 0.0172	0.0147 ± 0.0027	0.0163 ± 0.0135	0.0178 ± 0.0095
0.375	0.0373 ± 0.0184	0.0219 ± 0.0008	0.0316 ± 0.0198	0.0232 ± 0.0096
0.5	0.0348 ± 0.0230	0.0300 ± 0.0086	0.0390 ± 0.0243	0.0159 ± 0.0130
0.75	0.0147 ± 0.0143	0.0317 ± 0.0193	0.0260 ± 0.0221	0.0054 ± 0.0012
1	0.0167 ± 0.0028	0.0157 ± 0.0001	0.0043 ± 0.0018	0.0142 ± 0.0039
1.5	0.1475 ± 0.0764	0.0081 ± 0.0000	0.0011 ± 0.0008	0.1493 ± 0.0831

The three-time excess risk in Table 5 uses the actual source-time laws for $0 \to 0.5$, $0.5 \to 1$, and $0 \to 1$, with 2,500 test inputs per pair. It is a sum of three Monte Carlo averages, not the average over all six forecast lags. Its ranking is mixed: the kernel mean is 0.08593 ± 0.05213, versus 0.07679 ± 0.01731 for the composing map. At lag 0.5, a post-hoc diagnostic finds that the largest 1% of absolute input values contribute 93.0–98.4% of the compatible kernel's excess error across seeds. These inputs are included in all reported scores.

Table 7: Warped-OU density accuracy and aggregate central-90% coverage. KL is in nats; coverage is a percentage. Entries are means $\pm$ sample standard deviations over three data seeds.

	Conditional KL estimate		Coverage (%)	
Lag	Compatible	Mixture	Compatible	Mixture
0.25	0.0061 ± 0.0011	0.0087 ± 0.0053	90.09 ± 0.44	90.17 ± 0.78
0.375	0.0066 ± 0.0008	0.0092 ± 0.0011	90.13 ± 0.01	90.32 ± 0.17
0.5	0.0069 ± 0.0009	0.0085 ± 0.0030	90.23 ± 0.29	90.03 ± 0.09
0.75	0.0062 ± 0.0013	0.0081 ± 0.0019	89.79 ± 0.28	89.71 ± 0.35
1	0.0047 ± 0.0005	0.0062 ± 0.0009	89.60 ± 0.30	90.02 ± 0.68
1.5	0.0040 ± 0.0012	0.0443 ± 0.0142	89.86 ± 0.31	91.71 ± 0.79

Near-nominal aggregate coverage does not establish conditional calibration at every input. At lag 1.5, total forecast MSE against noisy targets is approximately 1.9955 for the compatible kernel and 1.9992 for the composing map. The 7.69 ratio concerns *excess* MSE, the quantity relevant to approximation of the Bayes conditional mean.

On the double well (Table 8), the mixture has lower NLL at every tested lag in every seed. This is consistent with the fixed-chart representation limit, but finite training and different model capacities prevent interpreting the measured gap as a certified approximation lower bound.

Table 8: Double-well transition NLL on the coupled refined simulation at step 5×10^{-4}. Entries are means $\pm$ sample standard deviations over three data seeds. Lower is better.

Lag	Compatible kernel	Mixture
0.25	0.4106 ± 0.0087	0.3159 ± 0.0065
0.375	0.5553 ± 0.0087	0.4605 ± 0.0096
0.5	0.6565 ± 0.0083	0.5612 ± 0.0113
0.75	0.7871 ± 0.0124	0.6820 ± 0.0132
1	0.8763 ± 0.0028	0.7613 ± 0.0073
1.5	1.0009 ± 0.0098	0.8984 ± 0.0336

G.4 Composition residuals

At splits $(0.25, 0.25)$, $(0.5, 0.5)$, and $(0.5, 1.0)$, the maximum point-composition residual of the composing model is 1.49×10^{-16}. Across both datasets, the compatible kernel's maximum latent multiplier residual is 1.11×10^{-16} and maximum relative covariance residual is 2.50×10^{-16}. These are numerical checks of identities enforced by the architecture. Their magnitudes are not accuracy errors against the true law.

For each density model, sampled direct/composed W_1 uses 24 evenly spaced initial states in $[-1.8, 1.8]$, 1,024 samples per state, and three sampling repetitions. Independent direct/direct draws supply a sampling baseline. Compatible-kernel direct/composed values are of the same order as this baseline; no unbiased correction is claimed by subtracting the two.

H Lorenz illustration and prediction objects

At the state illustrated in Figure 1, restarting a noisy Lorenz ensemble at its intermediate mean changes the final mean by about 34.75 in Euclidean distance. This is a physical-process illustration of the composition defect, rather than an evaluation of a trained predictor. The full observed state follows

$$dX_t = b(X_t)\,dt + dW_t, \qquad b(x_1, x_2, x_3) = \big(10(x_2 - x_1),\ x_1(28 - x_3) - x_2,\ x_1 x_2 - \tfrac{8}{3} x_3\big),$$

with independent Brownian components. With $h = 0.6$ and step $\delta = 0.002$, stochastic Heun uses

$$X^p = X + \delta b(X) + \Delta W, \qquad X^+ = X + \tfrac{\delta}{2}\{b(X) + b(X^p)\} + \Delta W, \quad \Delta W \sim \mathcal{N}(0, \delta I_3).$$

The same increment enters the predictor and corrector. Starting at $x_0 \approx (0.0793, 0.4290, 15.4355)$, we simulate 5,000 particles to h and retain their mean $\widehat{m}_h(x_0)$. One forecast continues this ensemble for another h; the other starts 5,000 new particles at its estimated mean and evolves them for h. The latter uses a data-dependent restart state and is not an unbiased nested estimate of $m_h(m_h(x_0))$.

Table 9: Mean locations in the Lorenz illustration, rounded to two decimals. The two terminal means differ by 34.75 in three dimensions and 23.19 in the displayed (x_1, x_3) projection (computed from unrounded means).

Distribution	Time	$\widehat{m}_1$	$\widehat{m}_2$	$\widehat{m}_3$
Intermediate ensemble	0.6	11.86	7.19	35.20
Continued ensemble	1.2	-10.17	-9.99	30.61
Restarted at the mean	1.2	13.00	15.89	29.46

Selection and numerical sensitivity. The initial state had the third-largest estimated gap among 200 pilot candidates, using 300 particles per ensemble. It was selected to make the mechanism visible, not to estimate its prevalence. Independent simulations at the same state gave separations of 34.73 with 2,048 particles and 34.59 with 8,192 particles at step 0.002; halving the step with 8,192 particles gave 34.61. These checks support the visible separation but do not provide a population uncertainty interval or a rigorous discretization-error bound. The chosen horizon does not test a small-time expansion.

The reported distances concern one selected state, not the population L^2 defect Δ in (3). A fixed initial state also need not satisfy the coverage assumption of Theorem 1; this illustration does not evaluate its lower bound.

Panel a shows a deterministic trajectory for geometric orientation. Its visual smoothness is not a property inferred from the noisy endpoint clouds in panels b–c. Their faint background is a separate long noisy trajectory; neither background is a stochastic-support boundary. The simulation parameters and state-selection procedure above specify this illustration. The supplementary notebooks reproduce the main analytic and forecasting comparisons.

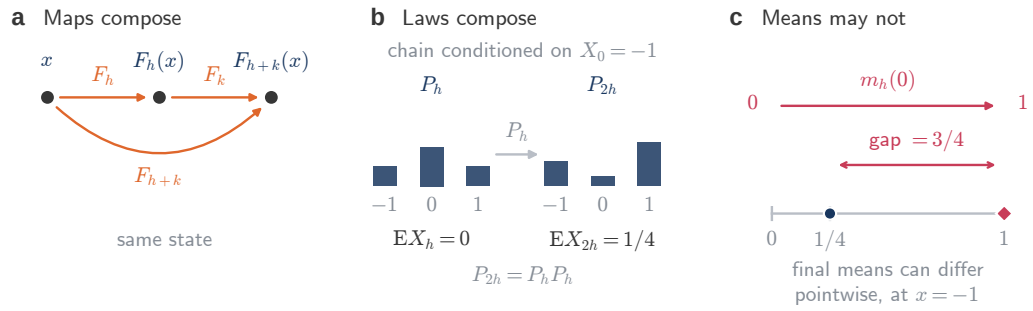

Figure 8: Compose the appropriate prediction object. A, Direct and composed maps agree on a sufficient deterministic state (schematic). **B,** For the paper's three-state chain started at -1, the intermediate law $(1/4, 1/2, 1/4)$ propagates to $(5/16, 1/8, 9/16)$, exactly matching the direct two-step law. Bar heights are probabilities and h denotes one chain step. **C,** Preserving the law gives final mean $1/4$, while restarting from its intermediate mean 0 gives $m_h(0) = 1$. The pointwise gap is $3/4$; it is distinct from the uniform-initial-law L^2 defect $\sqrt{3}/4$. B/C use exact arithmetic, not measured data.

I Oracle supervision and observed-state regression

The oracle experiments use three initialization seeds for each model and noise level. The Lorenz experiments use three independently generated datasets at each of two noise levels; each dataset is shared by four forecasting rules. Models and reference simulations use 64-bit arithmetic with TF32 disabled. Initialization variability in the oracle study and data variability in the Lorenz study are reported separately.

I.1 Oracle-mean supervision and numerical integration

Let $dZ = -Z\,dt + \sqrt{q}\,dW$, $Z_0 \sim \mathcal{N}(0,1)$, and $X = \psi_{3/5}(Z)$, where $\psi_c(z) = z + cz^3$. The initial laws for a pair starting at time s have latent variance $e^{-2s} + q(1 - e^{-2s})/2$. We minimize the sum of mean-prediction errors for $(s,h) = (0, 0.5), (0.5, 0.5), (0, 1)$, using the exact mean

$$m_h(x) = u + \tfrac{3}{5}(u^3 + 3uv), \qquad u = e^{-h}\psi_{3/5}^{-1}(x), \quad v = q(1 - e^{-2h})/2.$$

The cubic coefficient. The choice $c = 3/5$ fixes the numerical benchmark; it is not required for non-composition. For any $c > 0$, $\psi_c'(z) = 1 + 3cz^2 \geq 1$, so the chart is a smooth increasing bijection. Put $r_t = e^{-t}$, $v_t = q(1 - r_t^2)/2$, $z = \psi_c^{-1}(x)$ and $w = \psi_c^{-1}(m_h(x))$. Gaussian moments and the tower property give

$$m_t(x) = r_t^3 x + r_t(1 - r_t^2 + 3cv_t)\psi_c^{-1}(x),$$
$$m_k(m_h(x)) - m_{h+k}(x) = r_k(1 - r_k^2 + 3cv_k)(w - r_h z).$$

Indeed, applying the first identity at time h cancels the terms $r_k^3 m_h(x)$, while $\mathbb{E}[\psi_c^{-1}(X_h) \mid X_0 = x] = r_h z$. Since $\psi_c(w) = \psi_c(r_h z) + 3cr_h z v_h$, the second line is nonzero for $c, q, h, k > 0$ and $x \neq 0$. At $x = 0$ it vanishes by symmetry; $c = 0$ or $q = 0$ also restores mean composition. Thus the effect is not specific to the numerical coefficient.

Gauss–Hermite quadrature uses 96 nodes per pair for fitting, 160 for checkpoint validation, and 256 for final evaluation; nodes and weights are scaled by $\sqrt{2}$ and $1/\sqrt{\pi}$. The covariance intensities

$q = 0, 0.1, 0.3, 1$ and initialization seeds $0, 1, 2$ were fixed before the run. Validation excludes supplemental off-grid horizons.

The composing map is $F_h = \psi_c(e^{-ah}\psi_c^{-1}(\cdot))$, with learned $a, c > 0$. Its compatible-kernel extension adds a learned $\widehat{q} \geq 0$, latent variance $\widehat{q}(1 - e^{-2ah})/(2a)$, and evaluates the exact cubic conditional mean. We parameterize a, c exponentially and $\widehat{q}$ as the square of a real parameter, allowing zero noise. These models have two and three parameters, respectively. The compatible family contains the data-generating kernel. Unlike the finite tanh chart in Section G, its polynomial observation tails are correctly specified.

Both neural controls have two width-64 SiLU layers and 4,417 parameters. The direct model is $x + h\sqrt{10}\, f_\theta(x/\sqrt{10}, \log(1 + h))$. Neural autoregression evaluates this network only at lag 0.5 and uses its iterates for integer multiples, so composition is exact on that grid. The networks share initialization seeds and optimizer budgets. Both receive 6,000 Adam updates, learning rate 0.001 with cosine decay to 5%, gradient clipping at 10, and validation every 250 updates. All four models then use L-BFGS with a budget of 350 iterations in blocks of 25, step size 0.7, strong-Wolfe search, history size 40, gradient tolerance 10^{-11}, and change tolerance 10^{-14}. Chart models use only L-BFGS. Minimum validation risk selects the checkpoint. Positive fitted errors need not be global infima. The equal-objective control removes the density-versus-mean loss difference, but does not equalize parameter counts across the chart and neural families.

Table 10: Complete oracle three-pair risks, mean $\pm$ initialization-seed standard deviation. The compatible family contains the truth. Near-zero entries are numerical fitting error.

q	Cubic map	Compatible mean	Neural AR	Direct MLP
0	$(2.97 \pm 2.24) \times 10^{-15}$	$(1.76 \pm 0.20) \times 10^{-12}$	$(2.68 \pm 1.19) \times 10^{-6}$	$(1.50 \pm 0.41) \times 10^{-4}$
0.1	$(1.78 \pm 0.00) \times 10^{-4}$	$(1.66 \pm 0.54) \times 10^{-15}$	$(2.58 \pm 0.26) \times 10^{-5}$	$(1.30 \pm 0.16) \times 10^{-4}$
0.3	$(2.14 \pm 0.00) \times 10^{-3}$	$(3.13 \pm 0.73) \times 10^{-15}$	$(2.52 \pm 0.01) \times 10^{-4}$	$(1.39 \pm 0.62) \times 10^{-4}$
1	$(3.82 \pm 0.00) \times 10^{-2}$	$(4.89 \pm 0.82) \times 10^{-15}$	$(4.96 \pm 0.01) \times 10^{-3}$	$(9.03 \pm 7.84) \times 10^{-5}$

For the cubic map at positive q, seed standard deviations are below 1.2×10^{-14}; their displayed zeros reflect rounding. At $q = 1$, the target's mean-composition defect is nonzero, while neural AR has zero model defect by construction. The fitted compatible mean retains the target defect: its pointwise noncomposition is consistent with exact kernel composition. In the zero-noise control both chart models approach the deterministic target. Direct MLP fitting is less accurate than neural AR at $q = 0, 0.1$, and more accurate at $q = 0.3, 1$ in all three seeds. This illustrates why an asymptotic compatibility obstruction does not imply a universal finite-fit ordering.

The largest 128-versus-256-node risk discrepancy is 1.02×10^{-5} for the direct MLP, 7.68×10^{-7} for neural AR, 1.99×10^{-10} for the cubic map, and 2.40×10^{-17} for the compatible mean. The direct MLP's integration error can be material relative to its small risk; these are quadrature diagnostics, not rigorous error bounds. Per-pair predictions, quadrature weights, scalar parameters, per-horizon diagnostics, and objective histories are archived without selecting or discarding an initialization.

I.2 Lorenz data, features, and target construction

Use the drift in Section H, with noise amplitude $\eta \in \{0, 1\}$ and the full three-dimensional state. Stochastic Heun uses step 0.002, and observations are spaced by $\tau = 0.05$. Each trajectory starts at $(1, 1, 25)$ plus independent standard Gaussian coordinates and runs for 20 physical time units before collection. Sixty training trajectories provide 60,000 aligned initial-time pairs; 20 separate validation trajectories provide 4,000. Forecast leads are $1, 2, 4, 8, 16, 32$ observations. Pairs within a trajectory overlap and are not claimed independent. Test states come from 128 separate burn-in trajectories, independent of the training and validation data, rather than neighboring snapshots of one trajectory.

Normalize states using center $(0, 0, 23.5)$ and scales $(8, 9, 8.5)$. The common feature vector contains a constant, three linear terms, six quadratic terms, and 2,800 cosine random Fourier features: 700 frequencies at each bandwidth $0.08, 0.16, 0.32, 0.64$. Frequencies are Gaussian divided by bandwidth, phases are uniform on $[0, 2\pi]$, and Fourier features are scaled by $\sqrt{2/2800}$. Feature seed

1234 is fixed across all data seeds. There are 8,430 regression weights per map. AR stores one map; the other families store six (50,580 weights each).

For $G = \Phi^\top \Phi/N$, ridge strength is $\lambda = \alpha \operatorname{tr}(G)/2810$. Choose $\alpha \in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$ using one-step validation MSE, then use it for every method and horizon. All noiseless seeds select 10^{-7}; all noisy seeds select 10^{-4}. These are grid endpoints. Regularized Gram condition numbers are approximately 1.34×10^{10} and 1.34×10^{7}, respectively; solves use float64. The study uses a shared one-step tuning rule. Section J.4 tunes each method separately and tests a second feature budget.

All methods share the fitted one-step map f. AR iterates it at inference. At doubling $h \to 2h$, Direct regresses observed X_{2h}, PointBoot regresses $F_h(F_h(X_0))$, and TowerBoot regresses $F_h(X_h)$, each onto the same features of X_0. Teachers are fixed from the preceding level and the training trajectories are reused. PointBoot and TowerBoot thus differ only in the teacher's evaluation state at each doubling, though their teachers subsequently differ. Projection means PointBoot is not an exactly compositional map family. TowerBoot requires observed intermediate states at the applicable horizons; no arbitrary-horizon recovery from one fixed snippet length is claimed. The population identity (10) does not guarantee a finite-sample contraction for these reused-data fitted teachers.

The complete shared-fit results are reported in the tables below. The broader tuning and rollout controls are summarized in Figure 9.

I.3 Reference means, refinement, and all-horizon results

For each noisy test state, simulate two independent halves of 2,048 particles each. With means $\widehat{m}_A, \widehat{m}_B$, a fixed predictor F has per-state score

$$\widehat{\mathcal{E}}(x) = \big(F(x) - \widehat{m}_A(x)\big)^\top \big(F(x) - \widehat{m}_B(x)\big).$$

Conditional independence and unbiased reference means give $\mathbb{E}[\widehat{\mathcal{E}}(x) \mid x, F] = \|F(x) - m(x)\|^2$ for the discretized process. Scores may be negative and are not clipped. At $\eta = 0$, one deterministic trajectory per half suffices. Model fitting and reference simulations are independent; the same references are reused for paired comparisons across methods and horizons.

A paired reference uses step 0.001, splitting each coarse Brownian increment into two fine increments. Main results use this refined reference. With noise, the largest absolute coarse-to-fine change in a reported risk is 0.02676, small relative to the primary model gaps. At lag 1.6, the noisy pathwise paired endpoint RMS is 0.817–0.966 across seeds: trajectory refinement error is not negligible merely because mean-risk comparisons are stable. Without noise, the maximum risk change across arms is 2.65 for the very inaccurate PointBoot model. AR remains the best long-horizon predictor under both references. These checks do not certify continuous-time convergence.

Table 11: Lorenz without noise: refined-reference excess risk, mean ± sample standard deviation over three independently generated data seeds. Lower is better.

Lag	AR	Direct	PointBoot	TowerBoot
0.05	$(3.38 \pm 0.53) \times 10^{-7}$	$(3.38 \pm 0.53) \times 10^{-7}$	$(3.38 \pm 0.53) \times 10^{-7}$	$(3.38 \pm 0.53) \times 10^{-7}$
0.1	$(1.25 \pm 0.10) \times 10^{-6}$	$(1.27 \pm 0.08) \times 10^{-6}$	$(1.29 \pm 0.06) \times 10^{-6}$	$(1.27 \pm 0.07) \times 10^{-6}$
0.2	$(3.01 \pm 0.26) \times 10^{-6}$	$(3.15 \pm 0.33) \times 10^{-6}$	$(3.22 \pm 0.45) \times 10^{-6}$	$(3.17 \pm 0.38) \times 10^{-6}$
0.4	$(4.28 \pm 0.24) \times 10^{-6}$	1.297 ± 0.74	1.297 ± 0.74	1.298 ± 0.74
0.8	$(2.38 \pm 0.83) \times 10^{-5}$	19.01 ± 4.2	101.6 ± 15	19.15 ± 4.1
1.6	0.03173 ± 0.054	42.88 ± 4	$(1.29 \pm 0.55) \times 10^{5}$	42.42 ± 3.8

With noise, TowerBoot improves over AR and PointBoot at both primary horizons in each seed. Its lag-1.6 improvement over Direct also holds in every seed, but the 0.8 comparison is mixed. PointBoot's poor long-horizon behavior without noise shows that its failure cannot be attributed solely to a stochastic expectation gap: projection and rollout amplification also matter. Conversely, AR's much lower deterministic error demonstrates that an exact composition constraint can be appropriate for the prediction target.

Table 12: Lorenz with unit additive noise: refined-reference excess risk, mean ± sample standard deviation over three data seeds. All methods share the one-step fit.

Lag	AR	Direct	PointBoot	TowerBoot
0.05	0.002483 ± 0.00015	0.002483 ± 0.00015	0.002483 ± 0.00015	0.002483 ± 0.00015
0.1	0.004799 ± 0.00048	0.004727 ± 0.00038	0.003482 ± 0.00039	0.003669 ± 0.00033
0.2	0.01177 ± 0.0018	0.01385 ± 0.0008	0.00768 ± 0.0011	0.01011 ± 0.0017
0.4	0.1667 ± 0.042	1.659 ± 0.26	1.63 ± 0.14	1.669 ± 0.26
0.8	20.08 ± 4	3.354 ± 1.1	9.954 ± 1.8	3.291 ± 1.1
1.6	61.69 ± 9.3	5.528 ± 0.63	46.93 ± 2.8	4.32 ± 0.45

Table 13: Long-horizon conditional-mean excess risk for every data seed, with and without noise.

η	Seed	Lag	AR	Direct	PointBoot	TowerBoot
0	0	0.8	3.3276×10^{-5}	17.851	104.7	18.058
0	1	0.8	1.9548×10^{-5}	15.49	84.933	15.743
0	2	0.8	1.8474×10^{-5}	23.687	115.19	23.651
0	0	1.6	0.094 135	44.624	1.5227×10^{5}	45.239
0	1	1.6	0.000 939 57	38.323	66 521	38.063
0	2	1.6	0.000 105 95	45.692	1.6873×10^{5}	43.951
1	0	0.8	19.54	3.6528	8.1314	3.4499
1	1	0.8	24.363	4.2492	10.039	4.3207
1	2	0.8	16.349	2.1595	11.691	2.1025
1	0	1.6	57.974	5.8832	48.963	4.7502
1	1	1.6	54.851	5.8974	43.708	4.3567
1	2	1.6	72.23	4.8026	48.125	3.8539

J Motion-capture forecasting and Lorenz training controls

We compare forecasting models on two CMU protocols, vary their consistency penalties, and test intermediate mean collapse. The Lorenz controls vary the feature budget, ridge selection rule, and rollout training. There are 39 distinct CMU training configurations; shared baseline fits are reused across comparisons. Lorenz uses two feature budgets, two noise levels, and three data seeds.

J.1 CMU data, models, and evaluation

We use the public `mocap35.mat` preprocessing associated with ODE2VAE (Yıldız et al., 2019), derived from the CMU Graphics Lab Motion Capture Database. The arrays contain 16 training, three validation, and four test sequences, each of shape 300×50. We use the supplied split and channels, without new PCA or test-fitted preprocessing.

Setup 1 permits training frames 0–299, uses test context frames 0–2, and evaluates frames 3–299 (297 targets). Setup 2 permits training frames 0–199, supplies context frames 0–99, and evaluates frames 100–299 (200 targets). In both cases our models use only the last three available context poses; setup 2 does not assimilate the first 97. Time per frame is 0.1. Per-channel centering and scaling use permitted training frames only, with standard-deviation floor 0.1. Predictions are decoded into the original units for evaluation.

The model state is the concatenation of three standardized poses, ordered oldest to newest. Forecasts score only the last 50 coordinates of the predicted state. Training draws disjoint input and target delay windows, with end times separated by at least three frames. Half the origins use the benchmark context endpoint and half use other valid training times. Evaluation includes leads of one and two frames, which are outside these sampled training lags. The architecture does not enforce the physical overlap constraints of adjacent delay windows. Consequently, exact compatibility concerns the learned 150-dimensional state kernel; it is not a claim that the human motion process is Markov in three poses or that its 50-dimensional projection automatically has the same property.

The direct point model is an unconstrained residual MLP (55,574 parameters). The conjugate point map has 182,660 parameters and uses an invertible four-layer coupling chart with width-128 net-

Table 14: Paired Direct-minus-TowerBoot risk at noisy lag 1.6. Positive values favor tower. The within-run standard error uses the 128 paired test-state scores and includes reference sampling variation; it is not the between-seed standard deviation.

Seed	Paired difference	Within-run standard error
0	1.133	0.287
1	1.541	0.629
2	0.949	0.376

works and 75 damped two-dimensional latent rotation blocks with a learned equilibrium. The compatible kernel adds one positive block-isotropic diffusion rate per block (182,735 parameters). The loose Gaussian-chart model has an independently learned lag-dependent latent variance (186,330 parameters), while its affine mean evolution still composes. This supplies a within-family covariance control, not an unrestricted density oracle. For a block with decay $\lambda > 0$, frequency ω, and diffusion rate $q > 0$, $R_h = e^{-\lambda h}\,\mathrm{Rot}(\omega h)$ and $\Sigma_h = q(1 - e^{-2\lambda h})I_2/(2\lambda)$. The latent mean is $c + R_h(z - c)$. All four families use identical data, history length, and nominal update budgets; capacities and training objectives are not matched.

Point models minimize standardized full-state MSE and density models minimize conditional NLL per state coordinate. Adam uses learning rate 10^{-3}, batch size 128, and 6,000 updates; the matching validation objective is evaluated every 500 updates and selects the returned checkpoint. Test data do not select checkpoints or penalty strengths. Some point fits select the final checkpoint, while density fits often select earlier checkpoints. These budgets are not convergence certificates. Exact architectural identities hold in ideal arithmetic; implemented inverses, floating-point operations, and variance safeguards have numerical error.

The main stochastic comparison uses 1,024 particles in setup 1 and 100 in setup 2. For target pose $y \in \mathbb{R}^{50}$ and predictive law Q, the energy score is

$$\mathrm{ES}(Q, y) = \mathbb{E}_{U\sim Q}\|U - y\|_2 - \tfrac{1}{2}\mathbb{E}_{U,V\overset{\mathrm{iid}}{\sim}Q}\|U - V\|_2.$$

The implementation uses disjoint member pairs for the second expectation. A deterministic forecast yields Euclidean error, not coordinatewise MSE, under this score. MSE averages squared forecast-mean error over channels, horizons, and the four sequences; RMSE is its square root. Coverage uses marginal central-90% intervals, not simultaneous trajectory bands. Forecast-mean MSE includes the Monte Carlo error of the finite ensemble mean; with only 100 members, setup 2 can have a larger contribution than setup 1. Samplewise MSE is reported separately.

Table 15: CMU model size and interval diagnostics, means over three seeds. Width is in original channel units. Deterministic forecasts have degenerate intervals and are omitted here.

Setup	Model	Params.	Coverage (%)	Mean width
1	Compatible kernel	182,735	80.69	5.746
1	Loose density	186,330	73.55	5.869
2	Compatible kernel	182,735	87.29	6.444
2	Loose density	186,330	81.22	5.275

All three seeds share one fixed dataset split. Frames within a sequence are correlated; the four held-out sequences are the test data units. Three model seeds quantify initialization and fitting variability, not population uncertainty across datasets. Published scores are not included in the numerical comparison because their MSE/RMSE and samplewise/forecast-mean conventions do not establish a matched evaluation protocol.

J.2 Frozen-kernel collapse and Monte Carlo controls

For each setup-1 kernel, freeze the fitted parameters and initial delay state. At each forecast lead, compare direct sampling with compositions over 2, 4, or 8 equal subintervals. Full propagation keeps each decoded ensemble member and applies independent innovations at the next step. Collapse replaces the decoded 150-dimensional ensemble by its empirical mean at every intermediate

boundary; the final step still returns an ensemble. This is a mean in observation coordinates, not the affine latent mean. The one-step case has no collapse. The intervention compares predictions of the same fitted model against the same held-out targets, removing refitting and parameter-count differences.

Use 256 and 512 members and two Monte Carlo repeats per configuration. Average the repeats within each model seed before computing across-seed summaries. Direct and full-ensemble compositions have the same ideal model law; their finite-sample differences provide a numerical reference. In Figure 4a, percentage changes are computed within each seed before averaging. The main-text percentages instead compare the across-seed mean scores. Panel b averages paired MSE differences at each lead over seeds, repeats, and test sequences; its columns are separately partitioned forecasts, rather than successive states of one rollout.

Table 16: Frozen setup-1 kernels at 512 particles. Entries are means across three model seeds, after averaging two Monte Carlo repeats. Variance is average marginal predictive variance. The number of intermediate collapses is one less than the number of steps.

Steps	Mode	MSE	ES	Coverage (%)	Variance
1	Direct	14.2177	17.7978	80.55	7.0079
2	Ensemble	14.2195	17.7992	80.57	7.0122
2	Collapse	14.5321	18.4152	75.83	5.5778
4	Ensemble	14.2195	17.7946	80.57	7.0119
4	Collapse	14.9988	19.2511	69.77	4.2640
8	Ensemble	14.2141	17.7920	80.52	7.0122
8	Collapse	15.4048	20.2440	61.93	3.0235

Table 17: Paired eight-step collapse effects by model seed. Each entry subtracts full propagation from collapse, after averaging two Monte Carlo repeats. Positive values mean worse scores.

	256 particles		512 particles	
Seed	ΔMSE	ΔES	ΔMSE	ΔES
0	0.8865	2.1608	0.8846	2.1512
1	1.1271	2.3822	1.1181	2.3648
2	1.5706	2.8405	1.5696	2.8400

With 512 particles, the mean paired MSE increase is 1.1908; the descriptive seed-based t interval is $[0.3258, 2.0558]$ with only two degrees of freedom. This is not an interval over new motion sequences. After averaging seeds, one of four test sequences improves under collapse and three worsen: sequence MSE changes are -0.3670, 0.8410, 2.5005, and 1.7886. Thus the aggregate effect is consistent across model seeds, not across every test sequence. Predictive variance falls from 7.0122 to 3.0235, while coverage falls from 80.52% to 61.93%. The full ensemble already under-covers. Samplewise squared error can decrease because of the variance reduction even when forecast-mean MSE and the proper energy score deteriorate.

J.3 Consistency regularization

The point-model sweep adds a squared composition residual in standardized state coordinates to the MSE objective, with weights $0, 0.01, 0.1, 1$ warmed over 500 updates. The loose density sweep adds symmetric Gaussian KL per latent coordinate between the direct and composed variances to the NLL objective, with weights $0, 0.1, 1$. We report every setting, without selecting a penalty by test score. Measured validation defects average three held-out splits and three seeds; point squared error and symmetric KL are distinct units.

Increasing point regularization to 1 reduces mean MSE by 12.1% and the measured squared defect by 73.6%. The sweep therefore does not exhibit the adverse risk–consistency frontier of the analytic example. The loose density with weight 1 also slightly beats the compatible kernel's mean setup-1 MSE. Finite-sample regularization can be beneficial while the population compatibility statement remains true; this dataset does not establish its nonzero-defect or coverage assumptions for the physical conditional means.

Table 18: Setup-1 ablations. MSE and ES are means ± seed SD. Validation defect is a mean squared point residual for the direct model and symmetric Gaussian KL per coordinate for the loose density. Smaller values are better within each block.

Family	Penalty	MSE	ES	Val. defect
Direct point MLP	0	22.62 ± 1.03	30.66 ± 0.55	0.3396
Direct point MLP	0.01	22.25 ± 1.70	30.34 ± 1.39	0.2804
Direct point MLP	0.1	21.52 ± 2.41	30.19 ± 1.90	0.2129
Direct point MLP	1	19.88 ± 3.97	28.83 ± 2.84	0.0896
Loose density	0	19.76 ± 1.71	21.82 ± 1.02	0.0641
Loose density	0.1	19.07 ± 1.84	21.43 ± 1.10	0.0563
Loose density	1	13.24 ± 3.50	17.49 ± 2.37	0.0311

J.4 Lorenz controls: selection, capacity, and all horizons

The data-generation and cross-product risk protocol follows Section I, using 60 independent training trajectories (60,000 pairs), 20 validation trajectories (4,000 pairs), 128 held-out test states, and noise amplitudes 0 and 1. Each reference has two independent halves of 2,048 members. The cross-product of errors to the two reference means estimates excess conditional-mean risk without their positive Monte Carlo variance bias. Refinement couples the 0.002 and 0.001 integration steps. We report the finer reference. Test states and reference particles are not additional training seeds.

The feature families use 350 or 700 features at each of four bandwidths $(0.08, 0.16, 0.32, 0.64)$ plus ten polynomial features, for 1,410 or 2,810 features. The feature seed is fixed at 1234. The six forecast leads are $0.05, 0.1, 0.2, 0.4, 0.8, 1.6$. Ridge candidates are $10^{-9}, 10^{-8}, \ldots, 10^{-1}$. AR-shared uses one-step validation; AR-tuned selects one ridge by the joint six-horizon observed-target validation loss. AR-rollout then trains the same coefficient matrix through its rollout for 1,500 Adam steps, learning rate 10^{-4}, batch size 64, with validation every 150 steps; the initial candidate is eligible for selection. Direct, tower, and point heads each select their ridge independently at every horizon using observed validation targets. No test conditional mean enters tuning.

All six methods share the training data and feature basis within a configuration. The three multihead methods store six coefficient matrices; AR stores one. Rollout training and feature sensitivity do not remove that capacity difference. Noiseless rollout fits all retain the initial candidate; one noisy fit selects the final allowed update. The optimizer is therefore not evidence that the autoregressive function class has been globally optimized.

Table 19: Lorenz excess mean risk, 1,410 features, noise amplitude 0. Mean ± SD across three seeds.

Lag	AR-shared	AR-tuned	AR-rollout
0.05	$(3.34 \pm 0.52) \times 10^{-7}$	$(3.34 \pm 0.52) \times 10^{-7}$	$(3.34 \pm 0.52) \times 10^{-7}$
0.1	$(1.24 \pm 0.11) \times 10^{-6}$	$(1.24 \pm 0.11) \times 10^{-6}$	$(1.24 \pm 0.11) \times 10^{-6}$
0.2	$(3.02 \pm 0.25) \times 10^{-6}$	$(3.02 \pm 0.25) \times 10^{-6}$	$(3.02 \pm 0.25) \times 10^{-6}$
0.4	$(4.11 \pm 0.20) \times 10^{-6}$	$(4.11 \pm 0.20) \times 10^{-6}$	$(4.11 \pm 0.20) \times 10^{-6}$
0.8	$(1.88 \pm 0.63) \times 10^{-5}$	$(1.88 \pm 0.63) \times 10^{-5}$	$(1.88 \pm 0.63) \times 10^{-5}$
1.6	0.006 ± 0.010	0.006 ± 0.010	0.006 ± 0.010
Lag	Direct	Tower	Point
0.05	$(3.34 \pm 0.52) \times 10^{-7}$	$(3.34 \pm 0.52) \times 10^{-7}$	$(3.34 \pm 0.52) \times 10^{-7}$
0.1	$(1.24 \pm 0.10) \times 10^{-6}$	$(1.24 \pm 0.09) \times 10^{-6}$	$(1.25 \pm 0.09) \times 10^{-6}$
0.2	$(3.29 \pm 0.23) \times 10^{-6}$	$(3.29 \pm 0.24) \times 10^{-6}$	$(3.31 \pm 0.25) \times 10^{-6}$
0.4	3.958 ± 1.884	3.958 ± 1.884	3.958 ± 1.884
0.8	23.387 ± 3.351	23.352 ± 3.438	2627.373 ± 371.330
1.6	56.105 ± 3.047	55.894 ± 3.421	$459910.403 \pm 217227.655$

At unit noise, 2,810 features, and $h = 1.6$, the estimated conditional-variance contribution is 111.5961. Adding it to excess risk gives approximately 159.6374 for rollout AR and 116.0628 for tower, a 27.3% total-risk reduction. The corresponding 90.7% figure is specific to excess risk. Across the three seeds, halving the integration step changes the paired long-horizon AR-rollout minus tower gap by at most 0.0367, compared with the mean gap of 43.5746. These checks assess a tested refinement and Monte Carlo controls, not a rigorous global discretization bound.

Table 20: Lorenz excess mean risk, 1,410 features, noise amplitude 1. Mean ± SD across three seeds.

Lag	AR-shared	AR-tuned	AR-rollout
0.05	0.002 ± 0.000	0.002 ± 0.000	0.004 ± 0.001
0.1	0.003 ± 0.000	0.003 ± 0.000	0.011 ± 0.002
0.2	0.008 ± 0.002	0.008 ± 0.002	0.032 ± 0.003
0.4	0.153 ± 0.029	0.154 ± 0.030	0.341 ± 0.082
0.8	19.541 ± 3.655	19.614 ± 3.819	17.889 ± 2.450
1.6	60.294 ± 7.598	60.661 ± 8.013	54.077 ± 4.819
Lag	Direct	Tower	Point
0.05	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000
0.1	0.004 ± 0.000	0.003 ± 0.000	0.002 ± 0.000
0.2	0.010 ± 0.001	0.007 ± 0.001	0.005 ± 0.001
0.4	3.850 ± 0.606	3.870 ± 0.607	3.787 ± 0.627
0.8	5.781 ± 1.241	5.880 ± 1.113	14.866 ± 2.678
1.6	8.037 ± 0.090	7.238 ± 0.383	36.325 ± 1.730

Table 21: Lorenz excess mean risk, 2,810 features, noise amplitude 0. Mean ± SD across three seeds.

Lag	AR-shared	AR-tuned	AR-rollout
0.05	$(3.32 \pm 0.54) \times 10^{-7}$	$(3.32 \pm 0.54) \times 10^{-7}$	$(3.32 \pm 0.54) \times 10^{-7}$
0.1	$(1.23 \pm 0.12) \times 10^{-6}$	$(1.23 \pm 0.12) \times 10^{-6}$	$(1.23 \pm 0.12) \times 10^{-6}$
0.2	$(3.02 \pm 0.24) \times 10^{-6}$	$(3.02 \pm 0.24) \times 10^{-6}$	$(3.02 \pm 0.24) \times 10^{-6}$
0.4	$(4.08 \pm 0.27) \times 10^{-6}$	$(4.08 \pm 0.27) \times 10^{-6}$	$(4.08 \pm 0.27) \times 10^{-6}$
0.8	$(1.62 \pm 0.45) \times 10^{-5}$	$(1.62 \pm 0.45) \times 10^{-5}$	$(1.62 \pm 0.45) \times 10^{-5}$
1.6	0.006 ± 0.011	0.006 ± 0.011	0.006 ± 0.011
Lag	Direct	Tower	Point
0.05	$(3.32 \pm 0.54) \times 10^{-7}$	$(3.32 \pm 0.54) \times 10^{-7}$	$(3.32 \pm 0.54) \times 10^{-7}$
0.1	$(1.23 \pm 0.12) \times 10^{-6}$	$(1.23 \pm 0.12) \times 10^{-6}$	$(1.23 \pm 0.12) \times 10^{-6}$
0.2	$(3.02 \pm 0.24) \times 10^{-6}$	$(3.02 \pm 0.24) \times 10^{-6}$	$(3.01 \pm 0.24) \times 10^{-6}$
0.4	1.183 ± 0.545	1.183 ± 0.545	1.183 ± 0.545
0.8	18.982 ± 4.349	19.100 ± 4.221	580.213 ± 516.961
1.6	42.474 ± 4.320	41.992 ± 3.997	3723.716 ± 3121.677

J.5 A CONSERVATIVE COVERAGE CONSTANT FOR THE CUBIC ORACLE

For the oracle process, write $\psi(z) = z + cz^3$ with $c = 3/5$, $h = 0.5$, $S = e^{-2h}$, $v = q(1-e^{-2h})/2$, and $V = S + v$. At the first intermediate time, the actual latent marginal is $\mathcal{N}(0, V)$. For $U \sim \mathcal{N}(0, S)$, the latent coordinate of the conditional observation mean is

$$T(U) = \psi^{-1}(AU + cU^3), \qquad A = 1 + 3cv.$$

The chart Jacobian cancels from the density ratio between the law of this intermediate mean and the actual observation marginal. For $u \geq 0$, set $d = T(u) - u \geq 0$. Since

$$d\{1 + c[T(u)^2 + T(u)u + u^2]\} = 3cvu,$$

we have $d \leq 3cvu/(1 + 3cu^2) \leq \sqrt{3c}\, v/2$ and $T(u)^2 - u^2 \leq 2v$. The latter follows from $3u(T + u) \leq 2(T^2 + Tu + u^2 + 1/c)$ for $T \geq u$. Consequently

$$T'(u) = \frac{A + 3cu^2}{1 + 3cT(u)^2} \geq \frac{A}{2A - 1}.$$

Oddness supplies the same bounds on the negative half-line. Maximizing the resulting quadratic bound on $T(u)^2/(2V) - u^2/(2S)$ gives, for $v > 0$,

$$\frac{d\,\mathrm{Law}(T(U))}{d\mathcal{N}(0,V)}(T(u)) = \sqrt{\frac{V}{S}}\frac{\exp\{T(u)^2/(2V) - u^2/(2S)\}}{T'(u)} \leq \underbrace{\frac{2A-1}{A}\sqrt{\frac{V}{S}}\exp(3cv/8)}_{\overline{C}}. \quad (27)$$

For $q = 0$, $T = \mathrm{id}$ and $C = 1$. At $q = 1$, $\overline{C} = 1.994864$ and quadrature gives $\Delta = 0.1063744$. Combining this analytic upper bound with each saved global Lipschitz bound yields the last column of Table 1. The cubic fit has $L = 0.643303$ and lower bound 0.0033196. Neural autoregression has bounds $L = 17.24, 46.58, 23.47$ across seeds, giving mean lower bound 2.1110×10^{-5}. Their larger

Table 22: Lorenz excess mean risk, 2,810 features, noise amplitude 1. Mean ± SD across three seeds. These fits use the tuning protocol in Section J.4 and differ from the shared fits in Sections I and K.

Lag	AR-shared	AR-tuned	AR-rollout
0.05	0.002 ± 0.000	0.003 ± 0.000	0.006 ± 0.001
0.1	0.004 ± 0.001	0.005 ± 0.001	0.015 ± 0.001
0.2	0.011 ± 0.002	0.012 ± 0.001	0.044 ± 0.005
0.4	0.165 ± 0.043	0.164 ± 0.037	0.397 ± 0.156
0.8	20.087 ± 4.064	20.155 ± 4.107	19.468 ± 5.176
1.6	62.884 ± 8.004	61.907 ± 11.064	48.041 ± 3.580

Lag	Direct	Tower	Point
0.05	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000
0.1	0.005 ± 0.001	0.004 ± 0.001	0.003 ± 0.001
0.2	0.014 ± 0.001	0.011 ± 0.002	0.007 ± 0.001
0.4	1.619 ± 0.227	1.639 ± 0.229	1.563 ± 0.133
0.8	4.134 ± 1.727	3.822 ± 1.369	8.869 ± 2.368
1.6	5.723 ± 0.103	4.467 ± 0.315	31.126 ± 8.723

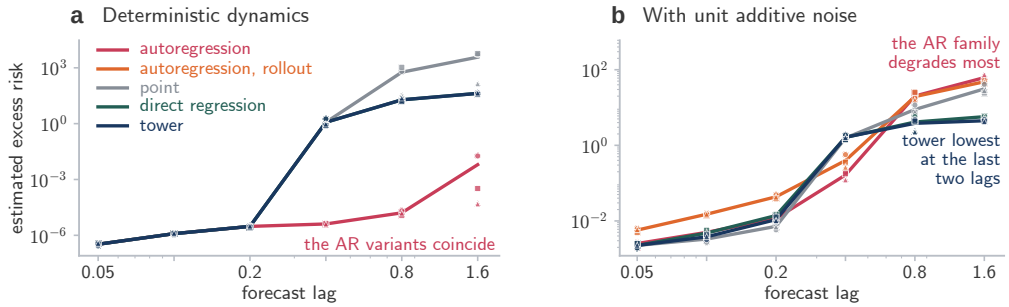

Figure 9: Noise changes the relative accuracy of forecasting rules. **a,** Without noise, autoregression is accurate and its three fitted variants coincide. **b,** With unit additive noise, tower regression has the lowest mean excess risk at the last two lags. All methods use 2,810 features; lines show means and symbols identify three seeds. Both risk axes are logarithmic, with different ranges. Point-bootstrap results are included, and direct and tower curves nearly coincide in a. The accompanying tables include all six methods and both feature budgets.

Lipschitz budgets explain the weaker certificate. The coverage bound is conservative and is not a grid-estimated supremum; reported risk and defect values are numerical quadrature evaluations.

The supplementary motion-capture notebook generates forecast means, variances, marginal quantiles, and scores. Means and quantiles determine MSE, RMSE, coverage, and interval widths; energy scores are computed from the sampled particles during execution.

K Stochastic Propagation of a Fixed Lorenz Map

K.1 Fitted Map and Residual Transition Kernel

For each of three data seeds, we reuse the noisy test states, training pairs, one-step ridge map f (8,430 coefficients), and reference means of Section I; nothing is refitted. This shared-fit comparison is separate from the independently tuned controls in Section J.4. For its 60,000 training pairs (x_i, y_i) define

$$r_i = y_i - f(x_i), \qquad \widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} r_i r_i^{\top}, \qquad K(x, \cdot) = \mathcal{N}(f(x), \widehat{\Sigma}).$$

This is the second-moment estimate for a zero-mean Gaussian innovation model; residuals are not recentered and the fitted conditional-mean map is unchanged. The covariance uses training data only. Its six distinct entries add a small fitted uncertainty model to the existing point forecaster. It may absorb both transition noise and one-step approximation error; no exact identification with the physical Lorenz transition law is assumed.

For each initial state, independently evolve M particles by $Y_{j+1}^{(b)} = f(Y_j^{(b)}) + L\xi_j^{(b)}$, where $LL^\top = \widehat{\Sigma}$ and the $\xi_j^{(b)}$ are independent standard Gaussians. The forecast is the particle mean. The 512-member evaluation uses a $10^{-12}I$ Cholesky safeguard; the 256-member full and collapsed arms both use the same unmodified covariance. Kernel powers K^n compose exactly on the observation grid $\tau = 0.05$. This is a classical ensemble forecasting construction (Leith, 1974); it defines transitions only at integer multiples of the observation interval. The control fixes the map rather than inference compute: propagating M particles requires M map evaluations per step instead of one.

K.2 Finite-ensemble prediction risk

We score each particle-mean forecast with the same two independent refined-reference halves as the deterministic baselines. Since model particles are independent of those references, the cross-product estimator in Section I evaluates the finite-ensemble forecaster's excess risk for the discretized reference process. We report raw finite-ensemble scores, without subtracting Monte Carlo variance.

Table 23: Shared-fit Lorenz predictors with unit noise: refined-reference excess risk, mean ± sample SD across three data seeds. The stochastic arm propagates the same one-step map used by AR, with a training-residual Gaussian innovation and 512 particles.

Lag	AR	Direct	Tower	Stochastic AR
0.05	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.003 ± 0.000
0.1	0.005 ± 0.000	0.005 ± 0.000	0.004 ± 0.000	0.005 ± 0.000
0.2	0.012 ± 0.002	0.014 ± 0.001	0.010 ± 0.002	0.011 ± 0.002
0.4	0.167 ± 0.042	1.659 ± 0.264	1.669 ± 0.256	0.064 ± 0.044
0.8	20.084 ± 4.035	3.354 ± 1.076	3.291 ± 1.118	0.286 ± 0.105
1.6	61.685 ± 9.265	5.528 ± 0.628	4.320 ± 0.449	0.603 ± 0.217

Stochastic propagation improves prediction in all three seeds at lag 1.6. The 0.603 mean risk is for the implemented 512-member forecast. Subtracting the independent-particle sampling term gives an estimated risk of 0.381 for its population ensemble mean; the main comparison uses the raw finite-ensemble score.

K.3 One collapse with matched particle count and randomness

At lag 0.8, replace the 256-member ensemble by 256 copies of its arithmetic mean, then continue stochastically to lag 1.6. The comparison keeps all 256 particles. Both variants use the same map, covariance, and Gaussian random numbers, including after the intervention. Their means agree at the collapse time; subsequent predictions differ because one variant discards the intermediate distribution. Both finish with a stochastic ensemble.

Table 24: One midpoint collapse at fixed map and 256 particles, lag 1.6. Raw finite-ensemble excess risk. The paired increase is positive in every data seed.

Seed	Full ensemble	One collapse	Paired increase
0	0.651	75.785	75.135
1	0.979	60.673	59.694
2	0.777	59.817	59.040
Mean ± SD	0.802 ± 0.166	65.425 ± 8.982	64.623 ± 9.109

The mean increase is about 64.62, from 0.802 ± 0.166 to 65.425 ± 8.982. Both variants use 256 particles, controlling the number of map evaluations. This measures the change in prediction risk after a propagation intervention. It does not estimate the mean-composition defect of either the fitted kernel or the physical process.

Table 25: More frequent intermediate collapses, with 256 particles. Means ± SD across three data seeds; all entries are raw finite-ensemble risks. Schedules use different random streams; only the one-collapse row has the matched full-ensemble comparison in Table 24.

Collapse interval	Collapses before 1.6	Risk at 0.8	Risk at 1.6
0.8	1	0.476 ± 0.140	65.425 ± 8.982
0.4	3	19.231 ± 5.016	60.468 ± 10.673
0.2	7	20.124 ± 4.688	64.908 ± 5.919
0.1	15	20.632 ± 5.063	62.258 ± 9.308

K.4 Monte Carlo corrections and mean-defect estimation

After a shared intermediate mean collapse, terminal particles are independent conditional on that random mean but dependent marginally. Subtracting their sample variance divided by M removes only conditional terminal sampling variation. It omits variation propagated from earlier estimated means, so we report finite-ensemble risk. With collapse at every observation, this forecaster converges to point AR as $M \to \infty$.

A nested estimate of the physical defect can restart trajectories at $\widehat{m}_h = (\widehat{m}_{h,A} + \widehat{m}_{h,B})/2$ and compare their continuation means with $\widehat{m}_{2h,A}, \widehat{m}_{2h,B}$. Reusing paths for the intermediate and final references introduces dependence; independent continuation halves remove neither this problem nor the nonlinear plug-in bias. For Brownian motion the true defect is zero, but the dependence still produces bias. If each reference half has n particles, write its mean increments over successive lags as U_A, V_A, U_B, V_B, independent with variance h/n. After averaging out the independent outer simulation, the diagnostic has expectation

$$\mathbb{E}\big[(\tfrac{1}{2}(U_B - U_A) - V_A)(\tfrac{1}{2}(U_A - U_B) - V_B)\big] = -\frac{h}{2n}.$$

Thus a nested cross product of this form is biased even when the true defect is zero. Attributing physical prediction error to Δ^2 requires independent reference streams, control of the error from estimating the intermediate mean, and the shifted-error bound in (22).

Independent-stream diagnostic in Figure 3. The displayed diagnostic instead uses two fresh independent ensembles to estimate $m_h(x)$, then independently continues from their two means. Each continuation is compared with a separate saved reference half at $2h$ in a cross product. With $n_C = 8{,}192$ intermediate particles per ensemble and 2,048 particles per continuation, its expectation is $\|\mathbb{E}[m_h(\widehat{m}_h(x))] - m_{2h}(x)\|^2$ for the discretized process, rather than the population defect at finite n_C. The intermediate-mean plug-in error remains. At $h = 0.8$, the estimate is 65.276 ± 9.142 across three data seeds; using $n_C = 2{,}048$ gives 64.993. This sensitivity check is not a rigorous bias bound. The diagnostic uses Heun step 0.002, while the fitted-model risks use refined step-0.001 references. These saved coarse/fine references use coupled Brownian increments, so the two plotted axes share correlated reference Monte Carlo noise. The raw pooled Pearson correlation with the paired risk increase is 0.984874, over 128 states per seed. Signed axes retain negative and small estimates. The association is descriptive, not an independent statewise verification of a population identity.

For full and collapsed predictions F, C, and physical conditional mean m, the exact risk difference is

$$R(C) - R(F) = \mathbb{E}\|C - F\|^2 + 2\mathbb{E}\langle F - m, C - F\rangle.$$

The cross term need not vanish for a fitted kernel. Moreover, the count comparison in Figure 3b changes collapse timing as well as frequency: its multi-collapse schedules all include the midpoint. It is not a controlled effect of count alone.

K.5 Moment-preserving replacements

We additionally replace a 256-member cloud once, at lag 0.4 or 0.8, then continue to 1.6 with the same fitted map and covariance. Three Monte Carlo repeats are run for each of three data seeds, each with 128 test states. Besides preserving or collapsing the cloud, we replace it by moment-matched isotropic, diagonal, full-covariance, or two-lobe ensembles. Within each comparison, future Gaussian innovations are shared. Replacements preserve the mean; covariance constraints depend on the

arm. Full-covariance replacement uses a Gaussian-based finite cloud corrected to match the original empirical mean and covariance. This changes higher moments and geometry. No particle clipping is used.

Table 26: Moment matching has a limited range of success in the fitted Lorenz kernel. Terminal excess risk at lag 1.6, mean ± SD over three data seeds after averaging three Monte Carlo repeats within each. A failed run contains at least one nonfinite state forecast among 128. For arms with failures, no pooled risk over surviving runs is reported.

Replacement	Reset at 0.4	Reset at 0.8
Keep cloud	0.837 ± 0.183	0.837 ± 0.183
Mean collapse	11.857 ± 1.346	64.565 ± 5.882
Isotropic	2/9 runs failed	6/9 runs failed
Diagonal	2.507 ± 0.225	6/9 runs failed
Full covariance	0.889 ± 0.149	9/9 runs failed
Two lobes	0.887 ± 0.116	1/9 runs failed

At the earlier reset, full-covariance replacement nearly restores the risk of keeping the cloud: the paired difference is 0.052 ± 0.062. At the later reset, all nine full-covariance runs become nonfinite. The mean and covariance matching errors remain at numerical precision (largest absolute mean discrepancy 2.7×10^{-13} and relative covariance discrepancy 3.6×10^{-15}). These failures concern continuation through the fitted nonlinear map, which has no global stability guarantee under reshaped distributions. They do not establish instability of the physical Lorenz process or a general impossibility of finite-moment closure. The two reset times also show why a large cost from one midpoint collapse does not imply the same cost from a collapse at any intermediate time.

L EXPLORATORY FORECASTS ON THREE NONLINEAR SYSTEMS

Figure 10 extends the point-versus-law comparison to Lorenz, Thomas and Chen dynamics, using one fitted model and 64 test states per system. This study compares forecasts across different geometries; it does not measure variation across training seeds.

Dynamics and fitting. Each process solves $dX_t = b(X_t)\,dt + \eta\,dW_t$ with three independent noise coordinates. The vector fields, writing the state as (x, y, z), are

$$b_{\mathrm{L}}(x, y, z) = (10(y - x),\ x(28 - z) - y,\ xy - 8z/3),$$
$$b_{\mathrm{T}}(x, y, z) = (\sin y - 0.18x,\ \sin z - 0.18y,\ \sin x - 0.18z),$$
$$b_{\mathrm{C}}(x, y, z) = (35(y - x),\ -7x - xz + 28y,\ xy - 3z).$$

Observation spacing is $\tau = h/16$. Training uses a stochastic Heun step $\tau/25$; reference forecasts use $\tau/50$. The same ridge family as in Section J.4 is fitted on 60,000 pairs from 60 trajectories, with the ridge selected on 4,000 pairs from 20 separate validation trajectories. The map has 2,810 features: constant, linear and quadratic terms, and 2,800 fixed random Fourier features. Ridge candidates are $10^{-9}, \ldots, 10^{-3}$. The residual second moment supplies the additive Gaussian innovation covariance.

Table 27: Settings and estimated excess risk at $2h$ for the exploratory study. Each risk is an average over 64 states for one training seed. $V = 3s^2$ and s are fixed noiseless calibration constants used only for normalization, not estimates of the noisy process's stationary variance. Ratios compare the two average risks.

System	h	η	s	AR risk	Law risk	AR/law
Lorenz	0.80000	1.00000	8.523	60.6691	0.4009	151.3
Thomas	20.46554	0.05261	2.268	8.6334	0.0507	170.2
Chen	0.35738	1.41594	8.066	59.5041	0.7550	78.8

Forecasts and scores. The fitted law uses 512 particles. Excess risk is estimated by $(F - \widehat{m}_A)^\top (F - \widehat{m}_B)$, with two independent 512-particle reference halves, at ten leads through $8h$, including the fitted ensemble's sampling error. For predicted particles $p_1, \ldots, p_M$ and independent reference particles $y_1, \ldots, y_N$, the energy-discrepancy estimator is

$$\widehat{D}_E = \frac{2}{MN} \sum_{i,j} \|p_i - y_j\| - \frac{\sum_{i \neq i'} \|p_i - p_{i'}\|}{M(M-1)} - \frac{\sum_{j \neq j'} \|y_j - y_{j'}\|}{N(N-1)}.$$

We use $M = N = 512$ for particle forecasts. For AR, $M = 1$ and the predicted within-sample term is zero. This is the full energy distance, without a factor of $1/2$. The point climatology baseline is the training mean; the distributional baseline is a 512-state cloud. Coverage concerns coordinatewise 5–95% intervals, not a joint state region or simultaneous trajectory band.

Scope and numerical qualifications. The implementation projects both point and particle forecasts into the training bounding box expanded by 25% on each side. The recorded long-horizon projection counts are zero for Lorenz and Thomas, and one for Chen. Projection changes a Gaussian transition's conditional mean, so the guarded point rollout is not exactly the kernel's conditional mean. These counts do not measure projection's effect on risk. All plotted arrays are finite; long-horizon integration sensitivity is not assessed. At $2h$, the law's estimated risk is smaller on 64, 64 and 63 states, respectively. The examples are illustrative; aggregate curves use all 64 states.

The signed normalized energy estimates at $8h$ are -1.89×10^{-5} for Thomas and -1.39×10^{-4} for Chen. These are finite-sample estimates of a nonnegative population discrepancy. The figure marks them at a display floor; ratios involving that floor are not used as evidence. Marginal coverage ranges from 88.77–89.99% for Lorenz, 85.76–89.69% for Thomas and 88.73–89.51% for Chen. The examples suggest that distribution propagation can remain informative after the point rollout diverges, while independent fits and unguarded controls are needed to establish the robustness of these gains.

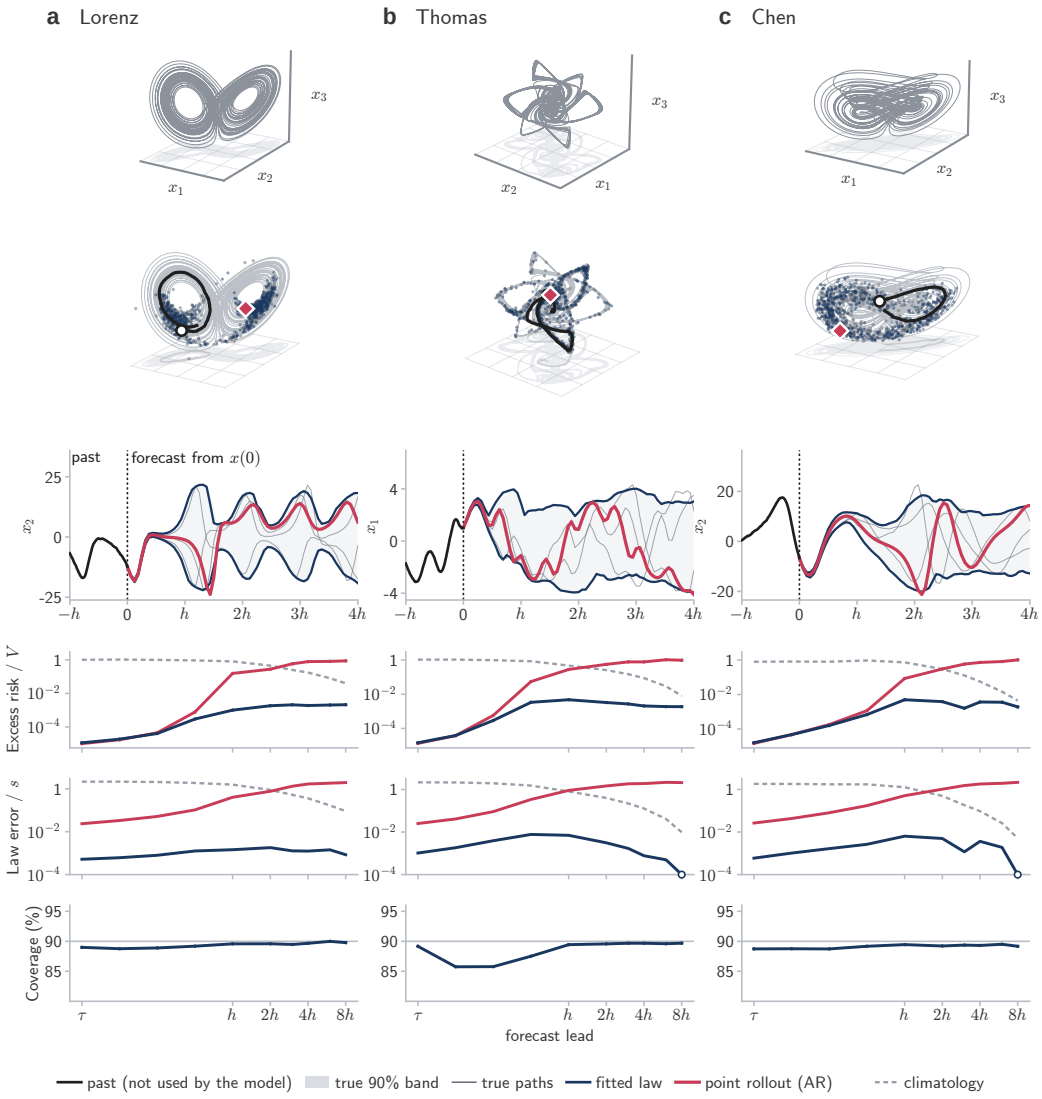

Figure 10: A point rollout and a law give different forecasts across three geometries. Exploratory results: one training seed and 64 test states per system. **First row,** noiseless trajectories for geometric orientation. **Second row,** observed past (black, unused by the predictor), reference particles (gray), and fitted particles (navy) at $2h$; the open circle marks the forecast origin and the crimson diamond marks AR. **Third row,** reference marginal 90% bands (gray), fitted marginal 90% bounds (navy), and reference sample paths through $4h$, with AR in crimson. Each example uses 2,048 reference and 512 fitted particles. **Bottom three rows,** mean excess risk divided by V, energy discrepancy divided by s, and marginal coverage, over all test states. The training mean is the point climatology baseline; the distributional baseline uses a state cloud. At $2h$, the average excess-risk ratios AR/law are 151, 170 and 79. Open circles in the energy row mark signed estimates below the 10^{-4} display floor; they are not positive measured errors. The protocol, normalization and recorded projection event are described in Section L.