Structural Blindness in Latent Data Assimilation: When Representation Geometry Misleads Sensor Design

Yongquan Qu[*]
Columbia University
yq2340@columbia.edu

Pierre Gentine
Columbia University
pg2328@columbia.edu

Abstract

Latent data assimilation estimates high-dimensional states by filtering in a learned low-dimensional representation. Accurate latent prediction and state reconstruction leave open how well the representation supports the assimilation of new observations. We study this question by fixing the representation and varying which observations the filter receives. In an exact linear–Gaussian benchmark, the averaged observation Fisher matrix retains the information needed to compare steady-state physical filtering errors across sensors. With a nonlinear decoder, two sensors can have identical averaged Fisher matrices and therefore tie under every score based only on that matrix, even when one yields a lower latent mean-squared error for the extended Kalman filter. Calibration trajectories let us examine the errors left by these updates and how the decoder maps them into physical space. Under moment and nondegeneracy conditions, we give a lower bound on the Pearson correlation between latent and physical mean-squared errors across sensors when their conversion rate varies little. On each of the four chaotic benchmarks with repeated independent tests, the best of five sensors selected using calibration errors had lower test RMSE on average than the best of five selected using posterior covariance scores. On two real-world datasets, assimilating validation-selected channels reduced error on the withheld channels relative to both a phase average and a decoder rollout without observation updates.

1 Introduction

Data assimilation (DA) combines model predictions with noisy observations to estimate the evolving state of a physical system [3]. To reduce the cost of these updates, researchers have developed ensemble Kalman filters [39] and particle filters [34] that operate in learned latent spaces. Variational methods can use learned dynamics as a prior [18] or alternate an observation update with a pretrained generative prior [44]. In atmospheric applications, learned representations have been used to combine model predictions with observations from different sources [16, 43]. A representation that reconstructs a state accurately can still pose a difficulty for assimilation. The filter often observes only a few components of the state, and it must use the differences between observed and predicted values to correct its estimates of the other components. How well does a representation trained for prediction and reconstruction support that correction?

We study this question by holding the learned dynamics and decoder fixed and varying the sensors. Each sensor configuration gives the filter a different view of the same physical system, and we can compare them by the error that remains in the full state after the update. The choice of what to observe is central to experimental design [11, 21, 40] and has motivated methods for placing sensors

[*]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

in assimilation systems [24, 56]. Reduced models have been used to select sensors for a Koopman observer [19] and to choose observation sites by applying pivoted QR to a reconstruction basis [38].

One way to compare sensors before running a filter is to ask which changes in the latent state each sensor can detect. At a given state, the local Fisher matrix measures how strongly the predicted observation responds to latent perturbations relative to the observation noise. Averaging this matrix over a common set of reference states gives a summary for each sensor.

Much of the work on averaged sensitivity asks which directions matter most. Wang and Ponce [53] chose latent coordinates for image inversion using a perceptual metric averaged over latent codes. A related construction for vector-valued functions averages Jacobian Gramians to identify influential input directions [57]. In Bayesian inverse problems, averaging local Fisher information over the prior identifies informative parameter directions before observations are collected [13]. Such averages can also guide a filter update or the choice of observations. Le Provost et al. [27] used averaged observation Jacobian Gramians to choose subspaces for an ensemble Kalman update, and Chen et al. [12] used averaged sensitivities to select observation channels by maximizing a lower bound on expected information gain. For sensor selection, we ask what the averaged observation Fisher matrix tells us about the error that remains after repeated filter updates.

In the exact linear–Gaussian model of Section 2, the observation sensitivity is constant. With the representation, dynamics, and process noise fixed, the averaged Fisher matrix determines the steady-state Kalman covariance and hence the minimum attainable physical mean-squared error. A nonlinear decoder allows the same sensor to have different sensitivities at different states. We construct two sensors with the same averaged Fisher matrix but different sensitivities at the forecast state of the extended Kalman filter (EKF). At that state, one sensor responds to latent perturbations and the other has zero slope. Any score based only on the averaged matrix assigns the two sensors the same value, even though one yields a lower stationary latent mean-squared error (Proposition 3.3).

On a calibration trajectory with known states, we can measure the latent errors that each sensor leaves after the filter update. How do these errors translate into errors in the physical state? The decoder stretches some latent directions more than others, as described by its local metric [4, 5]. Both calibration scores use the same post-update errors. One averages their squared lengths in latent space to obtain the latent MSE. The other weights the errors by direction using the mean decoder metric over a common set of reference states. Under the moment and nondegeneracy assumptions of Theorem 4.2, latent MSE and physical MSE have a Pearson correlation close to one across sensors when the rate of conversion between them varies little.

Across five chaotic benchmarks, we constructed sensor pairs with nearly equal averaged Fisher matrices. On the two-dimensional Navier–Stokes benchmark (NS2D), these small changes in the matrix were accompanied by substantially larger relative changes in the EKF's time-averaged spatial RMSE (Table 1). We also tested whether sensors selected on a calibration trajectory performed well on an independent trajectory. For each score, we compared the best physical RMSE among its five preferred sensors with the best in the full library. On all four systems with repeated tests, this gap was smaller on average for both calibration scores than for either posterior covariance score (§5.3).

The continuous positive airway pressure (CPAP) dataset [20] and the hydraulic-system dataset [22] contain simultaneous measurements from all channels. We can therefore evaluate how well the filter estimates the channels withheld from its updates. On CPAP, we selected the channel pair using the decoder-weighted calibration score on validation data. In every training run, that pair also had the lowest full-panel RMSE on held-out subjects. On Hydraulic, we selected the triple with the lowest measured full-panel RMSE on validation data for the withheld-channel evaluation. On both datasets, assimilating the selected channels gave lower test RMSE on the remaining channels than either a phase average or a decoder rollout without observation updates in every training run (§5.4).

2 Setup

The physical state $x_t \in \mathbb{R}^d$ evolves according to $x_{t+1} = \Phi(x_t) + w_t$, where w_t denotes process noise. We assume that the state process admits an invariant probability measure μ supported on a compact set $\mathcal{A}$. A sensor configuration is represented by a linear observation operator $H \in \mathbb{R}^{p\times d}$, with observations $y_t = Hx_t + \eta_t$, where $\eta_t \sim \mathcal{N}(0, R)$ and $R \succ 0$.

We encode the physical state as $z_t = E(x_t)$, using a fixed encoder $E : \mathbb{R}^d \to \mathbb{R}^k$ and decoder $D : \mathbb{R}^k \to \mathbb{R}^d$ trained in the deep Koopman framework [18, 29, 37]. The encoder maps the reference distribution μ to a latent distribution μ_z. The filter predicts the evolution of the latent state using $z_{t+1} = Mz_t + w_t^z$, with $w_t^z \sim \mathcal{N}(0, Q_z)$ and $Q_z \succ 0$, and updates the prediction $\hat{z}_{t|t-1}$ with y_t to obtain $\hat{z}_{t|t}$. We write $\hat{z}_t = \hat{z}_{t|t}$ and decode it as $\hat{x}_t = D(\hat{z}_t)$. The reconstruction error of the representation is measured by $\varepsilon_r^2 := \mathbb{E}_\mu \|x - D(E(x))\|^2$.

The sensor Gramian $H^\top H$ measures observation sensitivity in physical space: a perturbation δx produces a squared observation change $\|H\delta x\|^2 = \delta x^\top H^\top H \delta x$. Weighting this change by the inverse noise covariance gives the noise-weighted sensor Gramian $H^\top R^{-1} H$. In latent coordinates, the filter uses the observation map $z \mapsto HD(z)$, with Jacobian $HJ_D(z)$, where $J_D(z)$ is the decoder Jacobian. With Gaussian observation noise, the local Fisher information is therefore $J_D(z)^\top H^\top R^{-1} H J_D(z)$.

Definition 2.1 (Averaged observation Fisher matrix)**.** Averaging over μ_z gives

$$\bar{F}(H) := \mathbb{E}_{z\sim\mu_z}\left[J_D(z)^\top H^\top R^{-1} H J_D(z)\right].$$

Following the E-optimality criterion [40], we use the smallest eigenvalue of $\bar{F}(H)$ to define a sensitivity score for the weakest observed latent direction,

$$\omega_{\text{eff}}(H) := \sqrt{\lambda_{\min}(\bar{F}(H))\,\lambda_{\max}(R)}.$$

Under isotropic observation noise, $\omega_{\text{eff}}(H)$ is the smallest root-mean-square sensitivity of the predicted observation over unit latent directions.

With a linear decoder, the observation Jacobian is the same at every state, so averaging preserves the local Fisher information used in each update. To see what this implies for filtering error, we start with the exact linear-Gaussian embedding specified by hypothesis (LE) (App. B.1). Here the stationary state distribution is Gaussian, replacing the compact-support assumption above. The encoder and decoder are $E(x) = U^\top x$ and $D(z) = Uz$, with $U \in \mathbb{R}^{d\times k}$ and $U^\top U = I_k$. Under (LE), the dynamics matrix A is block diagonal with respect to the retained subspace and its orthogonal complement, so the retained coordinates evolve autonomously with $M = U^\top A U$. The sensors observe only the retained subspace, giving $Hx_t = HUz_t$. The projected model is

$$z_{t+1} = Mz_t + w_t^z, \qquad y_t = HUz_t + \eta_t.$$

The Kalman posterior covariance converges to $P^\star$, which solves the discrete-time algebraic Riccati equation (DARE) [3]

$$P^\star = \left[\left(Q_z + MP^\star M^\top\right)^{-1} + \bar{F}(H)\right]^{-1}. \tag{1}$$

For fixed latent dynamics and process noise, the representation U and sensor configuration H determine $P^\star$ through $\bar{F}(H) = U^\top H^\top R^{-1} HU$.

Theorem 2.2 (Linear-Gaussian filtering bound; App. B)**.** *Under* (LE)*, every square-integrable estimator $\hat{z}_t$ based on the observations available through time t satisfies*

$$\mathrm{FE}(\hat{z}) \;\geq\; \varepsilon_r^2 \;+\; \mathrm{tr}(P^\star), \tag{2}$$

where $\mathrm{FE}(\hat{z}) := \limsup_T T^{-1} \sum_{t=1}^T \mathbb{E}_\mu \|x_t - D(\hat{z}_t)\|^2$.

The two terms in the bound follow from the orthogonal decomposition

$$\|x_t - U\hat{z}_t\|^2 = \|(I - UU^\top)x_t\|^2 + \|z_t - \hat{z}_t\|^2.$$

The reconstruction term has expectation ε_r^2, and the Kalman filter attains the minimum latent mean-squared error $\mathrm{tr}P^\star$ in the steady-state limit (Proposition B.2). For a fixed representation, the reconstruction error is the same for every sensor configuration. Ranking sensors by $\mathrm{tr}P^\star$ therefore ranks their best attainable mean-squared error in physical space. In this linear-Gaussian setting, $\bar{F}(H)$ contains all the sensor-dependent information needed for this comparison. Appendix B.5 discusses lower bounds for nonlinear decoders.

3 Fisher–Gram impossibility

With a nonlinear decoder, observation sensitivity can vary across latent states. To study what averaging retains about an observation operator, we express $\bar{F}(H)$ as a linear function of the noise-weighted sensor Gramian $H^\top R^{-1} H$. Let $\mathrm{Sym}(d)$ denote the space of real symmetric $d \times d$ matrices. For a fixed decoder and the latent distribution μ_z shared by all candidate sensors, define the linear map

$$\pi_\mu : \mathrm{Sym}(d) \to \mathrm{Sym}(k), \qquad \pi_\mu(B) := \mathbb{E}_{z\sim\mu_z}[J_D(z)^\top B\, J_D(z)]. \tag{3}$$

For an observation operator H with constant noise covariance $R \succ 0$, $\bar{F}(H) = \pi_\mu(H^\top R^{-1} H)$. Under isotropic noise $R = \sigma^2 I_p$, this becomes (Proposition D.1)

$$\bar{F}(H) \;=\; \sigma^{-2}\, \pi_\mu(H^\top H). \tag{4}$$

With M and Q_z fixed, a sensor score can use the averaged Fisher matrix directly or use it to compute a steady-state covariance.

Definition 3.1 (Averaged-Fisher diagnostic class)**.** Write $F = \bar{F}(H)$ and let $\mathcal{V}$ be $\mathrm{Sym}_+(k)$ or $\mathrm{Sym}_{++}(k)$ according to the score's domain. The class $\mathcal{D}$ consists of scores of either form

$$s(H) = f(F) \qquad \text{or} \qquad s(H) = \varphi(F, P^\star(F)),$$

where $f : \mathcal{V} \to \mathbb{R}$ and $\varphi : \mathcal{V} \times \mathrm{Sym}_{++}(k) \to \mathbb{R}$ are continuous. Both forms require $F \in \mathcal{V}$. The second also requires a stabilizing solution $P^\star(F)$ of (1).

For positive definite F, examples include the A- and D-optimality criteria $\mathrm{tr}(F^{-1})$ and $\log\det F$ [40]. The E-optimality criterion $\lambda_{\min}(F)$, the sensitivity score ω_{eff}, and $\log\det(F + \lambda I)$ with $\lambda > 0$ are defined on the entire positive semidefinite cone. The posterior trace $\mathrm{tr}P^\star(F)$ is defined wherever the stabilizing solution exists.

Theorem 3.2 (Fisher–Gram impossibility)**.** *Assume a fixed observation-noise covariance $R \succ 0$ shared by all candidate observation operators, a latent distribution μ_z independent of H, and $\mathbb{E}_{\mu_z}\|J_D(z)\|_F^2 < \infty$. If two observation operators $H_1, H_2 \in \mathbb{R}^{p\times d}$ satisfy*

$$H_1^\top R^{-1} H_1 - H_2^\top R^{-1} H_2 \in \ker \pi_\mu,$$

then $s(H_1) = s(H_2)$ for every $s \in \mathcal{D}$ defined at both operators.

Proof. The identity $\bar{F}(H) = \pi_\mu(H^\top R^{-1} H)$ and linearity of π_μ give $\bar{F}(H_1) = \bar{F}(H_2)$, hence $s(H_1) = s(H_2)$ for every defined $s \in \mathcal{D}$. □

Sensors with the same averaged Fisher matrix form a *Fisher fibre*. Can their filtering errors differ? In the example below, the two decoder coordinates have equal average squared derivatives under a symmetric latent distribution. At zero, where the EKF predicts the state at every step, only the first coordinate has a nonzero derivative. Rotating the sensor therefore changes the sensitivity used by the filter while preserving the reference average.

Proposition 3.3 (EKF error along a Fisher fibre)**.** *Let Z_t be i.i.d. with density $\rho_a(z) \propto (1 - z^2/a^2)^4 \mathbf{1}_{\{|z|<a\}}$, where $a > 0$, and variance $Q = \mathbb{E}Z_t^2 = a^2/11$. Set $m = \mathbb{E}\sin^2 Z_t > 0$ and $v = \mathbb{E}(\cos Z_t - 1)^2 > 0$, and define*

$$D(z) = (z, \cos z/\sqrt{m})^\top, \qquad H_\theta = (\cos\theta, \sin\theta).$$

The observations are $y_t = H_\theta D(Z_t) + \eta_t$, where the noise variables are i.i.d. $\mathcal{N}(0, r)$, $r > 0$, and independent of the state sequence. The EKF uses a Gaussian transition model with $M = 0$ and variance Q, matching the variance of Z_t. Averaging with respect to ρ_a gives $\bar{F}(H_\theta) = r^{-1}$ for every θ. The stationary mean-squared error of the latent EKF estimate is

$$\mathrm{MSE}_z(H_\theta) = \frac{Qr}{r + Q\cos^2\theta} + \frac{Q^2\cos^2\theta\sin^2\theta\, v}{m(r + Q\cos^2\theta)^2}. \tag{5}$$

Both terms increase strictly on $0 < \theta \le \pi/4$, so the estimation error increases along a Fisher fibre on which every defined score $s \in \mathcal{D}$ is constant.

The first term in (5) is the posterior variance the EKF computes from the observation slope at its forecast, and the second is the additional MSE due to the nonlinear part of the innovation. For $Q = r = 1$, rotating the sensor from $\theta = 0$ to $\theta = \pi/4$ raises the latent MSE from $1/2$ to more than $2/3$, while the averaged Fisher matrix stays fixed. Appendix D.1 derives the error formula from the EKF update.

When $d > k$, the linear map π_μ sends $\mathrm{Sym}(d)$ into the lower-dimensional space $\mathrm{Sym}(k)$, so its kernel is nontrivial. Within a fibre, local observation sensitivities can differ across latent states, with the resulting information differences averaging to zero under μ_z. Appendix D.2 examines how the local Fisher matrix varies across reference states.

Can the information at the filter predictions change as the sensor varies along a Fisher fibre? Take $R = \sigma^2 I_p$. The distribution of the predicted state $\hat{z}_{t|t-1}$ can itself depend on the observation operator. When the predictions have a stationary law, we denote it by ν_B, where $B = H^\top H$, and define

$$\pi_\nu(B) := \mathbb{E}_{u\sim\nu_B}[J_D(u)^\top B J_D(u)].$$

Using this prediction average in (1) gives the Riccati equation for $P^\star_\nu(B)$,

$$P^\star_\nu(B) = \left[\left(Q_z + M P^\star_\nu(B) M^\top\right)^{-1} + \sigma^{-2}\pi_\nu(B)\right]^{-1}. \tag{6}$$

Under the hypotheses of Lemma D.5, the Riccati equation based on prediction-averaged information has a unique positive-definite solution $P^\star_\nu(B)$.

Fix a rank-p operator $H_0 \in \mathbb{R}^{p\times d}$, with $p \le d$, and set $B_0 := H_0^\top H_0$. The rank-p sensor Gramians form the manifold $\mathcal{M}_p := \{B \in \mathrm{Sym}_+(d) : \mathrm{rank}(B) = p\}$. Its tangent space at B_0, denoted $T_{B_0}\mathcal{M}_p$, describes the first-order variations of a rank-p Gramian. Directions in $T_{B_0}\mathcal{M}_p \cap \ker \pi_\mu$ preserve the averaged Fisher matrix to first order. Applying the rank–nullity theorem to π_μ restricted to this tangent space gives (Proposition D.2)

$$\dim(T_{B_0}\mathcal{M}_p \cap \ker \pi_\mu) \ge \max\{0, p(2d-p+1)/2 - k(k+1)/2\}.$$

Let $\mathcal{U}_0$ be a neighborhood of B_0 in $\mathcal{M}_p$, and write $g(J) := (Q_z + MJ^{-1}M^\top)^{-1}$ for the Riccati prediction operator. We assume:

- **(N1)** For each $B \in \mathcal{U}_0$, the EKF predictions have a stationary law ν_B. The resulting map π_ν is continuously differentiable and satisfies $\pi_\nu(B) \succeq c_\nu I_k$ for a constant $c_\nu > 0$. Writing $J^\star_\nu(B) = (P^\star_\nu(B))^{-1}$, we assume that $\mathrm{Id} - \mathrm{d}g(J^\star_\nu(B))$ is invertible on $\mathrm{Sym}(k)$, with its inverse uniformly bounded in operator norm over $\mathcal{U}_0$ (see the full conditions and sufficient criterion in Appendix D.2).
- **(N2)** There is a direction $\Delta \in T_{B_0}\mathcal{M}_p \cap \ker \pi_\mu$ such that $\pi'_\nu(\Delta) \neq 0$, where π'_ν denotes the derivative of π_ν at B_0.
- **(N3)** The restriction of π_μ to $\mathcal{M}_p$ has constant rank near B_0.

Under (N1)–(N2), $P^\star_\nu$ changes to first order in a direction that preserves the reference average to first order. Under (N3), this direction is tangent to a curve of rank-p Gramians with the same averaged Fisher matrix.

We also assume that the decoder satisfies the distance bounds

$$\ell_D\|z - z'\| \le \|D(z) - D(z')\| \le \ell_U\|z - z'\|, \qquad z, z' \in \Omega, \tag{7}$$

where $0 < \ell_D \le \ell_U < \infty$ and $\Omega \subseteq \mathbb{R}^k$ is a closed set containing $E(\mathcal{A})$. These bounds control how much decoding stretches or contracts latent distances.

Theorem 3.4 (Local separation of Riccati solutions under (N1)–(N3))**.** *Assume the hypotheses of Theorem 3.2, with $R = \sigma^2 I_p$ and $\rho(M) < 1$, and suppose the decoder satisfies* (7)*. Let $p \le d$ and let $H_0 \in \mathbb{R}^{p\times d}$ have rank p. Suppose* (N1)–(N3) *hold in a neighborhood $\mathcal{U}_0$ of $B_0 := H_0^\top H_0$ in $\mathcal{M}_p$. Then there exist $H_1, H_2 \in \mathbb{R}^{p\times d}$ with $\mathrm{rank}(H_i) = p$ and $H_i^\top H_i \in \mathcal{U}_0$ such that*

$$\bar{F}(H_1) = \bar{F}(H_2) \qquad \textit{but} \qquad P^\star_\nu(H_1^\top H_1) \neq P^\star_\nu(H_2^\top H_2) \quad \textit{in } \mathrm{Sym}_{++}(k). \tag{8}$$

The proof is given in Appendix D.

Relation to true filtering error. Suppose the joint state and filter process has a stationary law with finite error second moment. Write $P^\star_{\nu,H} := P^\star_\nu(H^\top H)$ and $\Sigma_H := \mathbb{E}[(z_t - \hat{z}_{t|t})(z_t - \hat{z}_{t|t})^\top]$. The discrepancy $C_H := \Sigma_H - P^\star_{\nu,H}$ measures the difference between the actual error second moment and the Riccati solution; it vanishes at the correctly specified linear-Gaussian steady state. For two observation operators, write $\Delta X = X_{H_1} - X_{H_2}$. Then

$$\Delta\Sigma = \Delta P^\star_\nu + \Delta C, \qquad \Delta\mathrm{MSE}_z = \mathrm{tr}(\Delta P^\star_\nu) + \mathrm{tr}(\Delta C). \tag{9}$$

In Proposition 3.3, the two terms of (5) are the scalar posterior variance $P^\star_\nu(H_\theta^\top H_\theta)$ and the correction C_{H_θ}. Both increase on $0 < \theta \le \pi/4$, so the actual latent MSE increases as well. Appendix D.1 also verifies (N1)–(N3) locally along this family.

Sensor comparisons. In the learned chaotic systems, we estimate π_μ from sampled states. A step in the kernel of this estimated map preserves the sample average, but projecting the Gramian back onto $\mathcal{M}_p$ can change it. Section 5.1 compares the relative change in $\bar{F}$ with the relative changes in EKF and EnKF errors.

4 From latent calibration error to physical accuracy

Through the observation map HD, the decoder determines how a latent perturbation affects the predicted observation. After the update, the decoder maps the remaining latent error into a displacement in the full physical state. On calibration trajectories with known states, we can compare latent and physical filtering errors directly and ask how closely they track each other across sensors. For sensor H, the long-time mean-squared error of the latent EKF estimate is

$$\mathrm{MSE}_z(H) \;:=\; \limsup_{T\to\infty} T^{-1} \sum_{t=1}^{T} \mathbb{E}\|z_t - \hat{z}^{\mathrm{EKF}}_{t|t}(H)\|^2.$$

Write $\mathrm{RMSE}^2(H)$ for the long-time mean-squared error of the decoded estimate. All sensors share the same mean-squared reconstruction error ε_r^2. For sensors with $\mathrm{MSE}_z(H) > 0$, we define the conversion rate

$$\alpha(H) := \frac{\mathrm{RMSE}^2(H) - \varepsilon_r^2}{\mathrm{MSE}_z(H)}.$$

If all sensors have the same positive rate, physical MSE is an affine function of latent MSE, and their Pearson correlation is one whenever latent MSE has finite, positive variance across sensors.

To see why the rate can vary, consider a small latent error u at z. The squared displacement after decoding is approximately $u^\top G_D(z)u$, where $G_D(z) := J_D(z)^\top J_D(z)$ is the local decoder metric. Thus the physical effect of a latent error depends on its direction and on the state where it occurs. We describe the mean and variation of the metric under μ_z by

$$\overline{G}_D := \mathbb{E}_{\mu_z}[G_D(z)], \qquad \delta_D^2 := \mathbb{E}_{\mu_z}\|G_D(z) - \overline{G}_D\|_F^2.$$

For each sensor H, assume that the physical state, latent EKF estimate, and updated covariance have a stationary joint law ν_H. Write $\Sigma_H = \mathbb{E}_{\nu_H}[\tilde{z}\tilde{z}^\top]$ for the second moment of the latent error $\tilde{z} = z - \hat{z}$. The mean metric weights this error by direction, giving $T_1(H) = \mathrm{tr}(\overline{G}_D\Sigma_H)$. Under the conditions below,

$$\mathrm{RMSE}^2(H) - \varepsilon_r^2 = T_1(H) + \mathcal{E}_{\mathrm{total}}(H).$$

The remainder accounts for metric variation across latent states, the Taylor remainder of the decoder, and the cross-term between the reconstruction residual and the decoded displacement (App. C, (27)).

Hypotheses for Lemma 4.1. Let Ω be a convex domain containing z_t and $\hat{z}_{t|t}(H)$ at every stationary time for each admissible sensor. On Ω, assume that D is C^2, satisfies the upper distance bound in (7), and has $\|\nabla^2 D\|_{\mathrm{op}} \le \eta$. Assume $\mathbb{E}_{\mu_z}\|G_D(z)\|_F^2 < \infty$. To control the cross-term, assume

$$|\langle x - D(E(x)), D(E(x)) - D(z')\rangle| \le \varepsilon_\perp \|x - D(E(x))\|\,\|D(E(x)) - D(z')\|, \tag{10}$$

for every $x \in \mathcal{A}$ and $z' \in \Omega$, with a common $\varepsilon_\perp \in [0, 1)$.

For every sensor, assume that ν_H has physical marginal μ and encoded marginal μ_z, with finite reconstruction MSE ε_r^2. From each admissible initial law, assume that the time averages of the

expected latent error outer product and the expected physical squared error converge to their stationary values. Taking the trace in the first limit gives $\mathrm{MSE}_z(H) = \mathrm{tr}\Sigma_H$ (Lemma C.7). For $M_j(H) = \mathbb{E}_{\nu_H}\|\tilde{z}\|^j$, assume $M_2(H) \geq m_\star^2 > 0$ and $M_4(H) \leq C_4^\star < \infty$ uniformly in H. Then $M_4(H)/M_2(H)^2 \leq \kappa_4^\star := C_4^\star/m_\star^4$. Appendix A.3 describes the evidence for these assumptions in the experiments.

Lemma 4.1 (A sufficient smoothness bound for $\mathrm{Var}_H(\alpha)$)**.** *Let $H \sim \mathcal{P}$ on the admissible sensor family $\mathcal{H}$, and let Var_H denote variance under $\mathcal{P}$. Under the hypotheses above, $\mathrm{MSE}_z(H) > 0$. Dividing the physical-error decomposition by $\mathrm{MSE}_z(H)$ gives*

$$\beta(H) := \frac{T_1(H)}{\mathrm{MSE}_z(H)} = \frac{\mathrm{tr}(\overline{G}_D\Sigma_H)}{\mathrm{tr}\Sigma_H},$$
$$\alpha(H) = \beta(H) + \frac{\mathcal{E}_{\mathrm{total}}(H)}{\mathrm{MSE}_z(H)}.$$

Define the alignment variance $A(\overline{G}_D, \{\Sigma_H\}) := \mathrm{Var}_H(\beta(H))$. *Then*

$$\begin{aligned}\mathrm{Var}_H(\alpha) \leq\ & 6\kappa_4^\star\delta_D^2 + 6\ell_U^2\kappa_4^\star\eta^2\,\mathbb{E}_{\mathcal{P}}[\mathrm{MSE}_z(H)] \\ & + 24\ell_U^2\varepsilon_\perp^2\varepsilon_r^2/m_\star^2 + 2A(\overline{G}_D, \{\Sigma_H\}).\end{aligned} \tag{11}$$

Appendix C derives the three error terms from the squared displacement and the reconstruction residual.

In (11), δ_D measures how the metric varies across states, while η and $\varepsilon_\perp$ control the Taylor remainder and the reconstruction–filtering cross-term. The alignment variance A captures how the directions of the filtering errors vary across sensors relative to the mean metric. For an affine decoder with exact reconstruction, $\mathcal{E}_{\mathrm{total}} = 0$ and $\mathrm{Var}_H(\alpha) = A$. Even in this case, sensors with equal latent MSE can have different physical errors if the decoder stretches their error directions by different amounts. Appendix C.9 gives bounds on A.

How close is the Pearson correlation between latent and physical MSE to one when the conversion rate varies across sensors?

Theorem 4.2 (Variance-controlled Pearson bound)**.** *Let $X(H), Y(H)$ satisfy $Y = \varepsilon_r^2 + \alpha(H)X$ for constant $\varepsilon_r^2 \geq 0$ and sensor-indexed $\alpha(H)$, and let $\mathcal{P}$ be a probability measure on the sensor space. Assume:* (i) $\bar{\alpha} := \mathbb{E}_{\mathcal{P}}[\alpha] \geq c_\alpha > 0$, (ii) $\mathrm{Var}_{\mathcal{P}}(X) \geq v_X > 0$, (iii) $\mathbb{E}_{\mathcal{P}}[X^4] \leq M_{X,4} < \infty$, *and* (iv) $K_\xi := \mathbb{E}_{\mathcal{P}}[(\alpha - \bar{\alpha})^4]/\mathrm{Var}_{\mathcal{P}}(\alpha)^2 \leq K_\xi^\star$ *when* $\mathrm{Var}_{\mathcal{P}}(\alpha) > 0$ *(with $K_\xi := 1$ otherwise). Set $\sigma_\alpha := \sqrt{\mathrm{Var}_{\mathcal{P}}(\alpha)}$ and*

$$a \ :=\ 2\,K_\xi^{\star\,1/4}\,M_{X,4}^{1/4}/(c_\alpha\sqrt{v_X}), \qquad \varepsilon_0 \ :=\ 1/a^2. \tag{12}$$

If $\mathrm{Var}_{\mathcal{P}}(\alpha) \leq \varepsilon \leq \varepsilon_0$, then $\rho_{\mathrm{Pearson}}(Y, X) \geq 0$ and

$$\left|1 - \rho_{\mathrm{Pearson}}(Y, X)\right| \ \leq\ a^2\,\varepsilon. \tag{13}$$

Proof in App. C.5.

Corollary 4.3 (Pearson correlation between latent and physical MSE)**.** *Suppose the hypotheses of Lemma 4.1 hold and the sensor distribution satisfies conditions* (i)–(iv) *of Theorem 4.2 for $X(H) = \mathrm{MSE}_z(H)$. Applying Theorem 4.2 with $Y := \mathrm{RMSE}^2$ and the variance bound from Lemma 4.1 gives $|1 - \rho_{\mathrm{Pearson}}(\mathrm{RMSE}^2, \mathrm{MSE}_z)| \leq a^2 B_{\mathrm{sm}}(\delta_D, \eta, \varepsilon_\perp, A)$ whenever $B_{\mathrm{sm}} \leq \varepsilon_0$, where B_{sm} is the right-hand side of* (11)*. Appendix C.5 gives the proof, and App. C.8 discusses the moment conditions.*

Remark 4.4 (Rank statistics and top-K regret). Two sensors with nearly equal physical MSE can switch order even when Pearson correlation is high. In a library of N sensors, let $Y(H) = \mathrm{RMSE}^2(H)$ and $Y_0(H) = \varepsilon_r^2 + \bar{\alpha}\,\mathrm{MSE}_z(H)$ with $\bar{\alpha} > 0$. Suppose $|Y(H) - Y_0(H)| \leq r$ for every sensor. For $1 \leq K < N$, if the $(K+1)$st smallest value of Y exceeds its minimum by more than $2r$, the K sensors with lowest latent MSE include every minimizer of Y (Proposition C.3). The factor of two comes from comparing two sensors: each approximation error can shift their score difference by up to r. Under this gap condition, sensors near the optimum can change order while every minimizer stays in the shortlist. Appendix C.6 also gives a sufficient gap for preserving the full ranking and a population Spearman bound under moment and density assumptions.

Table 1: Filtering errors for sensor pairs with nearly equal averaged Fisher matrices (SiLU MLP, $\sigma = 1$). For $\bar{F}$ and RMSE, we report median absolute relative changes. The blind fraction is the dimension of the subspace that preserves $\bar{F}$ to first order divided by the dimension of the Gramian tangent space. We report a lower bound. Sample sizes and intervals are in App. D.2.

System	(d, p, k)	blind $\geq$	$\Delta\bar{F}$	ΔRMSE EKF	ΔRMSE EnKF
L96	$(10, 5, 5)$	62.5%	0.96%	5.00%	3.49%
L96	$(40, 10, 10)$	84.5%	0.44%	1.46%	1.34%
KS	$(64, 16, 16)$	85.0%	0.95%	1.71%	1.82%
NS2D	$(256, 16, 16)$	96.6%	0.17%	17.98%	16.67%
β-BVE	$(256, 16, 16)$	96.6%	0.54%	3.05%	1.73%

On a calibration trajectory with known states, we can evaluate each sensor directly in physical space. We can also examine the directions of the latent errors after each update and use the decoder metric to estimate how much the decoder stretches errors in those directions. For sensor H, the errors $\tilde{z}_t(H) := z_t - \hat{z}_{t|t}(H)$ give the empirical second moment $\widehat{\Sigma}_H^{\mathrm{err}} := N^{-1} \sum_{t=1}^{N} \tilde{z}_t(H)\tilde{z}_t(H)^\top$. For a decoder weight W fixed across sensors, the calibrated score is $\widehat{T}_1^{\mathrm{CL}}(H) := \mathrm{tr}(W\widehat{\Sigma}_H^{\mathrm{err}})$. Taking $W = \widehat{\overline{G}}_D$, the sample mean of the physical decoder metric, gives a plug-in estimate of $T_1(H)$.

The linear-Gaussian model supplies a second score by replacing Σ_H with the steady-state posterior covariance $P^\star(\bar{F}(H))$, giving

$$T_1^{\mathrm{LG}}(H) := \mathrm{tr}(\overline{G}_D P^\star(\bar{F}(H))).$$

With the dynamics, decoder and reference law fixed, this score depends on H through $\bar{F}(H)$ and belongs to $\mathcal{D}$ wherever the stabilizing Riccati solution exists (App. E.5).

5 Experiments

We compare filtering errors on test trajectories for sensor pairs with nearly equal averaged Fisher matrices. We then examine how closely latent errors track physical errors across sensors and how well sensors selected on one trajectory perform on another.

We study Lorenz–96 (L96, $d = 10$ and 40), Kuramoto–Sivashinsky (KS, $d = 64$), and two vorticity systems, 2D Navier–Stokes (NS2D) and β-plane barotropic vorticity (β-BVE), both with $d = 256$. We run an EKF with a learned Koopman autoencoder and evaluate each sensor against the full simulated state. We use $p = k$, the smallest observation budget at which $HJ_D(z)$ can have full column rank (App. D.2). Simulation and training details are in App. E.

5.1 Sensor comparisons across five chaotic systems

We perturb each sensor Gramian in the kernel of the estimated Fisher map and project the result back to rank p. For each pair, we compare the change in the averaged Fisher matrix with the change in spatial RMSE averaged over the test trajectory. On NS2D, the median absolute relative change was about 18% for EKF RMSE and 0.2% for the averaged Fisher matrix. Within each EKF pair, we use the same observation noise in sensor coordinates, so rotating the sensor rows can change the measured RMSE even at a fixed Gramian. The EnKF gave a similar result on the matched subset (Table 1).

We also searched the fixed sensor libraries used for selection. Across the 13 libraries, no pair had a relative distance between averaged Fisher matrices at or below our prespecified threshold of 1.2% (App. E.5). We therefore evaluate the scores within each library by how well they rank its sensors.

5.2 Decoder geometry and calibration quality

Within each trained model, we rank the same sensors twice on one trajectory, once by latent MSE and once by physical MSE. Their Spearman correlation ρ_H measures agreement between the rankings. To compare decoders, we measure variation in the decoder metric across states in standardized output coordinates using $\widehat{\delta}_{D,\mathrm{rel}}$ (App. E.3).

Table 2: Decoder geometry and sensor rankings for four activations on L96 $d = 10$ (252 sensors). The first three columns after activation summarize decoder geometry and conversion-rate variability. The last four give Spearman correlations with physical RMSE across sensors. Here $\mathrm{tr}P^-$ is the forecast covariance trace, and Trace-CV is the coefficient of variation of the local Fisher trace across states for each sensor.

Activation	$\widehat{\delta}_{D,\mathrm{rel}}$	$\widehat{\eta}^{(1)}_{\mathrm{mean}}$	$\widehat{\alpha}$ CV	ω_{eff}	$\mathrm{tr}P^-$	Trace-CV	$\mathbf{MSE_z}$
ReLU	0.666	0 (p.w. aff.)	0.163	-0.039	-0.043	-0.014	$\mathbf{+0.849}$
tanh	0.677	121	0.341	$+0.094$	-0.113	-0.035	$\mathbf{+0.851}$
ELU	0.601	101	0.135	-0.023	$+0.012$	-0.006	$\mathbf{+0.925}$
SiLU	0.602	102	0.130	-0.053	-0.172	$+0.194$	$\mathbf{+0.944}$

On L96 $d = 10$, sensors with lower measured latent MSE generally had lower physical error for all four activations. The three other diagnostics had weak correlations with physical error on the same sensors (Tab. 2). Agreement was closest for ELU and SiLU. Both tanh and SiLU satisfy the smoothness condition in Lemma 4.1, but SiLU had less variation in the decoder metric across states and in the conversion rate across sensors.

Across 108 SiLU MLPs evaluated on the same 252 sensors, greater metric variation was associated with weaker agreement between latent and physical error rankings (Spearman $\rho = -0.79$ between $\widehat{\delta}_{D,\mathrm{rel}}$ and ρ_H, Appendix E.3).

In a separate analysis pooling both L96 systems and KS, the association between metric variation and ranking agreement was inconclusive when rankings were based on time-averaged spatial RMSE (Appendix E.3).

5.3 Sensor rankings and top-K selection

Rank correlation measures agreement across the full sensor library on one trajectory. For sensor selection, we ask whether the five candidates preferred on a calibration trajectory include a sensor with low physical error on a new trajectory. For a shortlist of size K, we define regret as

$$\mathrm{Regret}_K(s) := \min_{H \in \mathrm{top}_K(s)} \mathrm{RMSE}(H) \; - \; \min_{H \in \mathcal{H}} \mathrm{RMSE}(H), \tag{14}$$

where $\mathrm{top}_K(s)$ contains the K candidates preferred by the calibration score s in the library $\mathcal{H}$. Both minima use physical RMSE on the same independent test trajectory.

We kept the shortlists fixed for the independent tests. On all four systems with repeated tests, both calibration scores gave lower mean test regret than either posterior covariance score (Table 3). KS has one model and one trajectory pair. For the decoder-weighted calibration score $\widehat{T}_1^{\mathrm{CL}}$, we compute the weight in standardized output coordinates. The posterior scores use decoder Jacobians in physical coordinates.

Table 3: Independent-test top-five regret in physical RMSE units. Model scores use posterior covariances. Entries give means (sample standard deviations) across 15 model–trajectory combinations per system, with 14 on L96 $d = 40$ and one on KS. Trajectory pairs within each group share a trained model. Random gives the expected regret of five sensors drawn uniformly without replacement.

System	MSE_z	$\widehat{T}_1^{\mathrm{CL}}$	$\mathrm{tr}P^\star$	T_1^{LG}	Random
L96 $d = 10$	0.041(0.067)	0.054(0.080)	0.145(0.128)	0.196(0.202)	0.297(0.037)
L96 $d = 40$	0.055(0.060)	0.036(0.064)	0.290(0.095)	0.309(0.109)	0.309(0.029)
NS2D	1.264(1.532)	1.264(1.532)	3.118(1.426)	3.118(1.426)	3.827(0.678)
β-BVE	0.0034(0.0028)	0.0024(0.0029)	0.0121(0.0072)	0.0121(0.0072)	0.0077(0.0018)
KS	0.000	0.000	0.013	0.000	0.024

On NS2D, mean regret for the calibration scores rose from 0.02–0.03 on the calibration trajectories (Table 14) to 1.26 for either score on independent tests. The physical-error rankings agreed only weakly across trajectories (mean Spearman correlation 0.17).

5.4 Sensor selection on real-world data

Both datasets record all channels simultaneously, so we can evaluate each validation ranking and its first choice by test RMSE across the full panel. RMSE on the withheld channels tells us how well the filter estimates them from the selected observations.

On CPAP [20], we used the calibrated score to choose a pair from four pressure and flow channels. In every training run, the pair ranked first on validation also had the lowest full-panel RMSE on held-out subjects. Assimilating that pair roughly halved RMSE on the two withheld channels relative to the better of two baselines fitted on training data, a phase average and a decoder rollout without observation updates.

On Hydraulic [22], we ranked triples drawn from six pressure channels. In one run, the validation score had a Spearman correlation of 0.75 with test full-panel RMSE across triples, yet its first choice had 26% higher test RMSE than the best triple. For the withheld-channel experiment, we chose the triple with the lowest measured full-panel RMSE on validation. On held-out conditions, assimilating that triple reduced RMSE on the other three pressure channels relative to both baselines in every training run (Table 17, App. E.7).

6 Discussion

A representation trained for forecasting also shapes how a filter uses new observations. In a latent EKF, the decoder enters twice. At the forecast state, its Jacobian determines how a small change in the latent state affects the predicted observation. Two sensors can have the same averaged Fisher matrix yet differ in their sensitivity at that state, as Proposition 3.3 shows. After the update, the decoder determines how the remaining latent error translates into physical error. In atmospheric assimilation, Fan et al. [16] found an approximately linear relation between typical latent increments and the corresponding changes in physical space. Even with a linear decoder, two sensors with the same latent MSE can have different physical errors if the decoder stretches their error directions by different amounts.

A score can select a good sensor even when it misestimates the size of the filtering error. In the linear experiment with misspecified dynamics, the five sensors preferred by the posterior covariance trace included a minimizer of stationary error in each system (App. E.2). For a shortlist that will be used on a new trajectory, the choice also depends on how sensor rankings vary across trajectories. On NS2D, the rankings by physical error agreed only weakly across trajectories. Even the best sensor on a calibration shortlist could have substantially higher test error than the best sensor in the library (§5.3).

The equal-Fisher construction and the calibration analysis address complementary questions. The construction shows that averaging observation sensitivity can remove distinctions that affect latent filtering error. The calibration analysis gives conditions under which latent and physical mean-squared errors remain strongly correlated across sensors. Together, these results motivate evaluating a learned representation through both its observation sensitivities and the errors that remain after assimilation.

This perspective broadens how we evaluate representations of physical systems. Forecast and reconstruction accuracy leave open how effectively partial observations can correct the estimated state. Comparing representations with similar forecast accuracy under the same observations and filtering method would help isolate this capability. A further direction is to train representations with observation updates in the loop and evaluate their ability to reduce error in unobserved variables. This would connect representation learning more directly to the purpose of data assimilation: using limited observations to improve estimates of the full physical system.

Acknowledgments and Disclosure of Funding

We acknowledge funding from NSF through the Learning the Earth with Artificial intelligence and Physics (LEAP) Science and Technology Center (STC) (Award #2019625). The authors declare no competing interests.

References

[1] Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(110):3743–3773, 2014.

[2] Alen Alexanderian. A brief note on the Bayesian D-optimality criterion. *arXiv preprint arXiv:2212.11466*, 2022.

[3] Brian D. O. Anderson and John B. Moore. *Optimal filtering*. Prentice-Hall, Englewood Cliffs, NJ, 1979.

[4] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent Space Oddity: On the Curvature of Deep Generative Models. In *International Conference on Learning Representations*, 2018.

[5] Georgios Arvanitidis, Søren Hauberg, and Bernhard Schölkopf. Geometrically enriched latent spaces. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 631–639. PMLR, 2021.

[6] Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, 68 (3):255–278, 2013.

[7] Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 573–582. PMLR, 2019.

[8] A Bessell, B Ristic, A Farina, X Wang, and MS Arulampalam. Error performance bounds for tracking a manoeuvring target. In *Proceedings of the 6th International Conference on Information Fusion*, pages 903–910, 2003.

[9] Adrian N. Bishop and Pierre Del Moral. On the stability of Kalman–Bucy diffusion processes. *SIAM Journal on Control and Optimization*, 55(6):4015–4047, 2017.

[10] Matthieu Blanke, Yongquan Qu, Sara Shamekh, and Pierre Gentine. Strictly Constrained Generative Modeling via Split Augmented Langevin Sampling. In *International Conference on Learning Representations*, 2026.

[11] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: a review. *Statistical Science*, 10(3):273–304, 1995.

[12] Qiao Chen, Élise Arnaud, Ricardo Baptista, and Olivier Zahm. Coupled Input-Output Dimension Reduction: Application to Goal-oriented Bayesian Experimental Design and Global Sensitivity Analysis. *SIAM Journal on Scientific Computing*, 47(5):A2403–A2430, 2025.

[13] Tiangang Cui and Olivier Zahm. Data-free likelihood-informed dimension reduction of Bayesian inverse problems. *Inverse Problems*, 37(4):045009, 2021.

[14] Pierre Del Moral and Julian Tugaut. On the stability and the uniform propagation of chaos properties of ensemble Kalman–Bucy filters. *Annals of Applied Probability*, 28(2):790–850, 2018.

[15] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

[16] Hang Fan, Lei Bai, Ben Fei, Yi Xiao, Kun Chen, Yubao Liu, Yongquan Qu, Fenghua Ling, and Pierre Gentine. Physically consistent global atmospheric data assimilation with machine learning in latent space. *Science Advances*, 12(1):eaea4248, 2026.

[17] Hang Fan, Yi Xiao, Yongquan Qu, Juan Nathaniel, Fenghua Ling, Ben Fei, Lei Bai, and Pierre Gentine. Learning more physically realistic dynamics in machine-learning based weather forecasting with latent-space constraints. *arXiv preprint arXiv:2510.04006*, 2026. Version 2.

[18] Anthony Frion, Lucas Drumetz, Mauro Dalla Mura, Guillaume Tochon, and Abdeldjalil Aissa El Bey. Neural koopman prior for data assimilation. *IEEE Transactions on Signal Processing*, 72:4191–4206, 2024.

[19] John Graff, Albert Medina, and Francis D. Lagor. Information-based sensor placement for data-driven estimation of unsteady flows. *AIAA Journal*, 61(11):4864–4878, 2023.

[20] Ella Guy, Jennifer Knopp, and Geoff Chase. CPAP pressure and flow data from a local trial of 30 adults at the university of canterbury. PhysioNet, 2022. Version 1.0.1.

[21] Eldad Haber, Lior Horesh, and Luis Tenorio. Numerical methods for experimental design of large-scale linear ill-posed inverse problems. *Inverse Problems*, 24(5):055012, 2008.

[22] Nikolai Helwig, Eliseo Pignanelli, and Andreas Schütze. Condition monitoring of hydraulic systems. UCI Machine Learning Repository, 2015. Dataset.

[23] Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with Jacobian regularization, 2019.

[24] Yasutaka Ikuta and Hiromu Seko. Data-driven sparse sensor selection for observing-network optimization and its impact on data assimilation. *Monthly Weather Review*, 154(5):875–889, 2026.

[25] Thomas Kailath, Ali H Sayed, and Babak Hassibi. *Linear estimation*. Prentice Hall, 2000.

[26] Yuri Kinoshita, Kenta Oono, Kenji Fukumizu, Yuichi Yoshida, and Shin-ichi Maeda. Controlling posterior collapse by an inverse lipschitz constraint on the decoder network. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17041–17060. PMLR, 2023.

[27] Mathieu Le Provost, Ricardo Baptista, Youssef Marzouk, and Jeff D. Eldredge. A low-rank ensemble Kalman filter for elliptic observations. *Proceedings of the Royal Society A*, 478 (2266):20220182, 2022.

[28] John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2nd edition, 2012.

[29] Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature communications*, 9(1):4950, 2018.

[30] Krithika Manohar, Bingni W. Brunton, J. Nathan Kutz, and Steven L. Brunton. Data-driven sparse sensor placement for reconstruction: demonstrating the benefits of exploiting known patterns. *IEEE Control Systems Magazine*, 38(3):63–86, 2018.

[31] Jerrold E Marsden and Thomas JR Hughes. *Mathematical foundations of elasticity*. Courier Corporation, 1994.

[32] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, 2nd edition, 2009.

[33] Hanieh Mousavi and Jeff D. Eldredge. Low-order flow reconstruction and uncertainty quantification in disturbed aerodynamics using sparse pressure measurements. *Journal of Fluid Mechanics*, 1013:A41, 2025. doi: 10.1017/jfm.2025.10253.

[34] Nikolaj T. Mücke, Sander M. Bohté, and Cornelis W. Oosterlee. The deep latent space particle filter for real-time data assimilation with uncertainty quantification. *Scientific Reports*, 14: 19447, 2024.

[35] Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. ChaosBench: A Multi-Channel, Physics-Based Benchmark for Subseasonal-to-Seasonal Climate Prediction. In *Advances in Neural Information Processing Systems*, volume 37, pages 43715–43729, 2024.

[36] Mikhail Osipov. On geometry regularization in autoencoder reduced-order models with latent neural ode dynamics, 2026.

[37] Samuel E. Otto and Clarence W. Rowley. Linearly recurrent autoencoder networks for learning dynamics. *SIAM Journal on Applied Dynamical Systems*, 18(1):558–593, 2019.

[38] Suraj Pawar and Omer San. Equation-free surrogate modeling of geophysical flows at the intersection of machine learning and data assimilation. *Journal of Advances in Modeling Earth Systems*, 14(11):e2022MS003170, 2022.

[39] Mathis Peyron, Anthony Fillion, Selime Gürol, Victor Marchais, Serge Gratton, Pierre Boudier, and Gael Goret. Latent space data assimilation by using deep learning. *Quarterly Journal of the Royal Meteorological Society*, 147(740):3759–3777, 2021.

[40] Friedrich Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics. SIAM, 2006.

[41] Yongquan Qu and Xiaoming Shi. Can a machine-learning-enabled numerical model help extend effective forecast range through consistently trained subgrid-scale models? *Artificial Intelligence for the Earth Systems*, 2(1):e220050, 2023.

[42] Yongquan Qu, Mohamed Aziz Bhouri, and Pierre Gentine. Joint Parameter and Parameterization Inference with Uncertainty Quantification through Differentiable Programming. In *ICLR Workshop on AI4Differential Equations in Science*, 2024.

[43] Yongquan Qu, Juan Nathaniel, Shuolin Li, and Pierre Gentine. Deep Generative Data Assimilation in Multimodal Setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

[44] Yongquan Qu, Matthieu Blanke, Sara Shamekh, and Pierre Gentine. PnP-DA: Towards Principled Plug-and-Play Integration of Variational Data Assimilation and Generative Models, 2025.

[45] Konrad Reif, Stefan Günther, Engin Yaz, and Rolf Unbehauen. Stochastic stability of the discrete-time extended Kalman filter. *IEEE Transactions on Automatic Control*, 44(4):714–728, 1999.

[46] Berthold Schweizer and Edward F Wolff. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4):879–885, 1981.

[47] Tyler H. Summers and John Lygeros. Optimal Sensor and Actuator Placement in Complex Dynamical Networks. *arXiv preprint arXiv:1306.2491*, 2013.

[48] Petr Tichavský, Carlos H. Muravchik, and Arye Nehorai. Posterior Cramér–Rao bounds for discrete-time nonlinear filtering. *IEEE Transactions on Signal Processing*, 46(5):1386–1396, 1998.

[49] Ramon Van Handel. The stability of conditional Markov processes and Markov chains in random environments. *Annals of Probability*, 37(5):1876–1925, 2009.

[50] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. Wiley, 1968.

[51] Bart Vandereycken, P.-A. Absil, and Stefan Vandewalle. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. In *SSP09*, pages 389–392, 2009.

[52] Peter Walters. *An introduction to ergodic theory*, volume 79 of *Graduate Texts in Mathematics*. Springer, 1982.

[53] Binxu Wang and Carlos R. Ponce. A Geometric Analysis of Deep Generative Image Models and Its Applications. In *International Conference on Learning Representations*, 2021.

[54] Ruigang Wang, Krishnamurthy D. J. Dvijotham, and Ian R. Manchester. Monotone, bi-lipschitz, and polyak-łojasiewicz networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 50379–50399. PMLR, 2024.

[55] Nick Whiteley. Dimension-free Wasserstein contraction of nonlinear filters. *Stochastic Processes and their Applications*, 135:31–50, 2021.

[56] Zhaoyue Xu, Shizhao Wang, Xin-Lei Zhang, and Guowei He. Optimal sensor placement for ensemble-based data assimilation using gradient-weighted class activation mapping. *Journal of Computational Physics*, 514:113224, 2024.

[57] Olivier Zahm, Paul G. Constantine, Clémentine Prieur, and Youssef M. Marzouk. Gradient-Based Dimension Reduction of Multivariate Vector-Valued Functions. *SIAM Journal on Scientific Computing*, 42(1):A534–A558, 2020.

[58] Xingyu Zhu, Yongquan Qu, and Xiaoming Shi. LEX v1.6.0: a new large-eddy simulation model in JAX with GPU acceleration and automatic differentiation. *Geoscientific Model Development*, 19:1103–1120, 2026.

[59] Xingyu Zhu, Yongquan Qu, Naigeng Wu, Hualong Zhang, and Xiaoming Shi. Data-driven subgrid-scale turbulence parameterization: methods for improving model efficiency. *Machine Learning: Earth*, 2(2):025002, 2026.

Appendix

A Related work and limitations

A.1 Extended related work

What information guides sensor selection? In the steady-state Kalman design of Graff et al. [19], the information contributed by a fixed sensor array has a simple form for a linear decoder. In our notation, $D(z) = Uz$ and the latent observation matrix is $H_y = HU$. For observation-noise covariance $R \succ 0$,

$$H_y^\top R^{-1} H_y = U^\top H^\top R^{-1} H U = \bar{F}(H).$$

The last equality holds because the observation Jacobian is constant across states. Graff et al. also use the dynamics in the steady-state design and individual observation rows when adding sensors to the array. Our result applies to scores whose dependence on the array enters entirely through $\bar{F}(H)$. In the nonlinear example of Proposition 3.3, two sensors have the same reference average and different stationary latent EKF mean-squared errors.

QR pivoting selects observation sites from a reconstruction basis [30], an approach Pawar and San [38] use for a latent ensemble Kalman filter. Xu et al. [56] train a surrogate on the initial ensemble and use Grad-CAM to identify regions to observe before assimilation. Ikuta and Seko [24] rank GNSS observing sites by their contribution to reconstructing a field of precipitable water vapor, then evaluate selected groups in assimilation experiments. In §5.3, we rank candidate sensors by the latent filtering errors observed on a calibration trajectory.

Mousavi and Eldredge [33] map pressure observations and sensor coordinates to a latent estimate. They average the Gramian of the network's input Jacobian over noisy inputs to identify directions along which that estimate is sensitive. For a candidate sensor H, our map sends the latent state z to the predicted observation $HD(z)$. Its noise-weighted Jacobian Gramian describes sensitivity to perturbations in latent space.

With a generative prior, the update can also incorporate physical constraints. For example, Blanke et al. [10] enforce constraints during sampling in diffusion-based assimilation. Here the update rule is fixed, so the comparison across sensors isolates the effect of changing the observations.

Filtering error and stability. The posterior Cramér–Rao bound gives a lower bound on mean-squared estimation error under a specified stochastic model [48, 50]. Its tightness concerns how closely an estimator can approach that bound [8]. Sensor selection requires a comparison across observation models. If two sensors have the same $\bar{F}(H)$, every score in the class defined in §3 gives them the same value wherever the score is defined.

Filter stability results study how the effect of initialization evolves when filters use a common observation model [9, 49, 55]. In Appendix D, we vary the observation operator and assume that the information averaged over stationary predictions depends smoothly on the sensor Gramian. Under the local theorem's assumptions, the resulting Riccati matrix can change along a family of sensors with a fixed reference average.

Decoder geometry and physical error. For a small latent error e at z, the pullback metric gives the local approximation $\|D(z+e) - D(z)\|^2 \approx e^\top J_D(z)^\top J_D(z) e$ [4, 5]. The sensor affects which error directions remain after the update. Two sensors with similar latent MSE can therefore produce different decoded displacements. To compare the decoded estimate with the true state x, §4 also accounts for the reconstruction residual $x - D(E(x))$ and its inner product with the decoded displacement.

Forecast models also use latent coordinates to account for errors across physical variables. For example, Fan et al. [17] train a weather model with a loss in a learned latent space to account for dependence across atmospheric variables. Our comparison asks how the decoder weights the error left by each sensor after assimilation.

Jacobian regularization penalizes local sensitivity at the points used for training [23]. An upper Lipschitz bound limits how much a decoder can stretch a displacement, while an inverse Lipschitz constraint limits contraction [26], and bi-Lipschitz architectures control both [7, 54]. The effect of these constraints depends on how the representation is used. In an advection–diffusion–reaction experiment, Osipov [36] found that encouraging near-isometry during pretraining could increase

long-horizon rollout error. For the residual decoder in Appendix E.4, the architecture gives global distance bounds in standardized output coordinates. At sampled states we also estimate how much of the reconstruction residual lies in the local decoder tangent space.

Forecast dynamics and evaluation. In barotropic vorticity models, learned subgrid parameterizations can be trained through the forecast solver [41] and evaluated for both accuracy and computational cost [59]. Differentiable solvers also support learning turbulence closures in large-eddy simulations [58] and jointly estimating physical parameters and learned closures [42]. At the scale of global climate forecasts, ChaosBench evaluates multichannel predictions using deterministic, probabilistic, and physics-based metrics [35]. Our experiments hold the forecast dynamics fixed and evaluate how the choice of observed channels affects the updated state.

A.2 Limitations

Proposition 3.3 gives two sensors with the same averaged Fisher matrix and different latent EKF mean-squared errors. The factorization in §3 uses a common reference distribution, a decoder Jacobian with finite second moment, and a fixed positive-definite observation-noise covariance. Theorem 3.4 concerns the Riccati solution obtained by averaging observation information at the filter's predicted states. Under its hypotheses, this solution can vary among sensors with the same reference average. In particular, N2 requires a sensor perturbation that preserves the reference average to first order while changing the prediction average to first order. We have not verified N2 on the five chaotic benchmarks. The fitted NS2D dynamics also fall outside the theorem's stability range, $\rho(M) < 1$.

Section 4 relates latent and physical MSE through the conversion rate. The bound on its variance assumes a stationary joint law for the physical state, the latent EKF estimate, and its updated covariance. It requires the time averages of the expected latent error second moment and the expected physical squared error to converge to their stationary values, with uniform bounds on latent error moments across sensors. The tanh and SiLU MLPs satisfy the decoder smoothness condition. The experiments also include ReLU and ELU. To obtain a Pearson correlation bound, Theorem 4.2 imposes additional moment and nondegeneracy conditions across sensors, including a positive lower bound on the mean conversion rate. For convergence along a sequence of models, this lower bound must hold throughout the sequence (Appendix C.8).

On both $d = 256$ systems, directions that preserve the averaged Fisher matrix to first order span at least 96.6% of the Gramian tangent space. Yet in the fixed libraries we searched, no pair met our threshold of 1.2% for relative Fisher distance (App. E.5). To evaluate a shortlist chosen on simulated data, we compare its lowest physical RMSE with the lowest in the full library on the same independent test trajectory (App. E.6). Full-state values are available in simulation, with one model and one trajectory pair in the KS test. On Hydraulic and CPAP, all channels were recorded simultaneously. We can therefore evaluate estimates of the withheld channels on held-out operating conditions and subjects, respectively (App. E.7).

A.3 Assumptions in the numerical experiments

Table 4 summarizes the evidence for the assumptions in the sensor-pair experiments with SiLU MLPs.

Trained dynamics. Across the fitted models, the spectral radius of M was just below one on β-BVE and just above one on NS2D. On L96 and KS, we also examined temporal decorrelation in the encoded trajectories as an empirical check on ergodicity.

Decoder geometry. The smallest singular value of the decoder Jacobian at a sampled state describes local sensitivity (App. E.4). The distance bounds in (7) apply to pairs of states throughout the domain. On L96 $d = 10$, the estimated relative projection of the reconstruction residual onto the decoder tangent space was about 0.5. This projection describes alignment at the encoded state. Equation (10) bounds the absolute inner product of the residual with $D(E(x)) - D(z')$ by $\varepsilon_\perp$ times the product of their lengths, for every $x \in \mathcal{A}$ and $z' \in \Omega$.

For the residual decoder in Appendix E.4, the architecture gives analytic distance bounds in standardized output coordinates. Across three seeds, the constants were $\ell_D = 0.78 \pm 0.13$ and $\ell_U = 31.9 \pm 5.4$, and the relative tangent projection was 0.013 ± 0.004. For the finite-weight

Table 4: Assumptions for the five systems. Yes and No give the result of the spectral-radius check. *Sampled* indicates a check on sampled states, sensors, or trajectories. *Assumed* marks a requirement left open by the available checks. *Analytic* indicates a result derived for the stated model. The asterisk marks the local invertibility result at a stabilizing Riccati solution.

Condition	L96 $d = 10$	L96 $d = 40$	KS	NS2D	β-BVE
$\rho(M) < 1$	Yes	Yes	Yes	No	Yes
Encoded-process ergodicity	Sampled	Sampled	Sampled	Assumed	Assumed
Standard decoder distance bounds	Sampled	Sampled	Sampled	Sampled	Sampled
Approximate orthogonality	Sampled	Assumed	Assumed	Assumed	Assumed
Uniform positivity of $\bar{F}(H)$	Sampled	Sampled	Sampled	Sampled	Sampled
Decoder smoothness, SiLU MLP	Analytic	Analytic	Analytic	Analytic	Analytic
Joint law and averaging (aggregate)	Sampled	Sampled	Sampled	Assumed	Assumed
Fourth-moment bound	Sampled	Sampled	Sampled	Sampled	Sampled
N1(a), stationary prediction law	Sampled	Sampled	Sampled	Sampled	Sampled
N1(b), smooth sensor response	Sampled	Assumed	Assumed	Assumed	Assumed
N1(c), uniform positivity of $\pi_\nu(B)$	Sampled	Sampled	Sampled	Sampled	Sampled
N1(d), local DARE invertibility	Analytic*	Analytic*	Analytic*	Analytic*	Analytic*
N2	Assumed	Assumed	Assumed	Assumed	Assumed
N3	Sampled	Sampled	Sampled	Assumed	Assumed

MLP, convolutional, and residual blocks with tanh or SiLU activations, smoothness follows from the constituent maps, and bounded activation derivatives give a global Lipschitz bound. For a Transformer, regularity depends on its blocks.

Error moments. For four activations on each of the two L96 systems and KS, we checked how far the filter estimates lay from the encoded states and compared latent-error moments over different evaluation windows (Remark C.8). All sampled estimates lay within the tubes used for this check. On NS2D and β-BVE we estimated error moments for four SiLU architectures. The support and time-averaging conditions remain assumptions on those systems. Across all five systems, the observed fourth-moment ratios describe the sampled trajectories. The theory requires a stationary fourth-moment bound that holds uniformly over sensors. Standard filter-stability results address the related question of sensitivity to the initial condition [9, 14, 45].

Response to sensor perturbations. On L96 $d = 10$, we reran the filter after perturbing the sensor and recomputed the information average over its predicted states. Both the sensor and the prediction distribution therefore contribute to the measured change. The KS calculation keeps the prediction distribution fixed, so N1(b) for the full information average remains an assumption. To verify N2, one would need a direction $\Delta \in T_{B_0}\mathcal{M}_p \cap \ker \pi_\mu$ along which the derivative of π_ν is nonzero. For N3, we checked ranks at sampled points on L96 and KS (App. D.2). Constant rank near the reference sensor remains an assumption on NS2D and β-BVE.

Local invertibility in N1(d). At a positive-definite stabilizing posterior DARE solution P, set $X = Q_z + MPM^\top$, $S = \sigma^{-2}\pi_\nu(B)$, and $L = PX^{-1}M$. Then $\rho(L) < 1$. Differentiating $P^{-1} = X^{-1} + S$ gives

$$\delta P - L\delta P L^\top = -P\,\delta S\,P, \qquad \delta P = -\sum_{j=0}^{\infty} L^j P\,\delta S\,P(L^\top)^j.$$

Since $\rho(L) < 1$, the series converges and the Stein operator is invertible. In the coordinates $J = P^{-1}$, this gives invertibility of $\mathrm{Id} - \mathrm{d}g(J)$ at the reference sensor. Under N1(b), continuity and the implicit function theorem give a uniformly bounded inverse on a sufficiently small neighborhood.

B The filtering bound for a linear–Gaussian embedding

Under the exact linear–Gaussian embedding, Theorem 2.2 gives a lower bound on the long-run physical MSE of any square-integrable causal latent estimate. The projected Kalman filter attains the bound $\varepsilon_r^2 + \mathrm{tr}P^\star$. Projection onto the fixed subspace leaves a reconstruction error of ε_r^2, and estimating the retained coordinates contributes $\mathrm{tr}P^\star$ in the steady-state limit.

B.1 The projected model and its stationary prior

Hypothesis (LE). Let $U \in \mathrm{St}(d,k)$ have orthonormal columns, let $\Pi = UU^\top$, and let the columns of $U_\perp$ form an orthonormal basis for the complementary subspace.

(LE.i) The dynamics are linear, $\Phi(x) = Ax$, with $\rho(A) < 1$. The range of U reduces A, so A is block diagonal in the basis $[U, U_\perp]$.

(LE.ii) The encoder and decoder are $E(x) = U^\top x$ and $D(z) = Uz$.

(LE.iii) The observation operator satisfies $H = H\Pi$.

(LE.iv) The process noise is i.i.d. $\mathcal{N}(0, Q_x)$, where $Q_x \succeq 0$, $U^\top Q_x U = Q_z \succ 0$, and $U_\perp^\top Q_x U = 0$. The observation noise is i.i.d. $\mathcal{N}(0, R)$ with $R \succ 0$. The two noise sequences and the initial state are mutually independent.

We evaluate error under the stationary law by drawing $x_0 \sim \mu = \mathcal{N}(0, \Sigma_x)$. Since $\rho(A) < 1$, this law is unique and $\Sigma_x = A\Sigma_x A^\top + Q_x$ [25, §2.2]. The projected prior is therefore $z_0 \sim \mathcal{N}(0, U^\top \Sigma_x U)$.

Proposition B.1 (Linear-Gaussian latent reduction)**.** *Under* (LE) *and the stationary prior, $z_t = U^\top x_t$ follows a linear–Gaussian model with dynamics $M = U^\top AU$, observation matrix $H_z = HU$, and noise covariances Q_z and R. Its averaged observation Fisher matrix is $\bar{F} = H_z^\top R^{-1} H_z$. The Kalman estimate equals $\mathbb{E}[z_t \mid y_{1:t}]$, and its posterior covariance converges to the unique stabilizing solution $P^\star$ of* (1)*. The proof is in App. B.2.*

B.2 The projected Kalman filter

Write $z_t^\perp = U_\perp^\top x_t$ and $A_\perp = U_\perp^\top AU_\perp$. Projecting the dynamics onto the two subspaces and using the block structure in (LE.i) gives

$$z_{t+1} = Mz_t + U^\top w_t, \qquad z_{t+1}^\perp = A_\perp z_t^\perp + U_\perp^\top w_t. \tag{15}$$

The projected noises are jointly Gaussian with zero cross-covariance by (LE.iv), hence independent. The latent noise has covariance Q_z. By (LE.iii), $y_t = H\Pi x_t + \eta_t = H_z z_t + \eta_t$, so both the latent dynamics and the observations depend only on the retained state and their respective noises. Substituting $J_D = U$ into Definition 2.1 gives $\bar{F} = U^\top H^\top R^{-1} HU = H_z^\top R^{-1} H_z$.

With the projected Gaussian prior from App. B.1, the Kalman filter computes the conditional mean $m_t = \mathbb{E}[z_t \mid y_{1:t}]$ and its deterministic posterior covariance $P_{t|t}$.

Since $Q_z \succ 0$, the latent dynamics are stabilizable. All modes of M are stable by (LE.i), so (M, H_z) is detectable for every sensor. Under these conditions the Riccati recursion converges from any positive semidefinite initial covariance to the unique stabilizing solution $P^\star \succ 0$ of (1) [25, Thm. 14.7.1].

B.3 Proof of Theorem 2.2

Let $m_t = \mathbb{E}[z_t \mid y_{1:t}]$. The reconstruction residual $(I - \Pi)x_t$ is orthogonal to the range of U, and U preserves lengths. Thus the physical MSE is $\varepsilon_r^2 + \mathbb{E}_\mu \|z_t - \hat{z}_t\|^2$, where $\varepsilon_r^2 = \mathrm{tr}((I-\Pi)\Sigma_x(I-\Pi))$. For any square-integrable estimate based on $y_{1:t}$, the error $z_t - m_t$ is orthogonal in L^2 to $m_t - \hat{z}_t$. Since $\mathbb{E}_\mu \|z_t - m_t\|^2 = \mathrm{tr} P_{t|t}$, we obtain

$$\mathbb{E}_\mu \|x_t - U\hat{z}_t\|^2 = \varepsilon_r^2 + \mathrm{tr} P_{t|t} + \mathbb{E}_\mu \|m_t - \hat{z}_t\|^2.$$

The last term is nonnegative at every time. By Proposition B.1, $\mathrm{tr} P_{t|t}$ and its time average both converge to $\mathrm{tr} P^\star$. The limit inferior of the time-averaged physical MSE is therefore at least $\varepsilon_r^2 + \mathrm{tr} P^\star$, which also bounds $\mathrm{FE}(\hat{z})$ from below. For the Kalman estimate $\hat{z}_t = m_t$, the last term vanishes and the average converges to the bound.

B.4 Achievability for a closed linear projection

The same argument applies when the discarded coordinates depend on the retained ones. It is enough that the retained coordinates evolve autonomously and that the observations depend only on them.

Proposition B.2 (Latent Kalman achievability)**.** *Consider $x_{t+1} = Ax_t + w_t$ and $y_t = Hx_t + \eta_t$ with centered stationary Gaussian initial law $\mathcal{N}(0, \Sigma_x)$. The noise sequences are i.i.d. Gaussian with means zero and covariances Q_x and $R \succ 0$, respectively, and are independent of each other and of x_0. For $U \in \mathrm{St}(d, k)$ and $\Pi = UU^\top$, suppose*

$$U^\top A(I - \Pi) = 0, \qquad H(I - \Pi) = 0.$$

If the posterior covariance of the projected Kalman filter converges to P_∞, then

$$\mathrm{FE}_{\mathrm{Kalman}} = \varepsilon_U^2 + \mathrm{tr}P_\infty, \qquad \varepsilon_U^2 = \mathrm{tr}((I - \Pi)\Sigma_x(I - \Pi)).$$

Here ε_U^2 is the reconstruction error ε_r^2 for the projection $UU^\top$.

Proof. The two closure conditions give $z_{t+1} = (U^\top AU)z_t + U^\top w_t$ and $y_t = HUz_t + \eta_t$, with latent noise covariance $U^\top Q_x U$. With the projected prior, the latent MSE at time t is $\mathrm{tr}P_{t|t}$. By orthogonality, the physical MSE at that time is $\varepsilon_U^2 + \mathrm{tr}P_{t|t}$. Averaging and using $P_{t|t} \to P_\infty$ gives the stated equality. □

B.5 Lower bounds for nonlinear representations

For any causal estimate $\hat{z}_t$, write $z_t = E(x_t)$ and $e(x_t) = x_t - D(z_t)$. Squaring the sum of the reconstruction residual and the decoded displacement gives

$$\|x_t - D(\hat{z}_t)\|^2 = \|e(x_t)\|^2 + \|D(z_t) - D(\hat{z}_t)\|^2 + 2\langle e(x_t), D(z_t) - D(\hat{z}_t)\rangle. \tag{16}$$

Suppose $x_t \sim \mu$, the reconstruction error is finite, and $z_t, \hat{z}_t \in \Omega$ almost surely. Under (10), the cross-term is bounded below by $-2\varepsilon_\perp\|e(x_t)\|\,\|D(z_t) - D(\hat{z}_t)\|$. Applying $2ab \le a^2 + b^2$ gives

$$\|x_t - D(\hat{z}_t)\|^2 \ge (1 - \varepsilon_\perp)\big(\|e(x_t)\|^2 + \|D(z_t) - D(\hat{z}_t)\|^2\big). \tag{17}$$

The lower distance bound in (7) gives $\|D(z_t) - D(\hat{z}_t)\| \ge \ell_D\|z_t - \hat{z}_t\|$. Here $\mathbb{E}_\mu$ averages over the state and its observation history, with μ denoting the stationary law of the physical state. Taking expectations gives

$$\mathbb{E}_\mu\|x_t - D(\hat{z}_t)\|^2 \ge (1 - \varepsilon_\perp)\big(\varepsilon_r^2 + \ell_D^2\,\mathbb{E}_\mu\|z_t - \hat{z}_t\|^2\big). \tag{18}$$

Averaging (18) over time gives a bound on the long-run physical MSE. Taking the limsup on the left and the liminf on the right yields

$$\mathrm{FE}(\hat{z}) \ge (1 - \varepsilon_\perp)\,\varepsilon_r^2 + (1 - \varepsilon_\perp)\,\ell_D^2\,\liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_\mu\|z_t - \hat{z}_t\|^2. \tag{19}$$

For nonlinear representations, one can bound physical error using the full state model or bound latent error by treating the encoded state $E(x_t)$ as the quantity to be estimated.

(i) Full-state PCRB. Setting $\hat{x}_t := D(\hat{z}_t)$ and applying the Tichavský–Muravchik–Nehorai recursion [48] directly gives, under standard score regularity for the full-state Markov model,

$$\mathrm{FE}(\hat{z}) \;\ge\; \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathrm{tr}P^*_{\mathrm{full},t}, \tag{20}$$

where $P^*_{\mathrm{full},t}$ is the $d \times d$ full-state PCRB. The decoder can be arbitrary here. The observation contribution to the recursion is the physical-state Fisher matrix $H^\top R^{-1}H$.

(ii) Functional Van Trees. At a fixed time, let $p_t(x, y_{1:t})$ denote the joint density of the state and observation history, and define the Bayesian information matrix

$$\mathcal{J}_t := \mathbb{E}[s_t s_t^\top], \qquad s_t := \nabla_x \log p_t(x_t, y_{1:t}).$$

Under the Van Trees regularity and boundary conditions [50], with $\mathcal{J}_t \succ 0$,

$$\mathbb{E}\|E(x_t) - \hat{z}_t(y_{1:t})\|^2 \;\ge\; \mathrm{tr}\big(\mathbb{E}[J_E(x_t)]\mathcal{J}_t^{-1}\mathbb{E}[J_E(x_t)]^\top\big)\,.$$

This bound involves the full state–observation model and the averaged encoder Jacobian. Expressing it through $\bar{F}$ and the latent Riccati equation requires an additional relation between the physical and encoded models. For nonlinear decoders, Lemma 4.1 instead controls the cross-sensor variance of the latent-to-physical error conversion rate through the squared-error decomposition, decoder geometry, and error moments (App. C).

C How latent filtering error translates into physical error

For the orthonormal linear decoder in Appendix B, physical MSE equals reconstruction error plus latent MSE. With a nonlinear decoder, the physical effect of a latent error depends on its direction and where it occurs. For sensors with positive latent MSE, the conversion rate is $\alpha(H) = (\mathrm{RMSE}^2(H) - \varepsilon_r^2)/\mathrm{MSE}_z(H)$. We bound how much this rate varies across sensors and use the bound to relate latent and physical MSE through their Pearson correlation. For a finite sensor library, we give separate conditions for preserving the full ranking and for retaining every physical-error minimizer on a shortlist ranked by latent MSE.

C.1 Pointwise decomposition

Fix a sensor H. Writing $\hat{z}_t := \hat{z}_{t|t}$ for the updated estimate, let $z_t = E(x_t)$, $\tilde{z}_t = z_t - \hat{z}_t$, and $e(x_t) = x_t - D(z_t)$. By the squared-error identity (16),

$$\|x_t - D(\hat{z}_t)\|^2 = \|e(x_t)\|^2 + \|D(z_t) - D(\hat{z}_t)\|^2 + \underbrace{2\langle e(x_t),\, D(z_t) - D(\hat{z}_t)\rangle}_{\mathrm{CROSS}_t}. \tag{21}$$

Under the stationary support assumption, $z_t, \hat{z}_t \in \Omega$ almost surely. Equation (10) bounds the cross-term by $|\mathrm{CROSS}_t| \le 2\varepsilon_\perp \|e(x_t)\|\, \|D(z_t) - D(\hat{z}_t)\|$. Write $d_t = D(z_t) - D(\hat{z}_t)$ for the decoded displacement and $v_t = J_D(z_t)\tilde{z}_t$ for its linear approximation. The segment from z_t to $\hat{z}_t$ lies in Ω, so Taylor's theorem gives $\|d_t - v_t\| \le \eta\|\tilde{z}_t\|^2/2$. The upper distance bound in (7) gives $\|d_t\| \le \ell_U\|\tilde{z}_t\|$. Applying the same bound between z_t and $z_t - s\tilde{z}_t$ for $0 < s \le 1$ and taking the derivative at $s = 0$ from the right gives $\|v_t\| \le \ell_U\|\tilde{z}_t\|$. Since $\|v_t\|^2 = \tilde{z}_t^\top G_D(z_t)\tilde{z}_t$, the error in the squared displacement is $\mathrm{HOT}_t := \|d_t\|^2 - \|v_t\|^2$, with

$$|\mathrm{HOT}_t| = |\langle d_t - v_t, d_t + v_t\rangle| \le \eta\ell_U\|\tilde{z}_t\|^3.$$

C.2 Time averaging and the T_1 decomposition

Fix a sensor H and let ν_H denote the stationary joint law of the physical state, the latent estimate, and its updated covariance. Lemma C.7 assumes that two time averages converge from any admissible initial law. The first identifies the long-time average of the expected latent error outer product with its stationary value,

$$\Sigma_H \;:=\; \mathbb{E}_{\nu_H}[\tilde{z}\tilde{z}^\top] \;=\; \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\tilde{z}_t\tilde{z}_t^\top]. \tag{22}$$

Taking the trace of this limit gives $\mathrm{tr}\Sigma_H = \mathbb{E}_{\nu_H}\|\tilde{z}\|^2 = \lim_T T^{-1}\sum_t \mathbb{E}\|\tilde{z}_t\|^2 = \mathrm{MSE}_z(H)$. Weighting this second moment by the mean decoder metric gives

$$T_1(H) \;:=\; \mathrm{tr}(\overline{G}_D\,\Sigma_H). \tag{23}$$

The encoded state has marginal law μ_z under ν_H for every sensor. The mean decoder metric $\overline{G}_D$ is therefore common to all sensors, while Σ_H depends on the errors left by sensor H. The fluctuation $\Delta_t = G_D(z_t) - \overline{G}_D$ has mean zero, but it can depend on the error at the same state. Writing $M_j(H) = \mathbb{E}_{\nu_H}\|\tilde{z}_t\|^j$ for $j = 2, 3, 4$, Cauchy–Schwarz gives

$$\left|\mathbb{E}_{\nu_H}[\tilde{z}_t^\top \Delta_t \tilde{z}_t]\right| \le \sqrt{\mathbb{E}_{\nu_H}\|\Delta_t\|_F^2}\sqrt{\mathbb{E}_{\nu_H}\|\tilde{z}_t\|^4}.$$

The first factor is δ_D by the definition of the metric variance under μ_z. Under the assumptions of Lemma C.6, the empirical mean of $\|\Delta_t\|_F^2$ along an encoded trajectory converges almost surely to δ_D^2. Splitting $G_D(z_t)$ into its mean and its fluctuation now gives

$$\mathbb{E}_{\nu_H}[\tilde{z}_t^\top G_D(z_t)\tilde{z}_t] = T_1(H) + \mathcal{E}_J(H), \qquad |\mathcal{E}_J(H)| \le \delta_D\sqrt{M_4(H)}. \tag{24}$$

The CROSS term from (21) is bounded directly by (10) and (7) on Ω:

$$|\mathbb{E}_{\nu_H}[\mathrm{CROSS}_t]| \le 2\varepsilon_\perp\sqrt{\mathbb{E}_{\nu_H}\|e(x_t)\|^2}\,\sqrt{\mathbb{E}_{\nu_H}\|D(z_t) - D(\hat{z}_t)\|^2} \le 2\varepsilon_\perp\varepsilon_r\ell_U\sqrt{\mathrm{MSE}_z(H)}. \tag{25}$$

Here Cauchy–Schwarz is applied under ν_H, whose physical-state marginal is μ, so $\mathbb{E}_{\nu_H}\|e(x_t)\|^2 = \varepsilon_r^2$.

Taking the expectation of the Taylor remainder bound gives

$$|\mathbb{E}_{\nu_H}[\mathrm{HOT}_t]| \le \eta \ell_U M_3(H). \tag{26}$$

The second time-average assumption identifies the stationary physical loss with $\mathrm{RMSE}^2(H)$. Combining the three terms gives

$$\mathrm{RMSE}^2(H) - \varepsilon_r^2 = T_1(H) + \mathcal{E}_{\mathrm{total}}(H), \tag{27}$$

with $|\mathcal{E}_{\mathrm{total}}(H)| \le \delta_D \sqrt{M_4} + 2\varepsilon_\perp \varepsilon_r \ell_U \sqrt{M_2} + \eta \ell_U M_3$.

C.3 From the pointwise bound to $\mathrm{Var}_H(\alpha)$

We now draw the sensor from $\mathcal{P}$. The uniform lower bound $M_2(H) = \mathrm{MSE}_z(H) \ge m_\star^2 > 0$ lets us divide (27) by the latent MSE for every admissible sensor. Writing $\beta(H)$ for the contribution of the mean metric per unit latent MSE gives

$$\alpha(H) = \beta(H) + \frac{\mathcal{E}_{\mathrm{total}}(H)}{\mathrm{MSE}_z(H)}, \qquad \beta(H) := \frac{T_1(H)}{\mathrm{MSE}_z(H)} = \frac{\mathrm{tr}(\overline{G}_D \, \Sigma_H)}{\mathrm{tr}\Sigma_H}. \tag{28}$$

Let $\kappa_4(H) = M_4(H)/M_2(H)^2$. Dividing the three remainder bounds by $M_2(H)$ and using $M_3(H) \le \sqrt{M_2(H) M_4(H)}$ gives

$$|\alpha(H) - \beta(H)| \le \delta_D \sqrt{\kappa_4(H)} + \frac{2\varepsilon_\perp \varepsilon_r \ell_U}{\sqrt{M_2(H)}} + \eta \ell_U \sqrt{\kappa_4(H) M_2(H)}. \tag{29}$$

By Cauchy–Schwarz, the square of this sum is at most three times the sum of the squares. Averaging over sensors and using $M_2(H) \ge m_\star^2$ and $\kappa_4(H) \le \kappa_4^\star$ from the stationary moment bounds gives

$$\begin{aligned} \mathrm{Var}_H(\alpha - \beta) &\le \mathbb{E}_{\mathcal{P}}[(\alpha - \beta)^2] \\ &\le 3\kappa_4^\star \delta_D^2 + 12\varepsilon_\perp^2 \varepsilon_r^2 \ell_U^2 / m_\star^2 + 3\kappa_4^\star \eta^2 \ell_U^2 \mathbb{E}_{\mathcal{P}}[M_2(H)]. \end{aligned} \tag{30}$$

Finally, $\mathrm{Var}_H(\alpha) \le 2\mathrm{Var}_H(\beta) + 2\mathrm{Var}_H(\alpha - \beta)$. Substituting the alignment variance $A(\overline{G}_D, \{\Sigma_H\}) = \mathrm{Var}_H(\beta)$ and the bound (30) yields

$$\begin{aligned} \mathrm{Var}_H(\alpha) \le{}& 6\kappa_4^\star \delta_D^2 + 6\ell_U^2 \kappa_4^\star \eta^2 \, \mathbb{E}_{\mathcal{P}}[\mathrm{MSE}_z(H)] \\ &+ 24\ell_U^2 \varepsilon_\perp^2 \varepsilon_r^2 / m_\star^2 + 2A(\overline{G}_D, \{\Sigma_H\}). \end{aligned} \tag{31}$$

This proves (11) in Lemma 4.1.

Appendix C.9 gives bounds on the alignment variance.

Approximating the cross-term. For a sensor with positive latent MSE, one can also approximate the decoded displacement in the cross-term by $J_D(\hat{z})\tilde{z}$, expanding at the updated estimate. Taylor's bound and Cauchy–Schwarz show that the resulting change in the expected cross-term divided by $\mathrm{MSE}_z(H)$ is at most

$$\frac{\eta \, \mathbb{E}_{\nu_H}[\|e(x)\| \|\tilde{z}\|^2]}{\mathrm{MSE}_z(H)} \le \eta \varepsilon_r \sqrt{\kappa_4(H)}, \qquad \kappa_4(H) = M_4(H)/M_2(H)^2.$$

C.4 Pearson correlation under an additive perturbation

Lemma C.1 (Pearson correlation under an additive perturbation)**.** *Let $X, W \in L^2(\mathcal{P})$ with $\mathrm{Var}X > 0$, and let $\bar{\alpha} \in \mathbb{R}$. Set $Y := \bar{\alpha} X + W$. Then*

$$\mathrm{Var}X \cdot \mathrm{Var}Y - \mathrm{Cov}(X,Y)^2 \;=\; \mathrm{Var}X \cdot \mathrm{Var}W - \mathrm{Cov}(X,W)^2 \;\ge\; 0, \tag{32}$$

and, when $\mathrm{Var}Y > 0$,

$$1 - \rho_{\mathrm{Pearson}}(X,Y)^2 \;=\; \frac{\mathrm{Var}X \cdot \mathrm{Var}W - \mathrm{Cov}(X,W)^2}{\mathrm{Var}X \cdot \mathrm{Var}Y} \;\le\; \frac{\mathrm{Var}W}{\mathrm{Var}Y}.$$

If in addition $|\bar{\alpha}|\sqrt{\mathrm{Var}X} \ge 2\sqrt{\mathrm{Var}W}$ and $\bar{\alpha} \ne 0$, then $\mathrm{Var}Y \ge \frac{1}{4}\bar{\alpha}^2 \mathrm{Var}X$, $\mathrm{sign}(\bar{\alpha}) \, \rho_{\mathrm{Pearson}}(X,Y) \ge 0$, and

$$\big|1 - \mathrm{sign}(\bar{\alpha}) \, \rho_{\mathrm{Pearson}}(X,Y)\big| \;\le\; 1 - \rho_{\mathrm{Pearson}}(X,Y)^2 \;\le\; \frac{4 \, \mathrm{Var}W}{\bar{\alpha}^2 \, \mathrm{Var}X}. \tag{33}$$

Proof. Expanding $\mathrm{Cov}(X,Y) = \bar{\alpha}\mathrm{Var}X + \mathrm{Cov}(X,W)$ and $\mathrm{Var}Y = \bar{\alpha}^2\mathrm{Var}X + 2\bar{\alpha}\mathrm{Cov}(X,W) + \mathrm{Var}W$ gives (32). Cauchy–Schwarz makes its right-hand side non-negative; dropping the squared covariance gives the bound on $1 - \rho^2_{\mathrm{Pearson}}$.

Under this bound on $\mathrm{Var}W$, the reverse triangle inequality for centered variables gives

$$\sqrt{\mathrm{Var}Y} \geq |\bar{\alpha}|\sqrt{\mathrm{Var}X} - \sqrt{\mathrm{Var}W} \geq \tfrac{1}{2}|\bar{\alpha}|\sqrt{\mathrm{Var}X}.$$

Also $|\mathrm{Cov}(X,W)| \leq \sqrt{\mathrm{Var}X\mathrm{Var}W} \leq \frac{1}{2}|\bar{\alpha}|\mathrm{Var}X$, so $\mathrm{Cov}(X,Y)$ has the sign of $\bar{\alpha}$. Thus, with $s = \mathrm{sign}(\bar{\alpha})$, $0 \leq s\rho_{\mathrm{Pearson}} \leq 1$ and $|1 - s\rho_{\mathrm{Pearson}}| \leq 1 - \rho^2_{\mathrm{Pearson}}$. Combining this with $\mathrm{Var}Y \geq \bar{\alpha}^2\mathrm{Var}X/4$ proves (33). □

C.5 Proof of Theorem 4.2 (variance-controlled Pearson bound)

For filtering, take $X = \mathrm{MSE}_z(H)$ and $Y = \mathrm{RMSE}^2(H)$. A change in the conversion rate has a larger effect on physical error when latent error is large. The residual in the Pearson bound is therefore the product $(\alpha - \bar{\alpha})X$. The following argument allows the two factors to be dependent.

Proof. Let X, Y be as in Theorem 4.2, and set $\xi := \alpha - \bar{\alpha}$ and $W := \xi X$. Then $Y - \varepsilon_r^2 = \bar{\alpha}X + W$, and subtracting the constant ε_r^2 preserves variance and correlation. Write $\sigma_\alpha^2 := \mathrm{Var}_{\mathcal{P}}(\alpha)$. Hypothesis (iv) gives $\mathbb{E}_{\mathcal{P}}[\xi^4] \leq K_\xi^\star \sigma_\alpha^4$ when $\sigma_\alpha > 0$; when $\sigma_\alpha = 0$, $\xi = 0$ a.s., so the same bound holds. Cauchy–Schwarz and hypothesis (iii) therefore give

$$\mathrm{Var}W \;\leq\; \mathbb{E}_{\mathcal{P}}[\xi^2 X^2] \;\leq\; (\mathbb{E}_{\mathcal{P}}\xi^4)^{1/2}(\mathbb{E}_{\mathcal{P}}X^4)^{1/2} \;\leq\; K_\xi^{\star\,1/2}\,\sigma_\alpha^2\,M_{X,4}^{1/2}, \tag{34}$$

Using $\bar{\alpha} \geq c_\alpha$ from (i) and $\mathrm{Var}X \geq v_X$ from (ii), the smallness condition of Lemma C.1 is satisfied whenever

$$K_\xi^{\star\,1/2}\,\sigma_\alpha^2\,M_{X,4}^{1/2} \;\leq\; \tfrac{1}{4}\,c_\alpha^2\,v_X \qquad\Longleftrightarrow\qquad \sigma_\alpha^2 \;\leq\; 1/a^2, \qquad a \;:=\; 2\,K_\xi^{\star\,1/4}\,M_{X,4}^{1/4}/(c_\alpha\sqrt{v_X}).$$

Under $\sigma_\alpha^2 \leq \varepsilon \leq \varepsilon_0 := 1/a^2$, Lemma C.1 together with (34) gives $\mathrm{Var}Y \geq \bar{\alpha}^2\mathrm{Var}X/4 > 0$, $\rho_{\mathrm{Pearson}}(Y,X) \geq 0$, and

$$\big|1 - \rho_{\mathrm{Pearson}}(Y,X)\big| \;\leq\; \frac{4\,\mathrm{Var}W}{\bar{\alpha}^2\,\mathrm{Var}X} \;\leq\; \frac{4\,K_\xi^{\star\,1/2}\,M_{X,4}^{1/2}}{c_\alpha^2\,v_X}\,\sigma_\alpha^2 \;=\; a^2\,\sigma_\alpha^2 \;\leq\; a^2\,\varepsilon. \tag{35}$$

This proves (13). □

Under the moment assumptions of Theorem 4.2, Corollary 4.3 follows by taking $X = \mathrm{MSE}_z$, $Y = \mathrm{RMSE}^2$, and $\varepsilon = B_{\mathrm{sm}}$ from Lemma 4.1 whenever $B_{\mathrm{sm}} \leq \varepsilon_0$.

Simplification under independence. Suppose $\xi = \alpha - \bar{\alpha}$ and X are independent under $\mathcal{P}$, with finite second moments, $\bar{\alpha} \neq 0$, and $\mathrm{Var}X > 0$. Since $\mathbb{E}_{\mathcal{P}}\xi = 0$, we have $\mathrm{Cov}(X, \xi X) = 0$ and $\mathrm{Var}W = \sigma_\alpha^2\mathbb{E}_{\mathcal{P}}[X^2]$. Equation (32) then gives

$$1 - \rho^2_{\mathrm{Pearson}} = \frac{\mathrm{Var}W}{\bar{\alpha}^2\mathrm{Var}X + \mathrm{Var}W} \leq \frac{\sigma_\alpha^2\mathbb{E}_{\mathcal{P}}[X^2]}{\bar{\alpha}^2\mathrm{Var}X}.$$

C.6 Sensor rankings in finite libraries and continuous populations

A small change in physical error can reverse the order of two sensors whose errors are close. We first give sufficient conditions for preserving ranks in a finite sensor library. We then consider the population rank correlation for continuously distributed errors.

Throughout this subsection let

$$\begin{aligned} X(H) &:= \mathrm{MSE}_z(H), \qquad Y(H) := \mathrm{RMSE}^2(H), \quad \xi(H) := \alpha(H) - \bar{\alpha}, \\ Y_0(H) &:= \varepsilon_r^2 + \bar{\alpha}X(H), \quad R_\xi(H) := \xi(H)X(H), \end{aligned}$$

so $Y = Y_0 + R_\xi$ and $\bar{\alpha} > 0$ under Theorem 4.2. For a finite empirical library with minimum atom mass $p_{\min}$ and $X_{\max} := \sup_H X(H)$, the variance condition alone gives

$$\sup_H |R_\xi(H)| \;\leq\; X_{\max}\sqrt{\mathrm{Var}_{\mathcal{P}}(\alpha)/p_{\min}},$$

where $p_{\min} > 0$. If each residual has magnitude at most r, the difference between two sensors' errors can change by at most $2r$. For the full ranking we compare this change with the smallest gap in Y_0. For a shortlist of size k, we compare it with the gap between the best physical squared error and the $(k+1)$st best.

Proposition C.2 (Spearman preservation under a finite-library margin)**.** *Let $\mathcal{H}$ be a finite sensor library. Define*

$$\gamma_S := \min_{H \neq H'} |Y_0(H) - Y_0(H')|.$$

Assume $\gamma_S > 0$. If

$$\sup_{H \in \mathcal{H}} |R_\xi(H)| < \gamma_S/2,$$

then the ordering induced by Y is identical to the ordering induced by X, so the no-tie Spearman correlation between RMSE^2 and MSE_z is exactly 1. In particular, if $\sup_H |R_\xi(H)| \le B_S\sqrt{\varepsilon}$, the conclusion holds whenever $2B_S\sqrt{\varepsilon} < \gamma_S$.

Proof. Because $\bar{\alpha} > 0$, Y_0 and X induce the same ordering. For any distinct H, H',

$$|(Y(H) - Y(H')) - (Y_0(H) - Y_0(H'))| \le |R_\xi(H)| + |R_\xi(H')| < \gamma_S.$$

Since $|Y_0(H) - Y_0(H')| \ge \gamma_S$, the sign of every pairwise difference is preserved. Thus all ranks agree. □

Proposition C.3 (Top-k inclusion under a finite-library squared-error margin)**.** *Let $\mathcal{H}$ be finite with $|\mathcal{H}| = N$, and let $Y_{(1)} \le \cdots \le Y_{(N)}$ be the sorted oracle squared errors. Fix $k < N$ and assume the squared-error top-k margin*

$$\gamma_k^Y := Y_{(k+1)} - Y_{(1)} > 0.$$

Let S_k^X contain the k sensors with the smallest latent MSE, breaking ties at the cutoff arbitrarily. If

$$\sup_{H \in \mathcal{H}} |R_\xi(H)| < \gamma_k^Y/2,$$

then the shortlist S_k^X contains every minimizer of Y and has zero regret. In particular, if $\sup_H |R_\xi(H)| \le B_k\sqrt{\varepsilon}$, the conclusion holds whenever $2B_k\sqrt{\varepsilon} < \gamma_k^Y$.

Proof. Let $H_\star$ be any oracle minimizer of Y. Suppose $H_\star \notin S_k^X$. Then there are k sensors H in S_k^X with $X(H) \le X(H_\star)$. For each such H,

$$Y(H) - Y(H_\star) = \bar{\alpha}\big(X(H) - X(H_\star)\big) + R_\xi(H) - R_\xi(H_\star) < \gamma_k^Y,$$

because the first term is nonpositive and the residual difference is smaller than γ_k^Y. Thus these k sensors, together with $H_\star$, all have squared error strictly below $Y_{(1)} + \gamma_k^Y = Y_{(k+1)}$, contradicting the definition of $Y_{(k+1)}$. Therefore $H_\star \in S_k^X$. □

For error marginals with bounded densities, the following proposition gives a distributional Spearman bound of order $\sqrt{\varepsilon}$.

Proposition C.4 (Distributional Spearman rate under continuous marginals)**.** *Assume $Y = \varepsilon_r^2 + \alpha(H)X(H)$, $\bar{\alpha} \ge c_\alpha > 0$, and $\mathbb{E}_{\mathcal{P}}[X^4] \le M_{X,4} < \infty$. Assume also that the error distributions across sensors have bounded densities:*

(D1) *$X = \mathrm{MSE}_z$ has density f_X with $\sup f_X \le L_X < \infty$;*

(D2) *$Y = \mathrm{RMSE}^2$ has density f_Y with $\sup f_Y \le L_Y < \infty$.*

Continuity of F_X and F_Y from (D1)–(D2) *gives the CDF representation of the population Spearman coefficient [46] used in the proof. Then for any $\mathrm{Var}_{\mathcal{P}}(\alpha) \le \varepsilon$,*

$$\big|\, 1 - \rho_{\mathrm{Spearman}}(\mathrm{RMSE}^2, \mathrm{MSE}_z) \,\big| \;\le\; C_{\mathrm{Sp}}\sqrt{\varepsilon}, \qquad C_{\mathrm{Sp}} \;:=\; 12\big(L_X/c_\alpha + L_Y\big)\, M_{X,4}^{1/4}. \tag{36}$$

Proof. Let $U = F_X(X)$ and $V = F_Y(Y)$. Under (D1) and (D2), both are uniform on $[0, 1]$. The population Spearman formula [46] gives $\rho_{\mathrm{Sp}} = 12\mathbb{E}[UV] - 3$. Since $\mathbb{E}[U^2] = 1/3$ and $0 \le U \le 1$,

$$|1 - \rho_{\mathrm{Sp}}| = 12|\mathbb{E}[U(U - V)]| \le 12\mathbb{E}|U - V|.$$

The map from X to $Y_0 = \varepsilon_r^2 + \bar{\alpha}X$ is strictly increasing, so $U = F_{Y_0}(Y_0)$. The density of Y_0 is bounded by $L_X/\bar{\alpha} \le L_X/c_\alpha$, and F_Y is L_Y-Lipschitz. Adding and subtracting $F_Y(Y_0)$ therefore gives

$$\mathbb{E}|U - V| \le (L_X/c_\alpha) \int_{\mathbb{R}} |F_{Y_0}(y) - F_Y(y)|\, dy + L_Y\mathbb{E}|Y - Y_0|.$$

For each y, the integrand is at most $\mathbb{E}|\mathbf{1}_{\{Y_0 \le y\}} - \mathbf{1}_{\{Y \le y\}}|$. For each pair (Y_0, Y), the indicators differ on an interval of length $|Y - Y_0|$. Integrating and applying Tonelli's theorem bounds the integral by $\mathbb{E}|Y - Y_0|$. Finally, $Y - Y_0 = (\alpha - \bar{\alpha})X$, so Cauchy–Schwarz gives

$$\mathbb{E}|Y - Y_0| \le \sqrt{\mathrm{Var}_{\mathcal{P}}(\alpha)}\, (\mathbb{E}X^2)^{1/2} \le \sqrt{\varepsilon}\, M_{X,4}^{1/4}.$$

Combining the bounds gives the stated constant $12(L_X/c_\alpha + L_Y)M_{X,4}^{1/4}$. □

Remark C.5 (On (D2): a sufficient condition). Under (D1), suppose also that $X \ge m_\star^2 > 0$ almost surely and that the conditional density of α given $X = x$ is bounded by a common constant L for f_X-almost every x. At fixed x, the map $\alpha \mapsto \varepsilon_r^2 + x\alpha$ gives a conditional density for Y bounded by L/x. Averaging over X yields $f_Y(y) \le L/m_\star^2$ almost everywhere, as required by (D2).

Gaussian-sensor experiment. On L96 $d = 10$, we ran the EKF with a SiLU decoder for 2000 steps for each of 250 Gaussian sensors ($k = p = 5$, with independent entries of H drawn from $\mathcal{N}(0, 1/d)$). Across these sensors, the Spearman correlation between latent MSE and time-averaged spatial RMSE was 0.864. In a separate multi-trajectory evaluation of 252 coordinate sensors, the correlation between the two trajectory-averaged scores was 0.913 (Fig. 1).

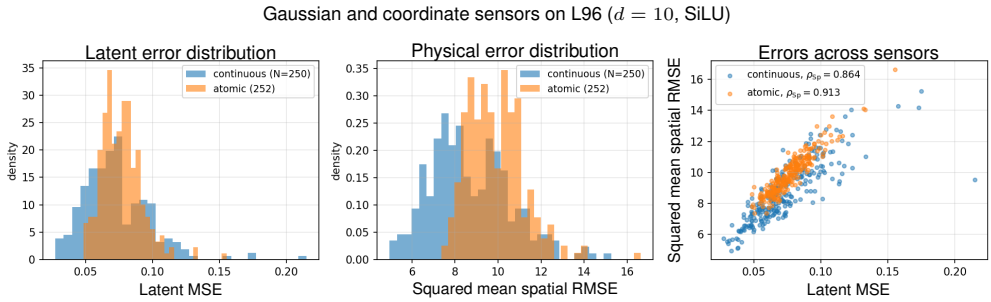


Figure 1: Latent and physical errors across Gaussian sensors (blue) and coordinate sensors (orange). For each Gaussian sensor, we compute spatial RMSE at each retained time step, average over time, and square the result for plotting. For each coordinate sensor, we square the mean RMSE from the multi-trajectory evaluation. Squaring these scores preserves their ranks.

C.7 Stationary moments and time averages

Along a stationary ergodic trajectory of the encoded state, the empirical mean of the squared decoder-metric fluctuation converges to its expectation under μ_z. Filtering errors also depend on the estimate. For each sensor, we assume that the time averages of the expected latent error second moment and the expected physical loss converge to their respective stationary values.

Lemma C.6 (Ergodic time-average)**.** *Assume ergodicity of the encoded process and that J_D is defined μ_z-almost everywhere, Borel measurable, and satisfies $\|J_D(z)\|_{\mathrm{op}} \le \ell_U$ there. The upper distance bound in* (7) *gives this Jacobian norm bound at interior points of Ω. The same bound holds at limits of interior points by continuity of J_D under the decoder smoothness assumption. For a merely Lipschitz decoder it must be checked on the support of μ_z: Lebesgue-almost-everywhere differentiability alone does not imply differentiability under a singular invariant law. The Jacobian bound and Jensen's inequality give $\|J_D^\top J_D\|_F, \|\overline{G}_D\|_F \le \sqrt{k}\ell_U^2$. Thus $\tilde{g}(z) := \|J_D(z)^\top J_D(z) -$*

$\overline{G}_D\|_F^2 \le 4k\ell_U^4$ *is bounded and Borel measurable* μ_z*-a.e., and the Birkhoff–Khinchin pointwise ergodic theorem [52, Thm. 1.14] applied to the encoded process gives*

$$\tfrac{1}{T}\sum_{t=0}^{T-1}\tilde{g}(z_t) \xrightarrow{T\to\infty} \int_{E(\mathcal{A})}\tilde{g}\,\mathrm{d}\mu_z = \delta_D^2 \qquad \textit{a.s.}$$

If the encoded process is itself a positive Harris recurrent Markov chain, the Markov-chain law of large numbers gives the same limit [32, Thm. 17.0.1].

Lemma C.7 (L^2/L^4 EKF error bounds under explicit moment hypotheses)**.** *Assume the following conditions on the joint law and the EKF error moments.*

- Stationary law and time averages. *For each sensor* H*, assume that the joint process* $(x_t, \hat{z}_{t|t}, P_{t|t})$ *has a stationary law* ν_H *with physical marginal* μ*. Write* $(x, \hat{z}, P)$ *for the stationary state, estimate, and updated covariance. The latent error* $\tilde{z} = E(x) - \hat{z}$ *has law* π_H *and second moment* $\Sigma_H = \mathbb{E}_{\nu_H}[\tilde{z}\tilde{z}^\top]$*. Assume*

$$0 \;<\; m_\star^2 \;:=\; \inf_{H\in\mathcal{H}} \mathrm{tr}(\Sigma_H) \quad \textit{and} \quad \sup_{H\in\mathcal{H}} \mathrm{tr}(\Sigma_H) \;<\; \infty.$$

  The encoded state has marginal μ_z *for every sensor. Assume that* $\hat{z} \in \Omega$ *almost surely under* ν_H*. From each admissible initial law, the time average of the expected error second moment converges to its stationary value,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\tilde{z}_t\tilde{z}_t^\top\big] \xrightarrow{T\to\infty} \Sigma_H \qquad \textit{(operator norm).}$$

  For the same admissible initial laws, the physical loss also satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|x_t - D(\hat{z}_{t|t})\|^2 \longrightarrow \mathbb{E}_{\nu_H}\|x - D(\hat{z})\|^2.$$

- Uniform fourth-moment bound. *For each admissible initialization, including initialization from* ν_H*, assume that the following bound holds with a common finite constant* $C_4^\star$*. Here* $\mathbb{E}_H$ *averages over the state and filter randomness for fixed* H*.*

$$\sup_{H\in\mathcal{H}} \limsup_{t\to\infty} \mathbb{E}_H\|\tilde{z}_t\|^4 \;\le\; C_4^\star. \tag{37}$$

Then $\mathrm{MSE}_z(H) = \mathrm{tr}\Sigma_H$ *for each sensor, and the stationary errors satisfy*

$$\sup_{H\in\mathcal{H}} \kappa_4(H) \;:=\; \sup_{H\in\mathcal{H}} \frac{\mathbb{E}_{\pi_H}\|\tilde{z}\|^4}{(\mathbb{E}_{\pi_H}\|\tilde{z}\|^2)^2} \;\le\; \kappa_4^\star \;:=\; \frac{C_4^\star}{m_\star^4}. \tag{38}$$

Proof of Lemma C.7. Taking the trace in the time-average limit gives $\mathrm{MSE}_z(H) = \mathrm{tr}\Sigma_H$. When the filter starts from the stationary joint law ν_H, the fourth moment equals $\mathbb{E}_{\pi_H}\|\tilde{z}\|^4$ at every time, so (37) bounds it by $C_4^\star$. Since $\mathbb{E}_{\pi_H}\|\tilde{z}\|^2 = \mathrm{tr}\Sigma_H \ge m_\star^2$, the ratio in (38) is at most $C_4^\star/m_\star^4$ for every sensor. □

Remark C.8 (Stationary support and convergence of the error moments). Under joint stationary initialization both time-average identities hold at every averaging horizon. For other initial laws, a sufficient condition is weak convergence of the time-averaged joint laws to ν_H, together with continuity and uniform integrability of both observables on the relevant support. The physical loss requires a separate condition because the reconstruction residual depends on x.

For the L96 and KS runs, we checked whether the estimates stayed within distance r of $E(\mathcal{A})$, with r set to the 99.9th percentile of the latent error norm. Across the 12 system–activation combinations, $r/\mathrm{diam}(E(\mathcal{A})) \le 0.5$, and all estimated states stayed within distance r of $E(\mathcal{A})$ during the 2,000-step runs. This check concerns the estimated states. The Taylor bound in §4 requires a Hessian bound throughout a convex domain containing every segment from an encoded state to its estimate. Comparisons at $T = 500$ and $T = 2000$ examined the stability of latent MSE. Convergence of the expected physical loss is a separate assumption.

Remark C.9 (The fourth-moment hypothesis). The Taylor remainder is quadratic in the latent error. Bounding its fourth moment by the Hessian estimate would therefore require an eighth moment of the error. We assume the uniform fourth-moment bound in (37) directly.

For each sensor, we estimated $\kappa_4(H)$ by dividing the sample mean of $\|\tilde{z}\|^4$ by the square of the sample mean of $\|\tilde{z}\|^2$. Each 2,000-step run contributed 1,800 steps after burn-in. On L96 $d = 10$, the median across sensors ranged from 2.9 to 4.1 across activations. Individual estimates ranged from 1.4 to 8.8. On L96 $d = 40$ the medians ranged from 1.4 to 1.7 and individual estimates from 1.3 to 2.2. On KS the medians were 1.4, with estimates from 1.2 to 2.0. For comparison, a centered Gaussian error with nonzero covariance has $\kappa_4 \leq 3$, with equality for rank-one covariance.

We compared the estimate from the first 500 retained steps with the estimate from all 1,800. Across 2,608 sensor and model combinations, the relative difference $|\widehat{\kappa}_{4,500} - \widehat{\kappa}_{4,1800}|/\widehat{\kappa}_{4,1800}$ had median 7.4%, 95th percentile 37%, and maximum 153%. These comparisons show how sensitive the estimated ratios are to the window length. The uniform stationary fourth-moment bound remains an assumption.

Proposition C.10 (Uniform lower bound on MSE_z: linear-Gaussian reference)**.** *Under the correctly specified linear-Gaussian reference model of Proposition C.14, take $D_0 = U \in \mathrm{St}(d,k)$, $x_t = Uz_t$, and $E(x) = U^\top x$. Thus the true latent dynamics and observations are $z_{t+1} = Mz_t + w_t^z$ and $y_t = HUz_t + \eta_t$, with independent Gaussian noises of covariances $Q_z \succ 0$ and $\sigma^2 I_p$, respectively. Assume $\rho(M) < 1$, stationarity, and a bounded sensor family $\sup_{H\in\mathcal{H}} \|H\|_F \leq R_{\mathcal{H}} < \infty$. The stationary latent Kalman error (equivalently, the EKF error in this model) satisfies, with $\ell_U = 1$,*

$$\inf_{H\in\mathcal{H}} \mathrm{MSE}_z(H) \;\geq\; \frac{k^2}{k\ell_U^2 R_{\mathcal{H}}^2/\sigma^2 + \mathrm{tr}(Q_z^{-1})} \;>\; 0. \tag{39}$$

For the nonlinear-decoder EKF, Lemma C.7 assumes the uniform lower bound $m_\star^2 := \inf_H \mathrm{MSE}_z(H) > 0$.

Proof. Write $J^\star = (P^\star)^{-1}$. In this reference model $\mathrm{MSE}_z(H) = \mathrm{tr}P^\star$. Since the predicted covariance is at least Q_z, the posterior information equation gives $J^\star \preceq Q_z^{-1} + \bar{F}(H)$. The sensor norm bound gives $\mathrm{tr}\bar{F}(H) \leq k\ell_U^2 R_{\mathcal{H}}^2/\sigma^2$. Applying Cauchy–Schwarz to the eigenvalues of $P^\star$ and their reciprocals gives $\mathrm{tr}P^\star \geq k^2/\mathrm{tr}J^\star$, which yields (39). □

Remark C.11 (Relation to EKF stability results). Local EKF stability analyses [45] and continuous-time Kalman–Bucy moment bounds [9, 14] motivate the moment assumptions. Our proof uses the stationary joint law and uniform error moment bounds of Lemma C.7. Appendix A.3 summarizes the empirical evidence for these assumptions.

C.8 Moment conditions for the Pearson bound

The decoder regularity conditions in §4 hold on a convex domain containing each segment from the true latent state to its estimate. We use the bilinear Hessian norm $\|\nabla^2 D(z)\|_{\mathrm{op}} = \sup_{\|u\|,\|v\|\leq 1} \|\nabla^2 D(z)[u,v]\|$ and a bounded sensor family $\sup_H \|H\|_F \leq R_{\mathcal{H}}$. Section 4 states the stationary error moment bounds used in Lemma 4.1. Appendix A.3 summarizes the evidence for these conditions on the trained models.

Remark C.12 (Moment bounds for the Pearson estimate). Equation (34) uses the standardized fourth moment of $\xi = \alpha - \bar{\alpha}$. This gives a mixed-moment bound proportional to $\mathrm{Var}_H(\alpha)$ even as that variance approaches zero.

Alternatively, assume that $\mathbb{E}_{\mathcal{P}}[\xi^2 X^2] \leq K_W \mathrm{Var}_{\mathcal{P}}(\alpha)$ for some $0 < K_W < \infty$. With this bound in place of (iv), the proof gives $|1 - \rho_{\mathrm{Pearson}}| \leq a_W^2 \varepsilon$ for $a_W = 2\sqrt{K_W}/(c_\alpha \sqrt{v_X})$ whenever $\mathrm{Var}_{\mathcal{P}}(\alpha) \leq \varepsilon \leq a_W^{-2}$. On a finite sensor library with minimum atom mass $p_{\min} > 0$, $\max_H \xi(H)^2 \leq \mathbb{E}_{\mathcal{P}}[\xi^2]/p_{\min}$ gives $K_\xi \leq p_{\min}^{-1}$ for every nonconstant α, as required in (iv). A bound $X \leq X_{\max}$ instead gives the mixed-moment estimate directly with $K_W = X_{\max}^2$, since $\mathbb{E}_{\mathcal{P}}[\xi^2 X^2] \leq X_{\max}^2 \sigma_\alpha^2$.

Let $X = \mathrm{MSE}_z(H)$. Suppose each model in a sequence satisfies the hypotheses of Lemma 4.1 and Theorem 4.2 with common constants $c_\alpha, v_X, M_{X,4}, K_\xi^\star$. The fourth-moment bound gives

$\mathbb{E}_{\mathcal{P}} X \le M_{X,4}^{1/4}$. If $\ell_U, \kappa_4^\star, \varepsilon_r/m_\star$ are uniformly bounded and $\delta_D, \eta, \varepsilon_\perp, A$ tend to zero, then (11) gives $\mathrm{Var}_H(\alpha) \to 0$. Since the threshold a^{-2} is common to the sequence, the Pearson bound gives

$$0 \le 1 - \rho_{\mathrm{Pearson}}(\mathrm{RMSE}^2, X) \le a^2 \mathrm{Var}_H(\alpha) \to 0$$

for all sufficiently large indices.

The alignment variance is zero when $\overline{G}_D$ is proportional to the identity. Under the reference-model and saturation assumptions in Appendix C.9, it also tends to zero whenever the right-hand side of (41) does.

On $\bar{\alpha} > 0$. The exact decomposition (21), divided by $\mathrm{MSE}_z(H)$, gives

$$\alpha(H) = \frac{\mathbb{E}\|D(z) - D(\hat{z})\|^2}{\mathrm{MSE}_z(H)} + \frac{\mathbb{E}[\mathrm{CROSS}]}{\mathrm{MSE}_z(H)}.$$

If the decoder also satisfies the lower distance bound in (7), the first ratio lies in $[\ell_D^2, \ell_U^2]$. The second has absolute value at most $2\varepsilon_\perp \varepsilon_r \ell_U / m_\star$ by (25). Thus $\ell_D^2 > 2\varepsilon_\perp \varepsilon_r \ell_U / m_\star$ is one sufficient condition for uniform positivity. For a common lower bound along a sequence of models, the difference between the two sides of this strict inequality must stay bounded away from zero. The Pearson theorem uses the weaker requirement $\bar{\alpha} \ge c_\alpha > 0$ directly.

C.9 Bounds on the alignment variance

The first bound uses the eigenvalues of the mean decoder metric and applies to the actual filtering error second moment. The second relates variation in the alignment across sensors to variation in their averaged Fisher matrices. It requires a linear–Gaussian model in which the Riccati covariance equals the error covariance, together with uniform saturation conditions.

Lemma C.13 (Bounds on the alignment variance)**.** *Let $\overline{G}_D = U \Lambda U^\top$ with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_k)$, $\Delta_\lambda := \lambda_{\max}(\overline{G}_D) - \lambda_{\min}(\overline{G}_D)$. Define the spectral weights $w_i(H) := [U^\top \Sigma_H U]_{ii} / \mathrm{tr}\Sigma_H \in [0,1]$ with $\sum_i w_i \equiv 1$.*

Unconditional bound. *For any $\Sigma_H \succeq 0$ with $\mathrm{tr}\Sigma_H > 0$,*

$$A(\overline{G}_D, \{\Sigma_H\}) \;\le\; \Delta_\lambda^2 / 4. \tag{40}$$

Lipschitz bound for the linear–Gaussian reference model. *Under the additional hypothesis of Proposition C.14 that Σ_H coincides with the steady-state DARE covariance $P^\star(\bar{F}(H))$ of the linear-Gaussian reference system in that proposition, and assuming the uniform observation-saturation constants*

$$\phi_\star \;:=\; \inf_{H \in \mathcal{H}} \lambda_{\min}\big(\bar{F}(H)\big) \;>\; 0, \qquad \kappa_\star \;:=\; \inf_{H \in \mathcal{H}} \kappa(H) \;>\; 1,$$

where $\kappa(H) := \lambda_{\min}(\bar{F}(H)) / (\|Q_z^{-1}\|_{\mathrm{op}} \|M\|_{\mathrm{op}})$ is the saturation ratio of Proposition C.16, with the same convention $\kappa(H) = +\infty$ when $M = 0$,

$$A \;\le\; \frac{4k\, \lambda_{\max}(\overline{G}_D)^2\, \mathrm{Var}_{\mathcal{P}}(\bar{F})}{(1 - \kappa_\star^{-2})^2 \cdot \phi_\star^4 \cdot (\inf_H \mathrm{tr}\Sigma_H)^2}, \tag{41}$$

with $\mathrm{Var}_{\mathcal{P}}(\bar{F}) := \mathbb{E}_{\mathcal{P}} \|\bar{F} - \mathbb{E}_{\mathcal{P}} \bar{F}\|_F^2$, the scalar variance in the Frobenius norm across $H \sim \mathcal{P}$.

The second inequality uses the following identification of the stationary error covariance.

Proposition C.14 (Linear-Gaussian reference for the Σ_H–DARE identification)**.** *Consider the linear-Gaussian reference system: latent dynamics $z_{t+1} = M z_t + w_t^z$ with $w_t^z \sim \mathcal{N}(0, Q_z)$, linear decoder $D(z) = D_0 z$ (so $J_D \equiv D_0$ and the averaged Fisher is $\bar{F}(H) = \sigma^{-2} D_0^\top H^\top H D_0$), physical state $x_t = D_0 z_t$, observation $y_t = H D_0 z_t + \eta_t$ with $\eta_t \sim \mathcal{N}(0, \sigma^2 I_p)$ and $\sigma^2 > 0$, and the linear Kalman filter on the latent chain. Assume $Q_z \succ 0$ and that $(M, H D_0)$ is detectable (in particular, $\rho(M) < 1$ suffices). Then the stationary error covariance Σ_H of the Kalman filter satisfies $\Sigma_H = P^\star(\bar{F}(H))$, where $P^\star$ is the steady-state DARE covariance of Proposition C.16. Consequently Σ_H depends on H only through $\bar{F}(H)$, and the derivative bound $\|\mathrm{d}\Sigma_H / \mathrm{d}\bar{F}\|_{\mathrm{op}\to\mathrm{op}} = O(\kappa^{-2})$ of Proposition C.16 applies for fixed Q_z and $M \neq 0$ as $\lambda_{\min}(\bar{F}) \to \infty$.*

Proof. Since $Q_z \succ 0$ and (M, HD_0) is detectable, the Kalman posterior covariance converges to the unique positive-definite stabilizing solution Σ_H of the covariance DARE [3, §7]. In information form,

$$\Sigma_H^{-1} = (Q_z + M\Sigma_H M^\top)^{-1} + \sigma^{-2} D_0^\top H^\top H D_0 = (Q_z + M\Sigma_H M^\top)^{-1} + \bar{F}(H).$$

Thus $\Sigma_H = P^\star(\bar{F}(H))$. Proposition C.16 gives the stated $O(\kappa^{-2})$ sensitivity rate for fixed Q_z and $M \neq 0$ as $\lambda_{\min}(\bar{F}) \to \infty$.

When the stationary prediction law in (N1) exists, the constant Jacobian $J_D \equiv D_0$ gives $\pi_\nu(B) = D_0^\top B D_0$, so $P_\nu^\star(B)$ satisfies the same DARE. □

Remark C.15 (Scope of the linear-Gaussian reference). In the nonlinear EKF, the observation Jacobian is evaluated at each predicted state. Here Σ_H is the stationary second moment of the actual filtering error. The spectral bound (40) applies to this moment. To use (41), one also needs the Riccati identification in Proposition C.14 and the stated uniform saturation conditions.

C.9.1 Spectral bound

In the eigenbasis of $\overline{G}_D$, the diagonal entries of the normalized moment matrix are nonnegative and sum to one. Thus the weighted average $\beta(H) = \sum_i \lambda_i w_i(H)$ satisfies $\beta(H) \in [\lambda_{\min}, \lambda_{\max}]$ for every H. By Popoviciu's variance inequality (for any random variable $X \in [a, b]$, $\mathrm{Var}(X) \leq (b-a)^2/4$),

$$A = \mathrm{Var}_H(\beta) \;\leq\; \frac{\Delta_\lambda^2}{4}. \tag{42}$$

C.9.2 Sensitivity bound

Assume the covariance identification and the uniform bounds $\phi_\star > 0$ and $\kappa_\star > 1$ in Lemma C.13. We first bound the change in $P^\star$ as $\bar{F}$ varies, then bound the change in β as a function of $P^\star$.

Step (i). Lipschitz bound for the Riccati solution. By Proposition C.16 applied uniformly with $\inf_H \kappa(H) \geq \kappa_\star > 1$ and $\inf_H \lambda_{\min}(\bar{F}(H)) \geq \phi_\star$,

$$\sup_{H \in \mathcal{H}} \|\mathrm{d}P^\star(\bar{F}(H))\|_{\mathrm{op}\to\mathrm{op}} \;\leq\; \frac{\phi_\star^{-2}}{1 - \kappa_\star^{-2}} \;=:\; C(\kappa_\star).$$

Proposition C.16 gives the same derivative bound in the Frobenius norm. Integrating along the line segment between the two Fisher matrices gives

$$\|\Sigma_{H_1} - \Sigma_{H_2}\|_F \;\leq\; C(\kappa_\star)\, \|\bar{F}(H_1) - \bar{F}(H_2)\|_F \quad \text{for all } H_1, H_2 \in \mathcal{H}$$

The segment stays in the region where the derivative bound holds because every convex combination of the two Fisher matrices is bounded below by $\phi_\star I$.

Step (ii). Lipschitz bound for β. For $H_1, H_2 \in \mathcal{H}$, consider the segment $\Sigma_s = (1-s)\Sigma_{H_1} + s\Sigma_{H_2}$, $0 \leq s \leq 1$. Along this segment, $\Sigma_s \succeq 0$ and $\mathrm{tr}\Sigma_s \geq \min\{\mathrm{tr}\Sigma_{H_1}, \mathrm{tr}\Sigma_{H_2}\} \geq \inf_H \mathrm{tr}\Sigma_H > 0$.

The trace-weighted spectral average $\beta(\Sigma) = \mathrm{tr}(\overline{G}_D\Sigma)/\mathrm{tr}(\Sigma)$ has Fréchet derivative

$$D\beta(\Sigma)[\Delta] = \frac{\mathrm{tr}(\overline{G}_D\Delta)}{\mathrm{tr}\Sigma} - \frac{\mathrm{tr}(\overline{G}_D\Sigma)\mathrm{tr}\Delta}{(\mathrm{tr}\Sigma)^2}.$$

Since $|\beta(\Sigma)| \leq \lambda_{\max}$, the Frobenius bounds $\|\overline{G}_D\|_F \leq \sqrt{k}\lambda_{\max}$ and $\|I\|_F = \sqrt{k}$ give

$$\|\nabla\beta(\Sigma)\|_F \;=\; \|D\beta(\Sigma)\|_{F\to\mathbb{R}} \;\leq\; \frac{2\lambda_{\max}\sqrt{k}}{\mathrm{tr}\Sigma} \;\leq\; \frac{2\lambda_{\max}\sqrt{k}}{\inf_H \mathrm{tr}\Sigma_H} \;=:\; L_\beta. \tag{43}$$

Integrating this derivative bound along Σ_s gives $|\beta(H_1) - \beta(H_2)| \leq L_\beta \|\Sigma_{H_1} - \Sigma_{H_2}\|_F$.

Step (iii). Bounding the alignment variance. Composing (i) and (ii),

$$|\beta(H_1) - \beta(H_2)| \;\leq\; L_\beta\, C(\kappa_\star)\, \|\bar{F}(H_1) - \bar{F}(H_2)\|_F. \tag{44}$$

For square-integrable X and an independent copy X', expanding the square gives $\mathbb{E}\|X - X'\|^2 = 2\mathbb{E}\|X - \mathbb{E}X\|^2$. Apply this identity to the scalar $\beta(H)$ and the matrix $\bar{F}(H)$, using the Frobenius norm for the latter. Squaring and averaging over $(H, H') \sim \mathcal{P} \otimes \mathcal{P}$ gives:

$$A = \mathrm{Var}_H(\beta) \;\leq\; L_\beta^2\, C(\kappa_\star)^2\, \mathrm{Var}_{\mathcal{P}}(\bar{F}). \tag{45}$$

The definitions of L_β and $C(\kappa_\star)$ now give (41).

C.10 Sensitivity of the steady-state Riccati covariance

Proposition C.16 (Matrix Riccati insensitivity, linear-Gaussian reference)**.** *Let $Q_z \succ 0$ and $\bar{F} \succeq 0$, and suppose that $J^\star(\bar{F})$ is the unique solution in $\mathrm{Sym}_{++}(k)$ of the information DARE $J = g(J)+\bar{F}$ with Riccati prediction operator $g(J) := (Q_z + MJ^{-1}M^\top)^{-1}$, and let $P^\star(\bar{F}) := J^\star(\bar{F})^{-1}$. For linear maps on $\mathrm{Sym}(k)$, write $\|\cdot\|_{\mathrm{op}\to\mathrm{op}}$ for the norm induced by the matrix spectral norm. Let $\mathrm{d}g(J^\star)$ denote the Fréchet derivative of g at $J^\star$, acting on $\mathrm{Sym}(k)$. For $\bar{F} \succ 0$, the derivative satisfies*

$$\|\mathrm{d}g(J^\star)\|_{\mathrm{op}\to\mathrm{op}} \;\le\; \frac{\|Q_z^{-1}\|_{\mathrm{op}}^2\,\|M\|_{\mathrm{op}}^2}{\lambda_{\min}(\bar{F})^2} \;=:\; \kappa^{-2}, \tag{46}$$

where the dimensionless saturation ratio *is defined as*

$$\kappa \;:=\; \lambda_{\min}(\bar{F}) \,/\, \big(\|Q_z^{-1}\|_{\mathrm{op}}\,\|M\|_{\mathrm{op}}\big),$$

with the convention $\kappa = +\infty$ when $M = 0$. If $\bar{F} \succ 0$ and $\kappa > 1$, then $\mathrm{Id} - \mathrm{d}g(J^\star)$ is invertible, with inverse norm at most $(1-\kappa^{-2})^{-1}$, and the Fréchet derivative of $\bar{F} \mapsto P^\star(\bar{F})$ at $\bar{F}$ is

$$\mathrm{d}P^\star[\Delta\bar{F}] \;=\; -\,J^{\star-1}\,\big(\mathrm{Id} - \mathrm{d}g(J^\star)\big)^{-1}[\Delta\bar{F}]\,J^{\star-1}.$$

The derivative satisfies

$$\|\mathrm{d}P^\star\|_{\mathrm{op}\to\mathrm{op}} \;\le\; \frac{\|J^{\star-1}\|_{\mathrm{op}}^2}{1-\kappa^{-2}} \;\le\; \frac{\lambda_{\min}(\bar{F})^{-2}}{1-\kappa^{-2}}.$$

The same bound holds for the Frobenius-induced operator norm $\|\mathrm{d}P^\star\|_{F\to F}$. For fixed $Q_z \succ 0$ and fixed $M \neq 0$, it is $O(\kappa^{-2})$ as $\lambda_{\min}(\bar{F}) \to \infty$. When $M = 0$, every $\bar{F} \succeq 0$ has the exact solution $J^\star = Q_z^{-1} + \bar{F}$, with $\mathrm{d}g = 0$ and $\mathrm{d}P^\star[\Delta\bar{F}] = -P^\star\Delta\bar{F}P^\star$. Both $\|\mathrm{d}P^\star\|_{\mathrm{op}\to\mathrm{op}}$ and $\|\mathrm{d}P^\star\|_{F\to F}$ are bounded by $\lambda_{\min}(Q_z^{-1}+\bar{F})^{-2}$.

Proof. Differentiating the prediction operator gives

$$\mathrm{d}g(J)[\Xi] = g(J)MJ^{-1}\Xi J^{-1}M^\top g(J).$$

For $\bar{F} \succ 0$, the inequalities $g(J^\star) \preceq Q_z^{-1}$ and $J^\star \succeq \bar{F}$ yield (46) in both the spectral- and Frobenius-induced norms. For $\kappa > 1$, the Neumann series therefore gives $\|(\mathrm{Id} - \mathrm{d}g(J^\star))^{-1}\| \le (1-\kappa^{-2})^{-1}$. The implicit function theorem now gives a differentiable solution near $\bar{F}$. Differentiating the fixed-point equation gives

$$(\mathrm{Id} - \mathrm{d}g(J^\star))\,\mathrm{d}J^\star[\Delta\bar{F}] = \Delta\bar{F}.$$

Combining this with $\mathrm{d}P^\star = -J^{\star-1}(\mathrm{d}J^\star)J^{\star-1}$ proves the derivative formula and norm bounds. For the Frobenius norm, use $\|AXB\|_F \le \|A\|_{\mathrm{op}}\|X\|_F\|B\|_{\mathrm{op}}$. For $M = 0$, $g(J) = Q_z^{-1}$ and inversion of $Q_z^{-1} + \bar{F}$ gives the stated derivative and bounds. This inverse is smooth on the open set $\{\bar{F} \in \mathrm{Sym}(k) : Q_z^{-1} + \bar{F} \succ 0\}$, which contains $\mathrm{Sym}_+(k)$. □

Proposition C.16 bounds the sensitivity of the reference posterior covariance, while the nonlinear EKF gain also depends on the predicted covariance and the observation Jacobian at the current prediction.

D Sensors with the same averaged Fisher information

The reference Fisher matrix averages observation sensitivity over a common distribution of latent states. The EKF evaluates the observation Jacobian at its own predictions. We first show when the Riccati solution based on predicted states can change along a family of sensors with the same reference average. In Appendix D.1, we compute the stationary latent EKF error for one such family.

When two sensors receive the same score

Proposition D.1 (Fisher factorisation)**.** *For any H-independent probability law μ_z, fixed $R \succ 0$, and $\mathbb{E}_{\mu_z}\|J_D(z)\|_F^2 < \infty$, every $H \in \mathbb{R}^{p\times d}$ satisfies*

$$\bar{F}(H) = \pi_\mu(H^\top R^{-1}H), \qquad \pi_\mu(B) := \mathbb{E}_{z\sim\mu_z}\big[J_D(z)^\top B\, J_D(z)\big].$$

For $R = \sigma^2 I_p$, this reduces to $\bar{F}(H) = \sigma^{-2}\pi_\mu(H^\top H)$.

Proof. Substituting the observation Jacobian $HJ_D(z)$ into the definition of $\bar{F}$ gives the identity. The second-moment assumption makes the expectation finite, and the integrand is linear in B.

□

The proof of Theorem 3.2 in §3 applies this factorization to scores that depend on the sensor through $\bar{F}$. Continuous scores of finite-horizon observability Gramians also belong to $\mathcal{D}$ when constructed from fixed latent dynamics and $\bar{F}$ [47]. Another example is Gaussian information gain computed from $\bar{F}$ with a fixed prior covariance [2].

Perturbing a sensor at fixed Fisher information

For the local theorem, take $R = \sigma^2 I_p$ and write $B = H^\top H$. A perturbation in $\ker \pi_\mu$ preserves the reference average. To represent a sensor with p independent rows, the perturbed matrix must also be positive semidefinite of rank p.

Proposition D.2 (Kernel dimension, level-set curve, and Gram factorisation)**.** *For $d > k$,*

$$\dim \ker \pi_\mu \;\geq\; \tfrac{d(d+1)}{2} - \tfrac{k(k+1)}{2} \;>\; 0.$$

Let $\mathcal{M}_p := \{B \in \mathrm{Sym}_+(d) : \mathrm{rank}(B) = p\}$ denote the manifold of $d \times d$ symmetric positive semidefinite matrices of rank exactly p (a smooth embedded submanifold of $\mathrm{Sym}(d)$ of dimension $p(2d-p+1)/2$; see 51, Prop. 2.1).

At $B_0 \in \mathcal{M}_p$, the restricted kernel has $\dim T_{B_0}\mathcal{M}_p \cap \ker \pi_\mu \geq \max\{0,\, p(2d-p+1)/2 - k(k+1)/2\}$, which is positive when $p(2d-p+1) > k(k+1)$.

Assume (N3)*: the restriction $F := \pi_\mu|_{\mathcal{M}_p}$ has constant rank r in a neighborhood U of B_0 in $\mathcal{M}_p$. Then for any $\Delta \in T_{B_0}\mathcal{M}_p \cap \ker \pi_\mu$ with $\|\Delta\|_F = 1$, there exists a smooth curve $B : (-\tau_0, \tau_0) \to \mathcal{M}_p$ with $B(0) = B_0$, $B'(0) = \Delta$, and $\pi_\mu(B(t)) = \pi_\mu(B_0)$ for all $|t| < \tau_0$. Each $B(t)$ factors as $B(t) = H(t)^\top H(t)$ with $H(t) \in \mathbb{R}^{p\times d}$ (spectral decomposition of a rank-p PSD matrix at each t).*

Proof. The first statement is the rank-nullity theorem applied to $\pi_\mu : \mathrm{Sym}(d) \to \mathrm{Sym}(k)$ with $\dim \mathrm{Sym}(d) = d(d+1)/2 > k(k+1)/2 = \dim \mathrm{Sym}(k)$ for $d > k$.

Restricting π_μ to $T_{B_0}\mathcal{M}_p$ gives a linear map into $\mathrm{Sym}(k)$. Since $\dim T_{B_0}\mathcal{M}_p = p(2d-p+1)/2$, rank–nullity gives

$$\begin{aligned}\dim(T_{B_0}\mathcal{M}_p \cap \ker \pi_\mu) &= \dim T_{B_0}\mathcal{M}_p - \mathrm{rank}\big(\pi_\mu|_{T_{B_0}\mathcal{M}_p}\big) \\ &\geq \frac{p(2d-p+1)}{2} - \frac{k(k+1)}{2}.\end{aligned}$$

The dimension is also nonnegative, giving the stated bound.

Linearity gives $\ker \mathrm{d}F_{B_0} = T_{B_0}\mathcal{M}_p \cap \ker \pi_\mu$. Under N3, the constant-rank theorem makes $F^{-1}(F(B_0)) \cap U$ a smooth manifold with this tangent space [28, Thm. 5.12]. Each tangent vector Δ is therefore the derivative at zero of a smooth curve in the level set [28, Prop. 3.23]. Following the straight line could give an infeasible Gramian. For example, with $B_0 = \mathrm{diag}(1, 0)$ and $\Delta = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, the determinant of $B_0 + t\Delta$ is $-t^2$. Along the level-set curve, each $B(t)$ has a rank-p spectral decomposition $V\Lambda V^\top$, so $H = \Lambda^{1/2}V^\top$ gives the required sensor at that point.

□

Why the filter law depends only on the Gramian

A Gramian can represent several observation matrices. With isotropic noise, those matrices differ only in the choice of observation coordinates, as the following coupling shows.

Proposition D.3 (EKF reduction)**.** *Let $R = \sigma^2 I_p$, assume the observation noises $\{\eta_t\}_{t\geq 1}$ are i.i.d. $\mathcal{N}(0, \sigma^2 I_p)$ and independent jointly of the initial true state, the dynamics noises $\{w_s\}_{s\geq 0}$, and the common, possibly random, filter initialisation $(\hat{z}_{0|0}, P_{0|0})$. For any $t \geq 1$ the EKF's joint law of $(\hat{z}_{t|t}, P_{t|t})$ under sensor H coincides (in distribution over the realised trajectory $(x_s)_{s\leq t}$ and the observation noise) with that under H' whenever $H^\top H = H'^\top H'$. In particular, conditional on the current prediction $(\hat{z}_{t|t-1}, P_{t|t-1})$, the covariance-update map depends on H only through $H^\top H$, and hence the law of $P_{t|t}$ does also.*

Proof. Equality of the Gram matrices implies $H' = OH$ for some $O \in O(p)$: the map $Hv \mapsto H'v$ is an isometry between their column spaces and extends to an orthogonal map on $\mathbb{R}^p$. Couple the two filters using the same true trajectory, the same initialisation, and observation noises $\eta'_t = O\eta_t$. Isotropic Gaussian noise is invariant under O, so this coupling has the required marginals and gives $y'_t = Oy_t$.

Suppose the predictions agree, and write $C_t = HJ_D(\hat{z}_{t|t-1})$ and $P = P_{t|t-1}$. Then $C'_t = OC_t$ and

$$K'_t = PC_t^\top O^\top \left[O(C_t P C_t^\top + \sigma^2 I_p)O^\top\right]^{-1} = K_t O^\top.$$

The innovations satisfy $y'_t - H'D(\hat{z}_{t|t-1}) = O[y_t - HD(\hat{z}_{t|t-1})]$, so the state updates coincide. The covariance updates coincide because $K'_t C'_t = K_t C_t$. Prediction uses the same M, Q_z in both filters, completing the induction and proving equality of their joint laws. The following identity also shows that the covariance update depends on H only through $H^\top H$:

$$P_{t|t} = \left(P^{-1} + \sigma^{-2} J_D^\top H^\top H J_D\right)^{-1}.$$

□

Remark D.4 (Filter stationarity depends only on B). Under isotropic observation noise, the coupling gives identical predictions for any two factors of B under corresponding initializations. Under N1(a) we can thus use the same stationary prediction law ν_B for either factor when forming $\pi_\nu(B)$.

How the Riccati solution changes along the sensor family

The average $\pi_\nu(B) = \mathbb{E}_{\nu_B}[J_D^\top B J_D]$ depends on B both through the integrand and through the distribution of predicted states. We first solve the Riccati equation for this average, then differentiate it along a direction that preserves the reference average.

Lemma D.5 (Existence, uniqueness, and C^1 dependence of $J^\star_\nu(B)$)**.** *Assume $\rho(M) < 1$,* (7)*, and conditions N1(a) and N1(c). Then for each $B \in \mathcal{U}_0$, the information DARE*

$$J \;=\; \left(Q_z + MJ^{-1}M^\top\right)^{-1} \;+\; \sigma^{-2}\,\pi_\nu(B), \qquad J \in \mathrm{Sym}_{++}(k), \tag{47}$$

has a unique solution $J^\star_\nu(B)$. If in addition N1(b) and N1(d) hold (see the full conditions in Appendix D.2), then the map $B \mapsto J^\star_\nu(B)$ is continuously Fréchet differentiable on $\mathcal{U}_0$, and correspondingly $P^\star_\nu(B) := J^\star_\nu(B)^{-1}$ is C^1 on $\mathcal{U}_0$.

Proof. Define the Riccati prediction operator $g : \mathrm{Sym}_{++}(k) \to \mathrm{Sym}_{++}(k)$ by $g(J) := (Q_z + MJ^{-1}M^\top)^{-1}$, and the forcing term $\phi(B) := \sigma^{-2}\pi_\nu(B)$. The DARE (47) is the fixed-point equation $J = T_B(J)$ with $T_B(J) := g(J) + \phi(B)$.

Existence and uniqueness. Fix B and write $\Phi = \phi(B)$, which is finite and positive definite by N1(c). Inversion reverses the Loewner order twice in g, so T_B is order-preserving. Also $0 \prec g(J) \preceq Q_z^{-1}$, hence T_B maps $\mathcal{I}_B = [\Phi, \Phi + Q_z^{-1}]$ into itself. Starting at $J_0 = \Phi$, the sequence $J_{n+1} = T_B(J_n)$ is increasing and bounded above by $\Phi + Q_z^{-1}$. It therefore converges in finite dimension to a positive-definite $J^\star$, and continuity gives $T_B(J^\star) = J^\star$.

For uniqueness, for every $c > 1$ and $J \succ 0$, $Q_z \succ 0$ gives

$$g(cJ) = \left(Q_z + c^{-1}MJ^{-1}M^\top\right)^{-1} \prec c\left(Q_z + MJ^{-1}M^\top\right)^{-1} = c\,g(J),$$

and therefore $T_B(cJ) \prec cT_B(J)$. If two fixed points J_1, J_2 existed, let $c = \lambda_{\max}(J_2^{-1/2} J_1 J_2^{-1/2})$, the smallest scalar with $J_1 \preceq cJ_2$. If $c > 1$, monotonicity and strict scaling would give $J_1 = T_B(J_1) \preceq T_B(cJ_2) \prec cJ_2$, contradicting minimality of c. Thus $J_1 \preceq J_2$. Reversing their roles gives $J_2 \preceq J_1$, proving uniqueness. Every positive-definite fixed point lies in $\mathcal{I}_B$, so this uniqueness is global on $\mathrm{Sym}_{++}(k)$.

C^1 dependence on B. Define $F : \mathrm{Sym}_{++}(k) \times \mathcal{U}_0 \to \mathrm{Sym}(k)$ by $F(J, B) = J - g(J) - \phi(B)$. The map g is smooth on $\mathrm{Sym}_{++}(k)$, and ϕ is C^1 by N1(b), so F is jointly C^1. By N1(d), $\partial_J F(J^\star_\nu(B), B) = \mathrm{Id} - \mathrm{d}g(J^\star_\nu(B))$ is invertible, with inverse bounded uniformly on $\mathcal{U}_0$. Applying the C^1 implicit function theorem in local coordinates on $\mathcal{M}_p$ (31, Thm. 3.3.1) gives a C^1 solution near each $B \in \mathcal{U}_0$. Uniqueness makes these local solutions agree on overlaps, yielding C^1 dependence on all of $\mathcal{U}_0$. Smooth matrix inversion then gives the same regularity for $P^\star_\nu$.

□

Proposition D.6 (Nonzero directional derivative under (N2))**.** *Under the hypotheses of Lemma D.5 and the non-degeneracy condition* (N2) *of Theorem 3.4, let $\Delta \in T_{B_0}\mathcal{M}_p \cap \ker \pi_\mu$ be any direction satisfying $\pi'_\nu(\Delta) \neq 0$. Then the Fréchet derivative $\mathrm{d}P^\star_\nu\big|_{B_0}[\Delta]$ of $B \mapsto P^\star_\nu(B)$ at B_0 is a nonzero element of* $\mathrm{Sym}(k)$*. Consequently, for any C^1 curve $B : (-\tau, \tau) \to \mathcal{U}_0$ with $B(0) = B_0$ and $B'(0) = \Delta$, we have $P^\star_\nu(B(t)) \neq P^\star_\nu(B(0))$ for all sufficiently small $t \neq 0$.*

Proof. Write $J = J^\star_\nu(B_0)$, $P = J^{-1}$, and let $\dot{J}, \dot{P}$ denote their derivatives along Δ. By Lemma D.5, the solution is differentiable. Differentiating the fixed-point equation and the identity $P = J^{-1}$ gives

$$(\mathrm{Id} - \mathrm{d}g(J))\dot{J} = \sigma^{-2}\pi'_\nu(\Delta), \qquad \dot{P} = -P\dot{J}P.$$

The right-hand side of the first equation is nonzero by (N2), so $\dot{J} \neq 0$. Since P is invertible, $\dot{P} \neq 0$ as well. Along any C^1 curve with $B(0) = B_0$ and $B'(0) = \Delta$, we have $P^\star_\nu(B(t)) - P = t\dot{P} + o(|t|)$. The two covariance matrices therefore differ for all sufficiently small $t \neq 0$. □

Proof of Theorem 3.4. Choose a direction Δ satisfying N2 and normalize it so that $\|\Delta\|_F = 1$. By N3 and Proposition D.2, there is a C^1 curve $B(t)$ in $\mathcal{U}_0$ with $B(0) = B_0$, $B'(0) = \Delta$, and $\pi_\mu(B(t)) = \pi_\mu(B_0)$. Proposition D.6 gives $P^\star_\nu(B(t)) \neq P^\star_\nu(B_0)$ for all sufficiently small $t \neq 0$. For any such t, factor the rank-p Gramian as $B(t) = H_1^\top H_1$ with H_1 of full row rank and take $H_2 = H_0$. Proposition D.1 gives $\bar{F}(H_1) = \bar{F}(H_2)$. □

The theorem distinguishes two Riccati matrices. To compare latent MSEs we also need their traces and the error correction in (9). In the scalar example below we can calculate both terms.

D.1 An exact calculation of stationary EKF error

Proof of Proposition 3.3. The true states Z_t are independent draws from ρ_a, and the physical state is $x_t = D(Z_t) = (Z_t, \cos Z_t/\sqrt{m})^\top$. The encoder $E(x) = x_1$ recovers Z_t exactly on the compact set $D([-a, a])$. The EKF uses a Gaussian transition with $M = 0$ and the same variance Q as the truth. Thus the reference average samples the whole latent distribution, while the filter always linearizes at zero.

To write the physical dynamics in the form used in the paper, take $\Phi(x) = \mathbb{E}D(Z)$ and $w_t = D(Z_{t+1}) - \Phi(x_t)$. The encoded chain is stationary and ergodic, and $ME(x) = E(\Phi(x)) = 0$. The scalar transition density extends by zero to a C^3 function with finite Fisher information, since $(\rho'_a)^2/\rho_a$ is of order $(a - |z|)^2$ near either endpoint.

Reference average. Substituting $u = z/a$ in the density integrals gives

$$Q = a^2 \frac{\int_{-1}^{1} u^2(1-u^2)^4\,du}{\int_{-1}^{1}(1-u^2)^4\,du} = \frac{a^2}{11}.$$

The density is positive on $(-a, a)$, so $m = \mathbb{E}\sin^2 Z$ and $v = \mathbb{E}(\cos Z - 1)^2$ are strictly positive. Symmetry gives $\mathbb{E}\sin Z = 0$. Writing $c = \cos\theta$ and $s = \sin\theta$, we obtain

$$J_D(z) = (1, -\sin z/\sqrt{m})^\top, \qquad \bar{F}(H_\theta) = \frac{1}{r}\mathbb{E}\left(c - \frac{s\sin Z}{\sqrt{m}}\right)^2 = \frac{c^2 + s^2}{r} = \frac{1}{r}.$$

In fact, $\pi_\mu(B) = \mathrm{tr}B$ for every symmetric 2×2 matrix B. Thus the entire unit-sensor family lies on one exact fibre.

EKF update and true error. Because the working model has $M = 0$, every forecast has $\hat{z}_{t|t-1} = 0$ and $P_{t|t-1} = Q$, independently of the previous update. The observation function is $h_\theta(z) = cz + s\cos z/\sqrt{m}$, with $h'_\theta(0) = c$. Therefore the gain and internal posterior covariance are

$$K_\theta = \frac{Qc}{r + Qc^2}, \qquad P^{\mathrm{EKF}}_{t|t} = \frac{Qr}{r + Qc^2}.$$

The prediction law is $\nu_B = \delta_0$, so the information term is c^2/r. Writing $P^\star_\nu(\theta) := P^\star_\nu(H_\theta^\top H_\theta)$, the Riccati equation gives

$$P^\star_\nu(\theta) = (Q^{-1} + c^2/r)^{-1} = P^{\mathrm{EKF}}_{t|t}.$$

The innovation and error retain the nonlinear part of the observation:

$$y_t - h_\theta(0) = cZ_t + \frac{s(\cos Z_t - 1)}{\sqrt{m}} + \eta_t,$$
$$Z_t - \hat{z}_{t|t} = (1 - K_\theta c)Z_t - \frac{K_\theta s(\cos Z_t - 1)}{\sqrt{m}} - K_\theta \eta_t.$$

The cross term $\mathbb{E}[Z(\cos Z - 1)]$ is zero by symmetry. All noise cross terms vanish by independence and zero mean. Hence

$$\Sigma_{H_\theta} = (1 - K_\theta c)^2 Q + K_\theta^2 r + \frac{K_\theta^2 s^2 v}{m}$$
$$= \frac{Qr}{r + Qc^2} + \frac{Q^2 c^2 s^2 v}{m(r + Qc^2)^2}.$$

The gain uses the slope c at the forecast state $z = 0$. The nonlinear term $s(\cos Z_t - 1)/\sqrt{m}$ in the innovation accounts for the additional term in the MSE. Each error depends only on the current i.i.d. pair (Z_t, η_t) and has finite second moment. Its distribution is stationary after the first forecast, and the displayed second moment is the true stationary latent MSE. This second moment includes any bias in the estimate.

Effect of rotating the sensor. On $0 < \theta \le \pi/4$, the posterior variance $Qr/(r + Q\cos^2\theta)$ increases strictly because its denominator decreases strictly. The correction $C(\theta) = \Sigma_{H_\theta} - P^\star_\nu(\theta)$ also increases strictly. Its numerator is proportional to $\cos^2\theta \sin^2\theta = \sin^2(2\theta)/4$, which increases strictly on this interval, and its positive denominator decreases. For $a = \sqrt{11}$ and $r = 1$, the stationary MSE rises from $1/2$ at $\theta = 0$ to $2/3 + v/(9m)$ at $\theta = \pi/4$, with the same averaged Fisher information at both angles.

Local hypotheses of Theorem 3.4. The decoder is smooth and globally bi-Lipschitz:

$$|z - z'| \le \|D(z) - D(z')\| \le \sqrt{1 + m^{-1}}\,|z - z'|.$$

Thus (7) holds with $\Omega = \mathbb{R}$, and $\rho(M) < 1$ holds because $M = 0$. Both Q and r are positive, and the bounded Jacobian has finite second moment under ρ_a. Every H_θ has rank one. For (N1), the prediction law is δ_0 for every sensor and $\pi_\nu(B) = B_{11}$ is smooth. Near any $B_\theta = H_\theta^\top H_\theta$ with $0 < \theta \le \pi/4$, B_{11} is bounded away from zero. Further, $g(J) = Q^{-1}$, so $\mathrm{d}g = 0$ and $\mathrm{Id} - \mathrm{d}g$ is invertible with inverse norm one.

For (N2), the tangent $\Delta_\theta = dB_\theta/d\theta$ belongs to $T_{B_\theta}\mathcal{M}_1$ (with $d = 2$) and satisfies

$$\pi_\mu(\Delta_\theta) = \frac{d}{d\theta}\mathrm{tr}B_\theta = 0, \qquad \mathrm{d}\pi_\nu[\Delta_\theta] = -\sin(2\theta) \neq 0.$$

For (N3), the trace map on the nonzero rank-one positive-semidefinite manifold has rank one everywhere: its derivative in the radial tangent direction B is $\mathrm{tr}B > 0$. Thus N1–N3 hold locally at each $0 < \theta \le \pi/4$.

□

D.2 Sensor pairs with nearly equal Fisher information

How much can filtering error change when two sensors have nearly equal reference averages? We first illustrate how the pairs are constructed on L96 $d = 10$ with a SiLU decoder, then compare their errors across five systems.

On L96 we estimate π_μ from 5000 latent states. The estimated map has numerical rank 15 on the 55-dimensional space of symmetric matrices, giving an ambient kernel of dimension 40. The tangent space of rank-five Gramians also has dimension 40. Since the image has dimension at most 15, at least 25 tangent directions preserve the estimated average to first order. The restricted derivative had numerical rank 15 at all 50 sampled base points. N3 requires constant rank in a neighborhood for the population map.

Starting from $B_0 = H_0^\top H_0$, draw a random Δ in the kernel of the estimated map and scale it to unit Frobenius norm. The step $B_0 + t\Delta$ preserves the estimated average exactly. To obtain a sensor,

retain steps with positive pth eigenvalue and form $B_t = V_p \Lambda_p V_p^\top$ from the leading p eigenpairs. This projection can change the average. We therefore measure the relative changes in $\bar{F}$ and RMSE for $H_t = \Lambda_p^{1/2} V_p^\top$.

For the 50 L96 reference sensors, we took steps of $t \in \{\pm 0.1, \pm 0.2, \pm 0.5\}$. The median absolute relative change in EKF RMSE was 5.0%, compared with 1.2% for the averaged Fisher matrix (Fig. 2).

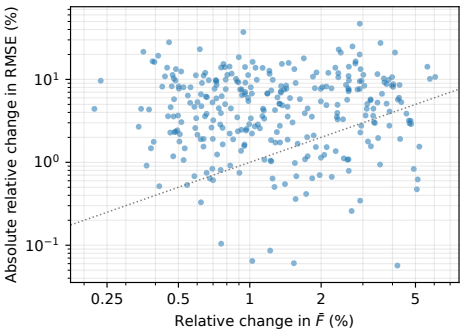


Figure 2: Each point compares a perturbed sensor with its reference sensor on L96 $d = 10$ (300 comparisons across 50 reference sensors). The axes show the relative change in the averaged Fisher matrix (Frobenius norm) and the absolute relative change in EKF RMSE, both in percent. The diagonal marks equal changes.

On KS, the derivative check holds the prediction law fixed and therefore omits the change in that law as the sensor varies. N2 concerns the full derivative and remains an assumption for the five chaotic systems. For the RMSE comparisons, we run the filter with each perturbed sensor.

Five-system comparison. Table 5 uses a separate sample of 132 base sensors, each with four signed perturbations. Within each EKF pair, we reuse the same noise vector in sensor coordinates. Even a rotation of the sensor rows can therefore change the realized RMSE while preserving the Gramian. NS2D has the smallest median relative change in $\bar{F}$ and the largest median absolute relative change in physical RMSE.

Table 5: EKF comparisons for constructed sensor pairs across five systems. n counts base sensors, each with four perturbations. Tangent intersection dimensions and blind fractions are lower bounds. Changes in $\bar{F}$ and RMSE are median absolute relative changes (%). "Ratio" is the median change in RMSE divided by the median change in $\bar{F}$.

System	(d, p, k)	$\dim T_{B_0}\mathcal{M}_p$	$\dim(T \cap \ker \pi_\mu) \geq$	blind $\geq$	n	$\Delta\bar{F}$	ΔRMSE [95% CI]	ratio
L96	$(10, 5, 5)$	40	25	62.5%	30	0.96	5.00 [3.76, 6.07]	5.2
L96	$(40, 10, 10)$	355	300	84.5%	30	0.44	1.46 [1.14, 2.25]	3.3
KS	$(64, 16, 16)$	904	768	85.0%	24	0.95	1.71 [1.41, 2.19]	1.8
NS2D	$(256, 16, 16)$	3976	3840	96.6%	24	0.17	17.98 [14.42, 23.03]	104.7
β-BVE	$(256, 16, 16)$	3976	3840	96.6%	24	0.54	3.05 [1.97, 3.87]	5.7

The percentile intervals in Table 5 describe variation across base sensors for the fitted models and trajectories. Each bootstrap draw resamples base sensors, keeping their four perturbations together (200000 draws).

Observation budget and state dimension. For a linear decoder $D(z) = Uz$ with $U^\top U = I_k$, positive-definite R gives

$$\lambda_{\min}(U^\top H^\top R^{-1} H U) = 0 \quad \Longleftrightarrow \quad \operatorname{range}(U) \cap \ker H \neq \{0\}.$$

Indeed, the quadratic form vanishes exactly when $HUv = 0$ for some $v \neq 0$. This includes PCA decoders and holds necessarily when $p < k$ by rank-nullity. Pointwise rank deficiency alone does not imply failure of dynamical detectability: distinct latent modes can become distinguishable over time.

The sensor budget p and latent dimension k play different roles. At each state, the observation Jacobian $HJ_D(z) \in \mathbb{R}^{p \times k}$ can have full column rank only when $p \geq k$. The five-system experiments

use $p = k$ to permit this rank while holding the budget fixed across decoders. Averaging over states can produce a full-rank $\bar{F}$ even when $p < k$, so the pointwise rank constraint alone does not imply that the averaged matrix is singular.

The budget also determines how many sensor perturbations can preserve the reference average to first order (Proposition D.2).

At $d = 10, k = 5$, the tangent dimensions for $p = 3$ and $p = 7$ are 27 and 49, giving lower bounds $27 - 15 = 12$ and $49 - 15 = 34$. Under the local constant-rank condition, Proposition D.2 integrates these tangent directions into exact-fibre curves. These counts concern continuously varying rank-p sensors; coordinate-subset libraries are assessed separately by their observed rankings.

At $d = 256$ and $p = k = 16$, the dimension bound covers over 96% of the tangent space on both NS2D and β-BVE. Yet their median absolute relative changes in EKF RMSE differ substantially (Table 5). The dimension count alone gives no estimate of the size of the error change. The matched-subset EnKF comparison shows the same contrast (Table 1).

Remark D.7 (Sensor-dependent observation-noise covariance). When the noise covariance varies with H, the averaged Fisher matrix is $\bar{F}(H) = \pi_\mu(H^\top R(H)^{-1} H)$, provided $R(H) \succ 0$. Two observation operators therefore have the same averaged Fisher matrix precisely when the difference between their noise-weighted Gramians lies in $\ker \pi_\mu$.

Correlated observation noise. We evaluated all 252 five-coordinate sensors on L96 $d = 10$ with a SiLU MLP under three observation-noise covariances. We used isotropic noise ($R = I_5$), diagonal noise with unequal variances ($R = \mathrm{diag}(0.25, 0.5, 1, 2, 4)$), and correlated noise with unit marginal variances ($R_{ij} = 0.6^{|i-j|}$). Across the three settings, the Spearman correlation with physical RMSE ranged from 0.86 to 0.90 for MSE_z and from 0.84 to 0.88 for $\widehat{T}_1^{\mathrm{CL}}$. The correlations for $\mathrm{tr} P^-$ and ω_{eff} had magnitude at most 0.20, as did the correlation for Fisher-trace CV, which uses variation across states. Under correlated noise, both calibration scores gave lower top-five regret than either the forecast covariance trace or the sensitivity score ω_{eff}.

Noise covariance. Whitening applies to every constant $R \succ 0$, including near-singular covariances; small eigenvalues affect numerical conditioning. For exactly singular R, the noisy subspace can be whitened, while $\ker R$ supplies noise-free observation constraints. Those constraints require a corresponding likelihood and filter formulation, so replacing R^{-1} with a pseudoinverse alone does not extend the present result.

Notation at a glance.

Object	Meaning
μ_z	fixed reference law (encoded invariant law when available)
ν_B	stationary law of $\hat{z}_{t\|t-1}$ at $B = H^\top H$
$\pi_\mu(B) = \mathbb{E}_{\mu_z}[J_D^\top B J_D]$	linear map averaging $J_D^\top B J_D$ over reference states
$\pi_\nu(B) = \mathbb{E}_{\nu_B}[J_D^\top B J_D]$	average under ν_B in DARE (6)
$P^\star(\bar{F})$	stabilizing solution of (1), when it exists (used in $\mathcal{D}$)
$P_\nu^\star(B)$	deterministic Riccati solution based on prediction-averaged information

Assumptions on the prediction law and Riccati equation (N1). Here stationarity is a hypothesis on the discrete-time EKF. Related stability results treat continuous-time diffusion filters [55] and continuous-time Kalman–Bucy filters [9, 14].

N1(a) There is an open $\mathcal{U}_0 \ni B_0$ in $\mathcal{M}_p$ and, for each $B \in \mathcal{U}_0$, a stationary law ν_B of the predicted EKF estimate $\hat{z}_{t|t-1}$ under any $H \in \mathbb{R}^{p \times d}$ with $H^\top H = B$.

N1(b) $B \mapsto \pi_\nu(B)$ is continuously Fréchet differentiable on $\mathcal{U}_0$, with derivative π_ν' at B_0.

N1(c) The expectation defining $\pi_\nu(B)$ is finite and satisfies $\pi_\nu(B) \succeq c_\nu I_k$ uniformly on $\mathcal{U}_0$, for a constant $c_\nu > 0$.

N1(d) *Local DARE invertibility*: at the stationary information matrix $J_\nu^\star(B)$, the linear operator $\mathrm{Id} - \mathrm{d}g(J_\nu^\star(B))$ on $\mathrm{Sym}(k)$ is invertible, with its inverse uniformly bounded in operator norm on $\mathcal{U}_0$.

Saturation gives a sufficient condition for N1(d). Let $\kappa(B) := \lambda_{\min}(\phi(B))/(\|Q_z^{-1}\|_{\mathrm{op}} \|M\|_{\mathrm{op}})$ (with the convention $\kappa(B) = +\infty$ when $M = 0$); if $\inf_{B \in \mathcal{U}_0} \kappa(B) > 1$, then N1(d) holds, with

the operator norm of the inverse bounded by $(1-\kappa_0^{-2})^{-1}$, where $\kappa_0 := \inf_{B\in\mathcal{U}_0}\kappa(B)$ (Proposition C.16). Condition N1(d) gives the local invertibility used to differentiate the Riccati solution in Lemma D.5.

Table 6: Riccati conditioning along the sampled sensor paths for SiLU decoders at $\sigma = 1$. Here $\sigma_{\min}$ is the smallest singular value of $\mathrm{Id} - \mathrm{d}g$. We compute the singular values in a Frobenius-orthonormal basis of $\mathrm{Sym}(k)$.

System	$\dim \mathrm{Sym}(k)$	κ range	$\sigma_{\min}$ range	$\|\mathrm{d}g\|_{F\to F}$ max
L96 d=10	15	$[0.42, 0.51]$	$[0.75, 0.78]$	0.25
L96 d=40	55	$[2.61, 3.08]$	$[0.96, 0.97]$	0.04
KS d=64	136	$\sim 2\times10^{-8}$	$\sim 1\times10^{-3}$	4.09
NS2D d=256	136	$[0.04, 0.21]$	$[0.44, 0.70]$	0.57
β-BVE d=256	136	$\sim 1\times10^{-5}$	$\sim 9\times10^{-3}$	1.005

On the sampled L96 $d = 10$ path, $\mathrm{Id} - \mathrm{d}g$ remained well conditioned even though $\kappa < 1$. On KS, its smallest singular value was much closer to zero, allowing greater amplification of perturbations in the information term by the linearized Riccati equation (Table 6). Appendix A.3 gives the local invertibility argument on a stabilizing branch.

Variance of the local Fisher matrix. In the L96 library of 252 sensors, the variance of the local Fisher matrix was weakly correlated with time-averaged spatial RMSE. The Spearman correlations ranged from -0.132 to -0.032 across the four decoder activations. For each sensor H, let $\mathcal{F}(z, H) = J_D(z)^\top H^\top R^{-1} H J_D(z)$. When $\mathcal{F}(z, H)$ has a finite second moment, its variance is $\delta_F^2(H) = \mathbb{E}_{\mu_z}\|\mathcal{F}(z, H) - \bar{F}(H)\|_F^2$.

In each of the five systems in Table 5, the median relative change in δ_F^2 was smaller than the median relative change in EKF RMSE. We normalize the absolute change in δ_F^2 by the larger of its two values and the absolute change in RMSE by the baseline RMSE.

E Experiments

We use simulations with known states to compare sensor scores. On the two real-world datasets, we use measurements from withheld channels to evaluate the filter estimates.

E.1 Simulation and filtering setup

Simulation settings. For the activation comparison on L96 with $d = 10$, we integrate Lorenz–96 at $F = 8$ using RK4 with step size 0.05. We discard 2000 simulation steps and save 50000 training states.

KS uses a periodic domain of length 22 with 64 grid points, Fourier pseudospectral discretization, and ETDRK4 integration with step size 0.25. Every integration step supplies one state snapshot, after 2000 simulation steps of spin-up.

NS2D and β-BVE use scalar vorticity on a doubly periodic $[0, 2\pi)^2$ domain with a 16×16 grid. Both use Fourier pseudospectral discretization with two-thirds dealiasing and RK4 integration. NS2D uses viscosity $\nu = 10^{-3}$, Kolmogorov forcing with velocity-forcing amplitude 1 and wavenumber 4, and integration step 0.01; every step supplies a snapshot. The β-BVE parameters are $\beta = 5$, $\nu = 10^{-4}$, linear drag $\mu = 0.05$, and random forcing with amplitude parameter 1 concentrated near wavenumber 8. Its integration step is 0.005, with one snapshot saved every ten steps, giving a sampling interval of 0.05. Both two-dimensional simulations discard 4000 integration steps before saving snapshots.

The β-BVE simulations use the sign convention

$$\partial_t\zeta + J(\psi, \zeta) - \beta\,\partial_x\psi = \nu\Delta\zeta - \mu\zeta + f, \qquad \Delta\psi = \zeta - \langle\zeta\rangle,$$

where $J(\psi, q) = \partial_x\psi\,\partial_y q - \partial_y\psi\,\partial_x q$ with $\langle\zeta\rangle$ denoting the spatial mean and $\langle\psi\rangle = 0$. The corresponding advecting velocity is $(U, V) = (-\partial_y\psi, \partial_x\psi)$. With this sign convention, $\beta = 5$ gives a background vorticity gradient $-\beta$ in these coordinates, with the transport term $J(\psi, \zeta - \beta y)$.

In these filtering sweeps, each update uses one saved state. The first 200 updates are discarded separately from simulation spin-up.

Training the L96 autoencoders. On L96 with $d = 10$, the encoder $E : \mathbb{R}^{10} \to \mathbb{R}^5$ and decoder $D : \mathbb{R}^5 \to \mathbb{R}^{10}$ each have two hidden layers of width 128 and a linear output layer. The encoder uses ReLU. We compare four decoder activations under the same training settings and random seed: ReLU, tanh, ELU with parameter 1, and SiLU. With this parameter, ELU is continuously differentiable and linear for $x > 0$. SiLU is given by $x\sigma(x)$ [15]. The latent predictor is $g_\theta(z) = Mz$, with $M \in \mathbb{R}^{5\times5}$ initialized to the identity and learned during training. Appendix E.3 compares eight activations with five training runs per activation.

$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}}$ with $\lambda_{\text{pred}} = 1$, where $\mathcal{L}_{\text{recon}} = \|x - D(E(x))\|^2$ and $\mathcal{L}_{\text{pred}} = \|E(x_{t+1}) - ME(x_t)\|^2$.

We train for 300 epochs with Adam on 50000 training states, using a learning rate of 10^{-3}, batches of 256, and gradient clipping at norm one. The training routine does not apply a spectral-radius projection. The evaluated L96 models satisfy $\rho(M) < 1$. Appendix A.3 summarizes which fitted models satisfy this stability condition.

Sensor libraries and error measures. We evaluate all $\binom{10}{5} = 252$ subsets of five L96 coordinates, using one observation matrix H per subset. Each filter run uses $R = I_5$, 2000 time steps, and a 200-step burn-in. For the four-activation results in Tables 2 and 9, the within-MLP sweep, and the architecture ablation in App. E.3, let $\mathcal{I}$ contain the $N = 1800$ retained time indices and define

$$q_H = \frac{1}{Nd}\sum_{t\in\mathcal{I}} \|x_t - D(\hat{z}_{t|t}(H))\|^2, \qquad m_H = \frac{1}{Nk}\sum_{t\in\mathcal{I}} \|E(x_t) - \hat{z}_{t|t}(H)\|^2.$$

For these comparisons, physical RMSE is $\sqrt{q_H}$, pooling squared errors over time and coordinates before taking the square root. The corresponding vector squared errors are dq_H in physical space and $km_H = \operatorname{tr}\widehat{\Sigma}^{\text{err}}_H$ in latent space, in the units used in §4. The constant factors d and k preserve the rankings within a model. Some extended analyses instead use $r_H = N^{-1}\sum_{t\in\mathcal{I}} \|x_t - D(\hat{z}_{t|t}(H))\|/\sqrt{d}$, the spatial RMSE averaged over time. This change in the order of averaging can change the sensor ranking.

EKF. At each step, the predicted observation is $HD(\hat{z}_{t|t-1})$. We compute $C_t = HJ_D(\hat{z}_{t|t-1})$ by automatic differentiation. The innovation covariance is $S_t = C_tP_{t|t-1}C_t^\top + R$, giving gain $K_t = P_{t|t-1}C_t^\top S_t^{-1}$. The filter adds $K_t(y_t - HD(\hat{z}_{t|t-1}))$ to the predicted state and updates the covariance to $(I - K_tC_t)P_{t|t-1}$. The next forecast uses $\hat{z}_{t+1|t} = M\hat{z}_{t|t}$ and $P_{t+1|t} = MP_{t|t}M^\top + Q_z$.

EnKF. For the constructed sensor pairs, the EKF and EnKF use the same true trajectory and generate observation noise within each filter run. The stochastic EnKF uses 100 members and perturbs their observations. It applies HD to each forecast member and estimates the gain from ensemble covariances. The ensemble size exceeds each latent dimension, and we use no localization.

Forecast and posterior scores. The prediction covariance P^- is obtained from the covariance DARE

$$P^- = M\big((P^-)^{-1} + \bar{F}\big)^{-1}M^\top + Q_z.$$

Applying the observation update gives the posterior covariance $P^\star = ((P^-)^{-1} + \bar{F})^{-1}$. The activation, architecture, noise-level, and correlated-noise comparisons and the calibration-trajectory shortlist comparison in Table 14 report $\operatorname{tr}P^-$. The real-data, constructed-pair, and finite-library studies use posterior DARE solutions, as does the independent-test comparison in Table 3.

Fisher-trace CV. Fix the decoder, sensor H, and observation-noise covariance R. For encoded trajectory states $z_1, \dots, z_n$, write

$$t_i = \operatorname{tr}\big(J_D(z_i)^\top H^\top R^{-1} H J_D(z_i)\big), \qquad \bar{t} = \frac{1}{n}\sum_{i=1}^n t_i, \qquad s_t^2 = \frac{1}{n}\sum_{i=1}^n (t_i - \bar{t})^2.$$

Here J_D is expressed in physical output coordinates, including the inverse standardization of the trained decoder. Under the empirical distribution assigning mass $1/n$ to each sampled state, the trace has coefficient of variation $s_t/\bar{t}$ for $\bar{t} > 0$. We compute $s_t/(\bar{t} + 10^{-15})$ for numerical stability.

At zero mean the standard CV is undefined, whereas this numerical convention returns zero. Trace-CV and FisherCV are abbreviations for this score in the tables. We sample reference states uniformly without replacement.

E.2 Sensor selection in a linear system

Setup. Can a reduced filter misestimate its error and still identify a good sensor? We retain three state variables z_t and omit nine variables u_t. The coupling γ controls how strongly the omitted variables enter the retained dynamics, and b controls their contribution to the observations. Write $x_t = (z_t, u_t) \in \mathbb{R}^{12}$, with $z_t \in \mathbb{R}^3$, encoder $E = [I_3, 0]$, and decoder $D = E^\top$. The dynamics and observation matrix are

$$A = \begin{bmatrix} M & \gamma B \\ 0 & N \end{bmatrix}, \qquad H = [\sqrt{1-b^2}\, C,\ bV], \qquad Q_x = 0.01 I_{12}, \quad R = 0.1 I_3.$$

The matrices M and N have independent random orthogonal eigenbases, with spectra $(0.70, 0.85, 0.95)$ and nine equally spaced values from 0.30 to 0.98, respectively. The Gaussian matrix $B \in \mathbb{R}^{3\times 9}$ is normalized to unit spectral norm. Each of six independent system instances has 64 candidates, generated from Gaussian $C \in \mathbb{R}^{3\times 3}$ and $V \in \mathbb{R}^{3\times 9}$ with unit-norm rows. The same C, V library is used in four cases: closed $(\gamma, b) = (0, 0)$, dynamic mismatch $(0.2, 0)$, observation mismatch $(0, 0.5)$, and combined mismatch $(0.2, 0.5)$. Every sensor has three unit-norm rows.

Reference and actual errors. The reference filter uses dynamics M, observation matrix HD, and the exact marginal covariance of the projection residual,

$$Q_{\mathrm{ref}} = 0.01 I_3 + \gamma^2 B \Sigma_u B^\top,$$

where Σ_u is the stationary covariance of u_t. The omitted state evolves over time, so its contribution to the retained dynamics can remain correlated across steps even with this choice of Q_{ref}. The model score is the posterior trace $\mathrm{tr}P^+$, obtained by updating the prediction covariance P^- with the current observation. We compute the actual stationary error covariance independently from the joint linear dynamics of the state and filter estimate. The observation covariance R contains the instrument noise; the contribution bVu_t remains an omitted component of the projected observation model. Orthogonal decoding gives $\mathrm{FE}(H) = \mathrm{tr}\Sigma_u + \mathrm{MSE}_z(H)$ in all four cases. In the closed case, the latent filter is correctly specified and $\mathrm{MSE}_z(H) = \mathrm{tr}P^+(H)$.

Sensor selection. Both scores evaluate the same reduced filter. Calibration uses $2{,}000$ samples of latent squared error per candidate. Each score selects five candidates, evaluated by their best physical RMSE on an independent $10{,}000$-step test trajectory. Regret is the difference from the best candidate in the full test library. Each trajectory discards $2{,}000$ initial steps; candidates share the truth and observation-noise streams within a trajectory. We average three independent calibration/test pairs within each system and report means and sample standard deviations across six systems. The random baseline is the expected regret from selecting five candidates uniformly without replacement.

Table 7: Top-five regret on independent test trajectories in the linear projection experiment (10^{-3} physical RMSE units). RMSE uses the squared vector norm averaged over time, without division by dimension, followed by the square root. Entries give the mean $\pm$ sample standard deviation across six systems after averaging three trajectory pairs within each system.

Case	Posterior Riccati	Latent-error calibration	Random
Closed	0	0.091 ± 0.120	2.508 ± 0.315
Dynamic mismatch	0.035 ± 0.054	0.042 ± 0.051	4.572 ± 1.910
Observation mismatch	0.117 ± 0.145	0.036 ± 0.089	4.349 ± 0.643
Combined mismatch	1.091 ± 2.246	0.096 ± 0.235	9.252 ± 3.843

Results. In the closed case, the posterior covariance agrees with the actual latent error covariance to numerical precision. Under dynamic mismatch the median relative discrepancy across candidates, $|\mathrm{tr}P^+ - \mathrm{MSE}_z|/\mathrm{MSE}_z$, averages 19.3% across systems. In every system the five candidates selected by the posterior trace include the sensor with the lowest stationary error. On a finite test trajectory, another sensor can have lower measured RMSE, which accounts for the small positive test regret. Calibration lowers mean test regret under observation and combined mismatch (Table 7). One of the six systems accounts for nearly all of the average gain under combined mismatch.

E.3 Decoder geometry and sensor rankings

A decoder can stretch the same latent displacement by different amounts at different states. We compare this variation across trained decoders with the agreement between latent and physical sensor rankings. For the activation and architecture sweeps, let D_{norm} denote the decoder in standardized output coordinates. At encoded states z_i, set $G_i = G_{D_{\mathrm{norm}}}(z_i) = J_{D_{\mathrm{norm}}}(z_i)^\top J_{D_{\mathrm{norm}}}(z_i)$ and $\bar{G} = N_J^{-1} \sum_i G_i$. We report the relative metric fluctuation

$$\widehat{\delta}_{D,\mathrm{rel}} = \frac{\left(N_J^{-1} \sum_i \|G_i - \bar{G}\|_F^2\right)^{1/2}}{\|\bar{G}\|_F}.$$

Writing the physical decoder as $D_{\mathrm{phys}}(z) = \mu + S D_{\mathrm{norm}}(z)$, with diagonal channel-scale matrix S, gives $J_{D_{\mathrm{phys}}} = S J_{D_{\mathrm{norm}}}$ and physical metric $J_{D_{\mathrm{norm}}}^\top S^2 J_{D_{\mathrm{norm}}}$. The theoretical δ_D in §4 is the unnormalized fluctuation of that physical metric.

For the Hessian estimates, let $h_i = \|\operatorname{unfold}_1(\nabla^2 D_{\mathrm{norm}}(z_i))\|_{\mathrm{op}}$, where the unfolding has shape $d \times k^2$. We report $\widehat{\eta}^{(1)}_{\mathrm{mean}} = N_H^{-1} \sum_i h_i$ and $\widehat{\eta}^{(1)}_{\max} = \max_i h_i$. At each state, this matrix norm bounds the bilinear Hessian norm; the theoretical η bounds the physical decoder's bilinear Hessian norm throughout Ω. The residual bi-Lipschitz comparison instead reports $\widehat{\eta}_{F,\max} = \max_i \|\nabla^2 D_{\mathrm{norm}}(z_i)\|_F$.

The main comparison of four activations uses 5000 training states for the metric and 200 for the Hessian. For the extended activation and 1D architecture comparisons, the corresponding test sample sizes are 2000 and 100. For the within-MLP comparison, we use 2000 test states for the metric and 80 for the Hessian. On the 2D systems we estimate the metric from 100 test states, except for one NS2D MLP run with 200. That run provides the only 2D architecture Hessian estimate, based on five states.

For the four-activation comparisons, the within-MLP sweep, and the architecture ablation, define the per-coordinate reconstruction MSE $e_r^2 = (Nd)^{-1} \sum_{t \in \mathcal{I}} \|x_t - D(E(x_t))\|^2$ on the same retained window and use

$$\widehat{\alpha}(H) = \frac{d(q_H - e_r^2)}{k\, m_H}.$$

We compute its mean, variance, and CV over each complete sensor library, using the library size as the variance divisor (252 for the four-activation comparisons).

We hold each trained decoder fixed. For geometry, we resample the training states used for each estimate, with 500 bootstrap draws. For conversion-rate CV and ranking agreement ρ_H, we resample the 252 sensors on the same evaluation trajectory, with 200 and 1000 draws, respectively. Each sensor retains its paired latent and physical errors. ELU and SiLU give closer ranking agreement than ReLU and tanh in this comparison (Table 8).

Table 8: Decoder geometry and sensor rankings by pooled physical RMSE on L96 $d = 10$. Brackets give 95% percentile bootstrap intervals for each fixed model.

Activation	$\widehat{\delta}_{D,\mathrm{rel}}$	$\widehat{\eta}^{(1)}_{\mathrm{mean}}$	$\widehat{\alpha}$ CV	ρ_H
ReLU	0.666 [0.661, 0.671]	0 (p.w. affine)	0.163 [0.146, 0.180]	0.849 [0.802, 0.885]
tanh	0.677 [0.668, 0.685]	121 [114, 127]	0.341 [0.316, 0.362]	0.851 [0.807, 0.884]
ELU	0.601 [0.597, 0.606]	101 [97, 106]	0.135 [0.121, 0.147]	0.925 [0.898, 0.942]
SiLU	0.602 [0.599, 0.606]	102 [98, 105]	0.130 [0.112, 0.148]	0.944 [0.924, 0.958]

Variation across trained models. We train each activation five times, with matched seeds across activations and a denoising reconstruction loss. We use batches of 256 and add Gaussian noise with standard deviation 0.1 to the standardized encoder input for the reconstruction loss; the prediction loss uses clean inputs. The four-activation comparison in Table 2 uses the same batch size and no input noise. Both ELU and SiLU have lower conversion-rate variance than either ReLU or tanh in all five paired repetitions (Table 9).

Across 108 SiLU models, greater variation in the decoder metric was associated with weaker agreement between latent and physical error rankings (Spearman $\rho = -0.791$) and greater variance in

Table 9: Variance of the conversion rate $\widehat{\alpha}$ across 252 sensors in each of five L96 $d = 10$ training repetitions. The last column divides the mean variance for ELU and SiLU by the mean variance for ReLU and tanh.

Repetition	ReLU	tanh	ELU	SiLU	Variance ratio
1	4573.03	2988.23	1867.31	1981.45	0.509
2	9026.88	4582.71	2576.46	2360.37	0.363
3	9226.73	2841.10	1620.02	1914.17	0.293
4	7584.02	4442.77	1445.95	1442.86	0.240
5	7139.93	3321.87	1189.09	2738.51	0.375

the conversion rate across sensors ($\rho = 0.763$). We trained each of 36 configurations three times, varying depth, width, and weight decay, and evaluated every model on the same 252 sensors.

Changing the decoder architecture. Latent and physical error rankings agreed least closely for the Transformer on both L96 systems and NS2D, and most closely for the Transformer on β-BVE (Table 10). On KS, all four architectures gave near-perfect ranking agreement in the single run per architecture. We compared architectures holding activation (SiLU), encoder (the ReLU MLP in Appendix E.1 on L96 $d = 10$), training settings (300 epochs using Adam with learning rate 10^{-3} and batch size 256, denoising $\sigma = 0.1$ on L96 d=10), and trajectories constant. The reconstruction RMSE column in Table 10 uses all 2000 states. For the conversion rate, we compute reconstruction and filtering errors on the same 1800 post-burn-in states. Fisher scores use 500 reference states per sensor on the 1D systems and 200 common reference states on the 2D systems; the geometry sample sizes are given above.

Table 10 summarizes five training runs per architecture on L96 with $d = 10$ and one per architecture on KS. On L96 with $d = 40$, NS2D, and β-BVE, we have four runs for the Transformer and five for each other architecture. The MLP maps the latent vector directly to the output through dense layers. The convolutional and attention models first project it into a spatial feature array.

- **MLP.** On L96 $d = 10$, the baseline has two width-128 SiLU hidden layers and a linear output.
- **CNN1D**: input projection $k \to (c_{\text{hidden}}, d)$ followed by three `Conv1d` layers with circular padding and kernel size 3. Circular padding exploits the ring topology of L96 sites.
- **ResNet1D**: same input projection as CNN1D followed by two residual blocks $x \mapsto x + W_2 \cdot \sigma(W_1 x)$ where W_1 and W_2 are `Conv1d` layers with circular padding and kernel size 3.
- **Transformer**: same input projection, followed by two pre-LayerNorm encoder blocks (multi-head self-attention with 4 heads + MLP of ratio 2). The positional encoding is rotary (*RoPE*) with d-periodic rotation angles $\theta_{j,i} = 2\pi\, j\, f_i/d$ for integer frequencies $f_i \in \{1, \ldots, \text{head_dim}/2\}$, making attention exactly equivariant under cyclic shifts of the length-d sequence and matching the L96 ring symmetry.

On L96 $d = 10$, all three non-MLP decoders use $c_{\text{hidden}} = 32$, giving decoder parameter counts in the same order of magnitude ($\sim$ 10–20k) across architectures. We use hidden width 64 for the non-MLP decoders on L96 $d = 40$ and KS, and 16 on the two vorticity systems. On the two-dimensional grids, convolution uses circular padding on both axes and attention uses two-dimensional periodic rotary position encoding.

Denoising on L96 $d = 40$. We retrained the Transformer with Gaussian input noise of standard deviation 0.1, as used on L96 $d = 10$ [1], and compared it with the model trained on clean inputs. In this comparison we rank sensors by time-averaged spatial RMSE r_H. Denoising reduced the sampled maximum Hessian norm by 62%, but ranking agreement rose only from 0.35 to 0.42. The other three architectures gave closer agreement on this target.

On β-BVE, the Transformer ranks highest in all four shared training repetitions. Reconstruction errors are similar across architectures here, so reconstruction accuracy alone gives little indication of which latent errors will best track physical errors.

Table 10: Decoder geometry, sensor-ranking agreement, and top-five regret across five systems. Parentheses give sample standard deviations across training runs. Within each shared training run, we also compared $\widehat{\delta}_{D,\mathrm{rel}}$ and ρ_H across the four architectures. Their Spearman correlation was negative in every run on L96 $d = 10$, NS2D, and β-BVE, and varied in sign on L96 $d = 40$. Regret is in physical RMSE units. Random gives the expected regret of five sensors drawn uniformly without replacement.

System	Arch.	$\widehat{\delta}_{D,\mathrm{rel}}$	$e_{r,\mathrm{all}}$	α CV	ρ_H	$\mathrm{Reg}_5(\mathrm{MSE}_z)$	Reg_5(rand.)
L96 d=10	MLP	0.591(0.015)	0.830(0.020)	0.156(0.022)	0.894(0.031)	0.013(0.029)	0.374(0.007)
	CNN1D	0.694(0.010)	0.802(0.014)	0.169(0.019)	0.881(0.018)	0.032(0.045)	0.274(0.027)
	ResNet1D	0.664(0.021)	0.796(0.010)	0.206(0.015)	0.894(0.022)	0.020(0.045)	0.300(0.049)
	Transformer	1.008(0.036)	0.783(0.013)	0.217(0.019)	0.747(0.060)	0.024(0.033)	0.281(0.065)
L96 d=40	MLP	0.788(0.067)	2.127(0.066)	0.210(0.096)	0.733(0.048)	0.087(0.081)	0.331(0.052)
	CNN1D	1.090(0.073)	2.163(0.143)	0.112(0.017)	0.877(0.024)	0.006(0.013)	0.110(0.040)
	ResNet1D	1.059(0.047)	2.166(0.090)	0.113(0.024)	0.857(0.047)	0.015(0.020)	0.099(0.012)
	Transformer	1.060(0.044)	2.677(0.028)	0.133(0.012)	0.377(0.068)	0.036(0.021)	0.067(0.016)
KS	MLP	0.230	0.025	0.019	0.999	0.000	0.018
	CNN1D	0.183	0.025	0.028	0.998	0.000	0.020
	ResNet1D	0.167	0.024	0.022	0.998	0.000	0.021
	Transformer	0.243	0.026	0.032	0.999	0.000	0.019
NS2D	MLP	0.539(0.005)	12.501(0.072)	0.236(0.028)	0.902(0.024)	0.226(0.346)	3.353(0.511)
	CNN2D	0.464(0.012)	11.104(0.081)	0.158(0.028)	0.953(0.013)	0.000(0.000)	2.921(0.616)
	ResNet2D	0.497(0.028)	11.032(0.064)	0.119(0.010)	0.973(0.006)	0.000(0.000)	3.035(0.709)
	Transformer2D	0.801(0.065)	11.044(0.257)	0.176(0.078)	0.408(0.294)	0.121(0.060)	0.220(0.057)
β-BVE	MLP	0.555(0.019)	0.134(0.001)	0.131(0.010)	0.860(0.025)	0.001(0.001)	0.009(0.001)
	CNN2D	0.495(0.015)	0.135(0.003)	0.131(0.009)	0.885(0.020)	0.002(0.003)	0.008(0.001)
	ResNet2D	0.522(0.014)	0.134(0.002)	0.133(0.008)	0.876(0.019)	0.001(0.001)	0.008(0.002)
	Transformer2D	0.454(0.041)	0.133(0.002)	0.100(0.015)	0.933(0.010)	0.000(0.000)	0.008(0.003)

Runs per architecture: five on L96 $d = 10$, one on KS, and five elsewhere (four for the Transformer).

Extended comparisons. Across 40 L96 models trained with eight decoder activations, greater variation in the decoder metric was associated with more variation in the conversion rate across sensors (Spearman $\rho = 0.56$) and weaker agreement between latent and physical error rankings ($\rho = -0.63$). Across 13 hypotheses, the Holm-adjusted p-values were 0.011 and 0.0026, respectively, and the Benjamini–Yekutieli q-values were 0.019 and 0.0083.

In the extended activation comparisons, we compute r_H by averaging spatial RMSE over the 1800 steps after burn-in. The per-coordinate latent MSE m_H uses the same steps, and we compute reconstruction RMSE $e_{r,\mathrm{all}}$ over all 2000 steps. The empirical conversion rate is $\widehat{\alpha}_{\mathrm{emp}}(H) = \max\{r_H^2 - e_{r,\mathrm{all}}^2, 10^{-10}\}/m_H$. The numerator floor was inactive in the four-activation comparisons on L96 and KS and in the eight-activation experiment.

Across four activations on each of three systems, we fitted a mixed-effects model with a system-specific intercept and slope [6]. The fitted mean slope of $\mathrm{CV}_H(\widehat{\alpha}_{\mathrm{emp}})$ against $\widehat{\delta}_{D,\mathrm{rel}}$ was 0.313 (model-based 95% Wald interval $[0.078, 0.547]$). For ranking agreement ρ_H, the interval spanned zero ($[-0.745, 0.200]$).

E.4 Decoder distance bounds and residual alignment

Even for the L96 decoders with the closest agreement between latent and physical error rankings (Table 2), the reconstruction residual had a substantial component along the decoder tangent space. We measure local stretching and alignment in standardized output coordinates. The distance and alignment assumptions in (7) and (10) concern finite displacements throughout the domain.

At 2000 sampled attractor points we compute the smallest singular value of the decoder Jacobian. At differentiable interior points, a lower distance bound implies the same lower bound on this singular value in the corresponding coordinates. To measure alignment, let u be the standardized input, z its encoding, and $e = u - D_{\mathrm{norm}}(z)$ the reconstruction residual. For nonzero e, its relative projection

onto the tangent space is

$$\hat{\varepsilon}_\perp(u) = \frac{\|P_{\mathrm{range}(J_{D_{\mathrm{norm}}}(z))}\, e\|}{\|e\|}.$$

We summarize the projection over 1000 sampled states.

Act.	ε_r	$\sigma_{\min}(J_D)$ min	$\sigma_{\min}(J_D)$ p5	$\varepsilon_\perp$ med	$\varepsilon_\perp$ p95	$\varepsilon_\perp$ max
ReLU	0.83	1.70	3.90	0.52	0.80	0.97
tanh	0.85	1.36	2.89	0.46	0.77	0.97
ELU	0.86	1.95	3.09	0.46	0.77	0.95
SiLU	0.83	1.79	3.38	0.47	0.79	0.97

A residual decoder with global distance bounds. On L96 with $d = 10$, the median tangent projection for the tanh MLP was about 35 times the mean of the three residual decoders' medians. We obtain global distance bounds for the residual decoder from its architecture, following BiLipNet and iResNet [7, 54]. In standardized output coordinates it takes the form

$$D_{\mathrm{norm}}(z) = W_{\mathrm{out}}\big(\mathrm{block}_L \circ \cdots \circ \mathrm{block}_1\big)(z) + b, \quad \mathrm{block}_i(h) = h + a_{\mathrm{res}}\, \widetilde{W}_{i,2}\, \mathrm{SiLU}(\widetilde{W}_{i,1} h + b_{i,1}),$$

where $W_{\mathrm{out}} = \sigma Q$, $\sigma > 0$, and Q has orthonormal columns. After training, we rescaled the residual-layer weights using their largest singular values so that $\|\widetilde{W}_{i,j}\|_{\mathrm{op}} \le 1$. Each residual branch then has Lipschitz constant at most $a_{\mathrm{res}} L_\phi$, where $L_\phi = \sup_x |\mathrm{SiLU}'(x)| \approx 1.10$. With $a_{\mathrm{res}} = 0.5$, this product is less than one. Each block scales distances by a factor between $1 - a_{\mathrm{res}} L_\phi$ and $1 + a_{\mathrm{res}} L_\phi$. Composing $L = 3$ blocks and applying W_{out} gives

$$\ell_D^{\mathrm{arch}} \;=\; \sigma\,(1 - a_{\mathrm{res}} L_\phi)^L, \qquad \ell_U^{\mathrm{arch}} \;=\; \sigma\,(1 + a_{\mathrm{res}} L_\phi)^L. \tag{48}$$

For physical outputs $D_{\mathrm{phys}} = \mu + S D_{\mathrm{norm}}$, the lower and upper bounds are $\min_j S_{jj}\, \ell_D^{\mathrm{arch}}$ and $\max_j S_{jj}\, \ell_U^{\mathrm{arch}}$, respectively. The smooth SiLU activation has bounded first and second derivatives, so the fitted decoder also has bounded Hessian.

The table reports the residual decoder's analytical lower bound and the MLPs' sampled minimum singular values. We measure Hessian Frobenius norms on 100 states per model and use the same sampled states for the residual decoder's tangent projections. Residual-decoder entries give means and sample standard deviations across the three runs.

Decoder	ℓ_D^{arch}	Sample min $\sigma_{\min}(J_D)$	$\varepsilon_\perp$ med	$\varepsilon_\perp$ max	$\widehat{\eta}_{F,\max}$
ReLU (MLP)	—	1.42	0.52	0.97	0 (a.e.)
tanh (MLP)	—	1.13	0.46	0.97	346
ELU (MLP)	—	1.93	0.46	0.95	253
SiLU (MLP)	—	1.49	0.47	0.97	216
Res. bi-Lip (n=3)	**0.78±0.13**	—	**0.013±0.004**	**0.050±0.016**	**15.5±2.5**

The standardized upper bound across the three runs was $\ell_U^{\mathrm{arch}} = 31.9 \pm 5.4$.

E.5 Fisher scores and filtering error

With $M, Q_z, \overline{G}_D$ and the reference measure fixed independently of H, $T_1^{\mathrm{LG}} = \mathrm{tr}(\overline{G}_D P^\star(\bar{F}))$ factors through $\bar{F}$ and therefore belongs to $\mathcal{D}$ on its stabilizing domain. In Proposition C.14, $\overline{G}_D = D_0^\top D_0$ and $\mathbb{E}\|D_0(z - \hat{z})\|^2 = T_1^{\mathrm{LG}}$. Its monotonicity also follows from the covariance recursion: prediction preserves the positive-semidefinite order, while $P \mapsto (P^{-1} + F)^{-1}$ is increasing in P and decreasing in F. Starting the two recursions from the same covariance and taking their stabilizing limits gives $P^\star(F_2) \preceq P^\star(F_1)$ for $F_2 \succeq F_1$, hence the same inequality for the weighted traces. These reference-model properties do not identify true nonlinear EKF error.

Constructed sensor pairs. Weighting the posterior Riccati covariance by decoder geometry gave a smaller median change than either latent MSE or physical RMSE in every system (Table 11). We used the sensor pairs from Table 5, physical-coordinate decoder Jacobians, a common reference sample within each pair, and a steady-state posterior covariance computed separately for each sensor.

Resampling whole base sensors gave disjoint 95% intervals for the model-score median and each error median in all five systems. The intervals describe variation across pairs for the fitted models and trajectories.

Table 11: Median absolute percentage changes in the posterior Riccati score, latent MSE, and time-averaged spatial EKF RMSE. Table 1 gives the corresponding percentage changes in $\bar{F}$ for the same EKF pairs.

System	T_1^{LG}	MSE_z	EKF RMSE
L96 $d = 10$	0.426	15.67	5.00
L96 $d = 40$	0.082	6.47	1.46
KS	0.166	3.62	1.71
NS2D	0.017	33.47	17.98
β-BVE	0.035	10.30	3.05

Rankings within fixed libraries. On the evaluation trajectory, latent MSE ranked sensor pairs more accurately than effective sensitivity in 79 of 81 trained models. The median within-model reduction in the fraction of misranked pairs was 28.4 percentage points.

For each model and score, we exclude pairs tied in the score or in pooled physical RMSE. Among the remaining pairs, we count the fraction for which the score favors the sensor with higher RMSE (Table 12). We favor higher ω_{eff} and lower MSE_z or Fisher-trace CV.

Table 12: Percentage of eligible sensor pairs for which the score favors the sensor with higher pooled physical RMSE. Entries give the median [minimum, maximum] across trained models within each system. Latent MSE uses reference states from the same evaluation trajectory.

System	Models	ω_{eff}	MSE_z	Fisher-trace CV
L96 $d = 10$	20	51.8 [46.8, 60.6]	15.3 [10.7, 26.5]	49.9 [43.2, 58.9]
L96 $d = 40$	19	42.9 [25.2, 46.1]	20.1 [12.9, 39.2]	43.0 [35.9, 53.9]
KS	4	20.5 [15.2, 37.0]	1.3 [1.2, 1.6]	15.7 [14.9, 16.0]
NS2D	19	49.1 [42.1, 56.6]	10.5 [5.9, 50.8]	50.3 [39.2, 58.5]
β-BVE	19	30.3 [26.0, 41.5]	14.8 [9.8, 18.3]	30.5 [24.9, 41.1]

Appendix E.6 evaluates sensor selection on independent trajectories.

Pairs with similar Fisher matrices. Across the 13 model-specific sensor libraries, the smallest relative distance between averaged Fisher matrices was 3.48% on NS2D, above the prespecified threshold of 1.2%.

For each trained SiLU MLP, we compared all unordered pairs in its library using

$$d_F(i,j) = \frac{\|\hat{\bar{F}}_i - \hat{\bar{F}}_j\|_F}{(\|\hat{\bar{F}}_i\|_F + \|\hat{\bar{F}}_j\|_F)/2}.$$

For each model we estimated the Fisher matrices using physical-coordinate decoder Jacobians, $R = I_p$, and a common sample from the training trajectory (500 states on L96 and KS, 200 on the two-dimensional systems). We used one model on KS and three on each of the other systems. L96 $d = 10$ includes all 252 five-coordinate subsets. On L96 $d = 40$ and KS, each library contains 100 uniformly sampled, 50 spaced, and 50 clustered configurations. On NS2D and β-BVE, each contains 50 uniform, 50 subgrid, and 50 clustered configurations. Appendix D.2 gives the separate construction of sensor pairs with nearly equal averaged Fisher matrices.

Sensitivity to observation noise. We vary the observation noise while holding the SiLU decoder and all 252 L96 $d = 10$ sensor configurations fixed. Both calibration scores track physical RMSE across the four noise levels. At each level, their five candidates also include a sensor with lower RMSE than the best candidate selected by any of the three comparison scores (Table 13). Both this scan and the correlated-noise comparison in Appendix D.2 use 500 evaluation states for the Fisher diagnostics.

Table 13: L96 $d = 10$ noise-level scan. At each σ, we rank all 252 sensors and evaluate top-five regret on the same trajectory using time-averaged spatial RMSE r_H.

	Spearman$(\cdot, \mathrm{RMSE})$					Top-5 regret				
σ	MSE_z	$\widehat{T}_1^{\mathrm{CL}}$	$\mathrm{tr}P^-$	ω_{eff}	FisherCV	MSE_z	$\widehat{T}_1^{\mathrm{CL}}$	$\mathrm{tr}P^-$	ω_{eff}	FisherCV
0.25	+0.89	+0.85	−0.02	−0.15	+0.24	0.000	0.000	0.376	0.583	0.397
0.50	+0.89	+0.86	−0.15	−0.10	+0.18	0.000	0.000	0.438	0.232	0.198
1.00	+0.87	+0.85	−0.13	−0.06	+0.18	0.086	0.086	0.658	0.159	0.305
2.00	+0.87	+0.86	−0.05	−0.15	+0.14	0.000	0.000	0.280	0.164	0.141

E.6 Sensor selection on independent trajectories

We choose the candidates on a calibration trajectory, then run the same sensors on an independent test trajectory.

Score information and coordinates. The calibration score $\widehat{T}_1^{\mathrm{CL}}$ weights the uncentered second moment of the latent EKF errors by $W = N_{\mathrm{ref}}^{-1} \sum_j J_{D_{\mathrm{norm}}}(z_j)^\top J_{D_{\mathrm{norm}}}(z_j)$. All sensors share this weight, computed in standardized output coordinates. Full-state calibration samples provide the reference states for W and the encoded states for measuring latent error. The posterior model scores use physical-coordinate decoder Jacobians. On each test trajectory, the true states provide the common filter initialization and the reference for measuring physical error.

Experimental setup. We evaluate three independently trained SiLU MLP models on each of L96 $d = 10$, L96 $d = 40$, NS2D and β-BVE, with up to five calibration/test trajectory pairs per model. This gives 15 model–trajectory combinations on L96 $d = 10$, NS2D and β-BVE, and 14 on L96 $d = 40$. KS uses one model and one trajectory pair. The libraries contain 252 sensors on L96 $d = 10$, 200 on L96 $d = 40$ and KS, and 150 on NS2D and β-BVE. Calibration trajectories determine the latent-error scores and process covariance. Each calibration or test filter starts from the encoded initial state of its trajectory with covariance $0.1I$. Each calibration trajectory and each test trajectory contains 10000 snapshots (2000 for KS), with sampling intervals for the PDE systems given in App. E.1. Each filter run discards the first 200 steps, leaving $N = 9800$ (1800 for KS) error samples. With $\mathcal{I}_{\mathrm{test}}$ denoting the retained time indices, physical RMSE is

$$\mathrm{RMSE}^{\mathrm{test}}(H) = \left[\frac{1}{Nd} \sum_{t \in \mathcal{I}_{\mathrm{test}}} \| x_t - D(\hat{z}_{t|t}(H)) \|^2 \right]^{1/2}. \tag{49}$$

Calibration RMSE is computed by the same expression on the calibration trajectory. Calibration regret compares the best sensor in the shortlist with the best sensor in the full library, using RMSE on the same calibration trajectory. For a score s with calibration shortlist $\mathrm{top}_K(s)$, the split regret is

$$\mathrm{Regret}_K^{\mathrm{split}}(s) = \min_{H \in \mathrm{top}_K(s)} \mathrm{RMSE}^{\mathrm{test}}(H) - \min_{H \in \mathcal{H}} \mathrm{RMSE}^{\mathrm{test}}(H).$$

The random baseline averages regret over uniformly sampled size-K subsets without replacement, computed exactly.

Five candidates per shortlist. On the calibration trajectories, both latent-error scores gave lower mean regret on the two L96 systems than the forecast covariance trace or either Fisher diagnostic (Table 14). For independent tests, Table 3 compares the calibration scores with the two posterior covariance scores. For each model, all sensors share a training reference sample for the averaged Fisher matrix (500 states on L96 and KS, 200 on the two-dimensional systems). The process covariance and physical decoder metric for the posterior scores use the calibration trajectory.

Averaging trajectory pairs within each trained model, both calibration scores have lower mean top-5 regret than both posterior scores for all three models on L96 $d = 40$, NS2D and β-BVE, and for two of the three models on L96 $d = 10$. The remaining L96 $d = 10$ model favors the posterior scores.

Table 15 reports independent-test regret across shortlist sizes.

Table 14: Top-5 regret on the calibration trajectories, measured in physical RMSE units. Entries are means (sample standard deviations) over 15 model–trajectory combinations per system, except for L96 $d = 40$ (14 combinations) and KS (one model and one trajectory pair).

Diagnostic	L96 d=10	L96 d=40	KS	NS2D	β-BVE
MSE_z	**0.000(0.000)**	**0.048(0.072)**	**0.000**	**0.031(0.122)**	**0.0014(0.0021)**
$\widehat{T}_1^{\mathrm{CL}}$	**0.016(0.043)**	**0.041(0.069)**	**0.000**	**0.020(0.077)**	**0.0006(0.0011)**
$\mathrm{tr}P^-$	0.215(0.123)	0.316(0.164)	0.000	4.719(2.301)	0.0117(0.0076)
ω_{eff}	0.266(0.116)	0.178(0.168)	0.002	5.916(2.339)	0.0171(0.0077)
Fisher-trace CV	0.314(0.163)	0.310(0.378)	0.003	4.557(1.962)	0.0063(0.0025)
Random baseline	0.297(0.044)	0.295(0.040)	0.018	3.852(0.833)	0.0073(0.0019)

Table 15: Independent-test regret in physical RMSE units for shortlists of size K. Entries give means (sample standard deviations) across 15 model–trajectory combinations per system, with 14 on L96 $d = 40$ and one on KS. Random gives the expected regret of K sensors drawn uniformly without replacement.

System	Diagnostic	$k = 1$	$k = 5$	$k = 10$	$k = 20$
L96 $d = 10$	MSE_z	0.118(0.122)	0.041(0.067)	0.022(0.046)	0.008(0.022)
	$\widehat{T}_1^{\mathrm{CL}}$	0.169(0.135)	0.054(0.080)	0.022(0.046)	0.008(0.022)
	$\mathrm{tr}P^-$	0.323(0.253)	0.185(0.136)	0.141(0.125)	0.117(0.111)
	ω_{eff}	0.403(0.132)	0.253(0.125)	0.171(0.116)	0.114(0.099)
	Fisher-trace CV	0.583(0.216)	0.312(0.135)	0.221(0.152)	0.155(0.124)
	Random	0.536(0.042)	0.297(0.037)	0.231(0.038)	0.174(0.039)
L96 $d = 40$	MSE_z	0.255(0.237)	0.055(0.060)	0.024(0.043)	0.007(0.013)
	$\widehat{T}_1^{\mathrm{CL}}$	0.166(0.133)	0.036(0.064)	0.016(0.035)	0.002(0.009)
	$\mathrm{tr}P^-$	0.670(0.260)	0.338(0.149)	0.147(0.105)	0.076(0.076)
	ω_{eff}	0.415(0.433)	0.186(0.181)	0.107(0.102)	0.063(0.061)
	Fisher-trace CV	1.088(1.221)	0.274(0.322)	0.186(0.159)	0.119(0.122)
	Random	0.852(0.133)	0.309(0.029)	0.211(0.029)	0.140(0.031)
NS2D	MSE_z	7.195(4.957)	1.264(1.532)	0.969(0.865)	0.693(0.805)
	$\widehat{T}_1^{\mathrm{CL}}$	7.404(5.040)	1.264(1.532)	0.969(0.865)	0.735(0.841)
	$\mathrm{tr}P^-$	15.228(23.629)	3.118(1.426)	2.695(1.296)	2.237(1.244)
	ω_{eff}	17.246(23.203)	4.702(2.222)	2.983(1.285)	2.260(1.046)
	Fisher-trace CV	11.106(6.504)	5.124(2.521)	4.167(2.161)	3.001(1.487)
	Random	10.258(1.179)	3.827(0.678)	2.594(0.603)	1.666(0.555)
β-BVE	MSE_z	0.0104(0.0040)	0.0034(0.0028)	0.0015(0.0021)	0.0007(0.0014)
	$\widehat{T}_1^{\mathrm{CL}}$	0.0056(0.0048)	0.0024(0.0029)	0.0012(0.0018)	0.0002(0.0006)
	$\mathrm{tr}P^-$	0.0225(0.0066)	0.0121(0.0072)	0.0078(0.0026)	0.0006(0.0010)
	ω_{eff}	0.0225(0.0066)	0.0170(0.0063)	0.0077(0.0030)	0.0033(0.0036)
	Fisher-trace CV	0.0104(0.0040)	0.0063(0.0028)	0.0041(0.0017)	0.0032(0.0020)
	Random	0.0221(0.0024)	0.0077(0.0018)	0.0057(0.0017)	0.0042(0.0016)
KS	MSE_z	0.013	0.000	0.000	0.000
	$\widehat{T}_1^{\mathrm{CL}}$	0.013	0.000	0.000	0.000
	$\mathrm{tr}P^-$	0.017	0.013	0.000	0.000
	ω_{eff}	0.003	0.003	0.000	0.000
	Fisher-trace CV	0.012	0.007	0.007	0.004
	Random	0.189	0.024	0.012	0.008

E.7 Sensor selection on recorded data

We evaluate every candidate subset in three training runs on a common split for each dataset. The filter receives only the selected channels. We rank subsets by full-panel RMSE and evaluate recovery using RMSE on the withheld channels. Both metrics pool squared errors over recordings, retained times, and target channels in coordinates standardized using training-data statistics.

Data and partitions. Hydraulic uses the six pressure channels PS1–PS6 from the UCI Condition Monitoring of Hydraulic Systems dataset [22]. We retain the first ten stable cycles for each combination of cooler, valve, pump, and accumulator conditions. We randomly assign the 144 combinations

to training, validation, and test sets (100, 22, and 22 combinations, respectively). Averaging blocks of 20 samples reduces each 60-second recording from 100 Hz to 300 observations at 5 Hz. We discard the first 50 observations of each cycle when computing errors. The library is all $\binom{6}{3} = 20$ pressure triples.

CPAP uses PhysioNet version 1.0.1 [20], with channels $(P1, P2, Q1, Q2)$ denoting pressure and flow at two Venturi stations. The processed release contains 238 recordings from 27 subjects. We removed three duplicate recordings before splitting the data. We split the data by subject into $17/5/5$ training/validation/test subjects ($139/48/48$ recordings). For each recording, we average the first 2000 samples in blocks of ten to obtain 200 observations at 10 Hz. We evaluate after the first 20 steps. The library is all six channel pairs. The PEEP/rate condition is known for initialization and baseline construction.

Training and filtering. Both studies use latent dimension $k = 3$, two width-64 hidden layers in each encoder and decoder, a ReLU encoder, a SiLU decoder, and bias-free linear latent dynamics initialized to the identity. We optimize equally weighted reconstruction and one-step latent prediction losses with Adam (learning rate 10^{-3}, batch size 1024, 60 epochs, gradient clipping at norm one). Transitions stay within recordings. After training, we rescale M if its spectral radius exceeds 0.995. Channel standardization, process covariance, and initial latent moments use training data only. Hydraulic resets each cycle to the training phase-zero latent mean/covariance; CPAP uses training moments for the corresponding PEEP/rate condition. Neither initializes from the full validation or test state.

The EKF uses observations from the selected channels without added observation noise, Joseph covariance updates, and $R = rI_p$ as a filter regularizer. For each dataset we choose one r by minimizing validation full-panel RMSE averaged over candidates and training runs. Hydraulic uses $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and selects 0.1; CPAP adds 1 to this grid and selects 0.01. We use the validation ranking and the same selected r for every test candidate.

Scores and endpoints. Let $W = N^{-1} \sum_j J_{D_{\mathrm{norm}}}(z_j)^\top J_{D_{\mathrm{norm}}}(z_j)$, using 4096 training points sampled without replacement. The metric matches the standardized evaluation coordinates. We compute the model score as $T_1^{\mathrm{LG}} = \mathrm{tr}(WP^\star)$ from the exact posterior DARE, and $\widehat{T}_1^{\mathrm{CL}} = \mathrm{tr}(W\widehat{\Sigma}^{\mathrm{err}}_{H,\mathrm{val}})$ from uncentered validation latent-error second moments. We apply an eigenvalue floor to the training-residual process covariance before solving the DARE and symmetrize each iterate. Lower values of both scores predict lower test full-panel RMSE. Table 16 compares the full ranking with the performance of its first choice.

Table 16: Held-out full-panel ranking. Spearman correlations compare validation scores with test RMSE across every candidate. Relative top-1 regret is $100(\mathrm{RMSE}_{\mathrm{selected}}/\mathrm{RMSE}_{\mathrm{oracle}} - 1)$ within the candidate library.

Dataset	Run	$\rho(T_1^{\mathrm{LG}})$	$\rho(\widehat{T}_1^{\mathrm{CL}})$	CL top-1 regret (%)
Hydraulic	1	0.239	0.893	0.00
Hydraulic	2	0.367	0.520	3.52
Hydraulic	3	0.665	0.747	25.81
CPAP	1	0.429	0.943	0.00
CPAP	2	0.486	0.486	0.00
CPAP	3	0.943	0.829	0.00

On both datasets, the calibrated score had a higher median Spearman correlation with test full-panel RMSE across the three training runs than the posterior model score (Table 16).

For withheld-channel recovery, we select the CPAP pair by the calibrated validation score and the Hydraulic triple by validation full-panel RMSE. On held-out subjects (CPAP) or conditions (Hydraulic), we compare the selected filter with phase climatology and a decoder rollout without updates, both fitted on training data. The phase means use training cycles on Hydraulic and the corresponding PEEP/rate condition on CPAP. Table 17 reports $100(1 - \mathrm{RMSE}_{\mathrm{selected}}/\mathrm{RMSE}_{\mathrm{baseline}})$ using the lower withheld-channel test RMSE of the two baselines.

For the full-panel comparison, we bootstrap the paired difference between the selected subset and the subset ranked last by the same validation rule. We resample the 22 held-out conditions on Hydraulic using condition-level RMSE. On CPAP we average recording-level MSE within each subject and resample the five held-out subjects with equal weight. All six intervals favor the selected subset.

Table 17: Validation-selected subsets on held-out data. Improvement is measured on withheld channels relative to the stronger baseline. The paired 95% interval gives the difference in full-panel error (selected subset minus the subset ranked last on validation data), using condition-level RMSE on Hydraulic and subject-level MSE on CPAP.

Dataset	Run	Selected channels	Withheld-channel RMSE reduction (%)	Full-panel difference (selected minus last, 95% interval)
Hydraulic	1	PS3/PS4/PS5	55.05	$[-2.480, -1.641]$
Hydraulic	2	PS1/PS3/PS5	36.14	$[-1.003, -0.749]$
Hydraulic	3	PS1/PS3/PS5	36.61	$[-1.880, -1.474]$
CPAP	1	P1/Q2	54.52	$[-0.285, -0.107]$
CPAP	2	P2/Q2	51.34	$[-0.213, -0.082]$
CPAP	3	P2/Q2	50.70	$[-0.206, -0.026]$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 summarize the exact sensor example (Proposition 3.3), the conditional Pearson bound (Theorem 4.2), and the experimental findings. Section 3 gives the conditions for local separation of the Riccati solutions (Theorem 3.4).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix A.2 describes the scope of the theoretical and empirical results. Appendix A.3 distinguishes assumptions checked on the trained models from those that remain hypotheses. Section 6 discusses how sensor rankings can change on a new trajectory.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each theorem states its hypotheses or points to their full statement. App. A.3 distinguishes analytic results, direct computations on trained-model matrices, finite-sample evidence, and remaining assumptions. The proof of Theorem 3.2 follows its statement in §3. Appendices B–D give the remaining proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: App. E.1 gives simulation and sampling parameters, training settings, filter updates, and sensor libraries. App. E.6 gives the calibration/test procedure, and App. E.7 describes the real-data preparation and evaluation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is not released with this submission. App. E.1 describes the simulation, training, and filtering methods and the sensor libraries; App. E.7 identifies the public datasets and their preparation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Optimizer, batch size, training length, number of training runs, and architecture details are given in App. E.1 and App. E.3. Simulation spin-up and snapshot intervals are in App. E.1; simulation and real-data splits are in App. E.6 and App. E.7, respectively.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We report sample standard deviations across training runs for the architecture comparisons and across model–trajectory combinations for the calibration/test comparisons (App. E.3, E.6). Sensor-pair 95% intervals resample base sensors with their perturbations (App. D.2). Real-data intervals resample operating conditions for Hydraulic and subjects for CPAP (App. E.7). App. E.3 reports Holm and Benjamini–Yekutieli corrections and a cross-system mixed-effects fit.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We ran the simulations, autoencoder training, and Kalman filtering on CPUs. The individual experiments report run counts and training budgets. A complete hardware specification and total runtime across all experiments are unavailable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The experiments use simulated trajectories and two public datasets. CPAP is a secondary analysis of de-identified recordings, with consent and ethics approval reported by the original study [20]; no participants were recruited for this work.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is foundational theoretical and methodological work on sensor-selection diagnostics in latent data assimilation, with no direct path to societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No released data or models with risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper credits both dataset sources but omits their terms of use in App. E.7.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new dataset or model is released as a contributed artifact.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No new participant study or crowdsourcing was conducted; we reuse the public de-identified CPAP recordings described in App. E.7.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new participant study was conducted. The original CPAP study reports ethics approval and consent for its public de-identified release [20].

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology; any use was limited to writing and editing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.