OBSERVABLE GEOMETRY FOR DERIVATIVE-FREE OPTIMIZATION OF DYNAMICAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Derivative-free optimization of dynamical systems avoids differentiating through a simulator, yet it often reduces structured outputs to scalar losses and discards their observation-wise responses. These responses retain optimization-relevant structure that scalarization hides. We exploit this structure through *observable mapping* (OM), which constructs local response-informed coordinates, and instantiate it in *Observable-Mapping Ensemble Kalman Inversion* (OM-EKI) as a derivative-free optimizer for structured-response dynamical-system problems. We call the resulting task-relevant response structure *observable geometry* and hypothesize that useful coordinates exhibit three properties: compactness, residual reachability, and local persistence. Across settings with available, costly, and unavailable exact gradients, OM-EKI achieves competitive held-out accuracy and runtime on a deterministic Neural ODE, reduces median time-to-quality by $7.72\times$ relative to Adam on a 19,587-parameter stiff UDE, and approaches surrogate-gradient validation performance on a 50,176-parameter hard-reset recurrence for Spiking Heidelberg Digits (SHD). Mechanism-focused controls further support this hypothesis. Together, these results show that structured forward responses can be used not only to evaluate derivative-free search in dynamical systems, but also to organize it.

1 INTRODUCTION

Derivative-free optimization can act on a simulator without differentiating through its implementation. Yet each evaluation may return structured observations—trajectories, fields, or multichannel measurements—that common perturbation-based methods reduce to a scalar objective (Hansen & Ostermeier, 2001; Salimans et al., 2017; Hazelden et al., 2023). For example, a Neural ODE solve produces a trajectory (Chen et al., 2018; Rubanova et al., 2019). Scalarization preserves aggregate objective change but discards the observation-wise pattern that produces it.

As Figure 1 illustrates, a scalar loss collapses an observation-wise response into an aggregate change. Structured responses preserve the correspondence

$$q_j \longmapsto \Delta\mathcal{F}_j \in \mathbb{R}^M,$$

from a parameter direction to its observation-wise effect. Across perturbations, this mapping distinguishes directions by observable behavior, reveals whether their responses concentrate, and determines whether the retained responses can express the current residual correction. Structured forward information can therefore define an optimization representation rather than merely score candidates.

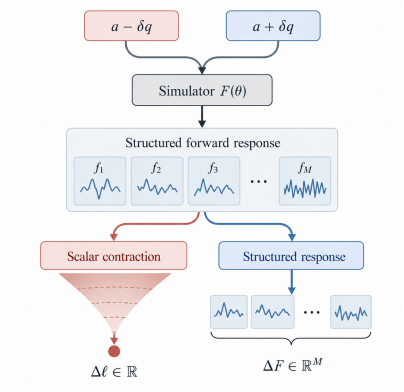

Figure 1: **Structured responses retain observable information.** A paired response $\Delta\mathcal{F} \in \mathbb{R}^M$ retains what $\Delta\ell \in \mathbb{R}$ discards.

Existing approaches capture complementary parts of this opportunity. Some retain structured residuals but operate within full or externally specified parameter spaces; others construct low-dimensional search spaces from random projections, scalar evaluations, or derivative information (Iglesias et al., 2013; Cartis & Roberts, 2019; Cartis et al., 2019;

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

Constantine et al., 2014; Cartis & Roberts, 2023; Gross &
Parks, 2022). What is missing is a derivative-free mechanism through which measured structured responses determine the search representation itself.

Paired perturbations estimate the local parameter-to-observation response within a probe span. This response operator induces a local geometry over parameter directions according to their measurable effects; we call it *observable geometry*. The resulting secant operator recovers projected gradient and Gauss–Newton geometry, and organizes high-dimensional parameters through a compact representation of their measurable effects.

Observable geometry is therefore a local optimization representation. Its utility depends on whether the measured responses admit compact coordinates, whether those coordinates can express the current residual correction, and how long they persist as optimization moves. We therefore formulate the *observable-geometry hypothesis*: useful forward-derived coordinates should be compact, residual-reachable, and locally persistent.

We introduce *observable mapping* (OM) to turn this measured geometry into response-aligned coordinates whose capacity follows the current residual. OM-EKI realizes this principle as a derivative-free optimizer for structured-response problems: it performs ensemble Kalman residual correction in the observable coordinates, verifies finite proposals with fresh evaluations, and renews the map as the geometry changes. By deriving its search representation from forward responses, OM-EKI applies whether backward derivatives are faithful, costly, or unavailable without surrogate approximation (Gholami et al., 2019; Zhuang et al., 2020; Neftci et al., 2019).

Contributions.

1. We show that structured forward responses expose a local parameter-to-observation geometry that can define the search representation.
2. We formulate the observable-geometry hypothesis: useful observable coordinates require compact response structures, residual reachability, and low geometry drift, making them economical, corrective, and persistent.
3. We introduce *observable mapping* (OM), which makes the effective search dimension follow observable structure rather than ambient parameter count, and develop OM-EKI as a general derivative-free optimizer for structured-response dynamical-system problems.
4. We evaluate OM-EKI across three gradient-access regimes, where exact gradients are readily available, costly, or unavailable, and examine the roles predicted by the observable-geometry hypothesis.

2 Related Work

Structured-residual optimization uses vector outputs as optimization signals. EKI forms updates from empirical parameter–observation covariation and has been applied to Neural ODEs, physics-based ML parameterizations, and physics-informed neural networks (Iglesias et al., 2013; Kovachki & Stuart, 2019; Böttcher, 2023; Lopez-Gomez et al., 2022; Pensoneault & Zhu, 2024). Derivative-free nonlinear least-squares methods interpolate residual vectors to construct local Gauss–Newton models (Cartis & Roberts, 2019; Cartis et al., 2019). Together, these works establish structured responses as an optimization interface, either in the full parameter space or within finite ensemble/model spans (Chada et al., 2018; Harris & Schillings, 2025).

Low-dimensional optimization approaches scalability through the search representation. Random-subspace DFO and randomized Gauss–Newton methods select a span and construct models or steps within it (Cartis & Roberts, 2023; Cartis et al., 2022); OMoRF and ASEBO learn search or sensing directions from scalar evaluations or estimated-gradient history (Gross & Parks, 2022; Choromanski et al., 2019). Jacobian metrics, active subspaces, vector-valued dimension reduction, and likelihood-informed methods identify influential or data-constrained directions from derivatives, samples, or priors (Transtrum et al., 2011; Constantine et al., 2014; Zahm et al., 2020; Cui et al., 2014). Together, these lines of work preserve structured residuals or learn low-dimensional search spaces, but do not derive the search representation from measured structured responses. What remains missing is a derivative-free mechanism that turns those responses into search coordinates.

Ours. OM closes this gap by deriving the search representation directly from structured forward responses, rather than from random or prescribed reductions. OM-EKI operationalizes this response-derived geometry as a derivative-free optimizer for structured-response dynamical-system problems across gradient-access regimes.

3 Observable Geometry of Derivative-free Optimization

The local parameter-to-observation operator identifies parameter changes visible to observations and available for optimization. Structured forward responses measure this operator, exposing an observation-dependent search space that can be far smaller than the ambient parameterization.

3.1 Structured forward responses expose local optimization geometry

Let $\theta \in \mathbb{R}^P$ parameterize a rollout $\mathcal{R}_\theta$, and let $\mathcal{O}$ collect the task-relevant quantities. We define

$$\begin{aligned} &\mathcal{F}(\theta) = \mathcal{O}(\mathcal{R}_\theta) \in \mathbb{R}^M, \quad e(\theta) = \Gamma^{-1/2}[y - \mathcal{F}(\theta)], \\ &\ell_2(\theta) = \tfrac{1}{2}\|e(\theta)\|_2^2, \end{aligned} \tag{1}$$

where $\Gamma \succ 0$ fixes the units of the observation channels. At anchor a_t, the normalized response operator $\bar{J}_t = \Gamma^{-1/2} D\mathcal{F}(a_t)$ gives $\nabla \ell_2(a_t) = -\bar{J}_t^\top e(a_t)$ and Gauss–Newton geometry $\bar{J}_t^\top \bar{J}_t$.

A scalar loss collapses the response into an aggregate change. Centered structured secants retain the observation-wise response geometry within the probed span. For orthonormal probes $Q = [q_1, \dots, q_K] \in \mathbb{R}^{P\times K}$ and $P_Q = QQ^\top$, define

$$h_j = \Gamma^{-1/2} \frac{\mathcal{F}(a_t + \delta q_j) - \mathcal{F}(a_t - \delta q_j)}{2\delta}, \qquad H_\delta = [h_1, \dots, h_K]. \tag{2}$$

Centered finite differences give

$$\begin{aligned} H_\delta &= \bar{J}_t Q + O(\delta^2), \\ -QH_\delta^\top e(a_t) &= P_Q \nabla \ell_2(a_t) + O(\delta^2), \\ H_\delta^\top H_\delta &= Q^\top (\bar{J}_t^\top \bar{J}_t) Q + O(\delta^2). \end{aligned} \tag{3}$$

Hence the measured response operator recovers the projected gradient and Gauss–Newton geometry within the probe span. Local smoothness gives the $O(\delta^2)$ accuracy; at finite scale, H_δ remains directly measurable from forward simulator evaluations. Appendix C.2 gives the precise conditions and discusses the nonsmooth case.

3.2 Observable directions define the effective local parameter space

The normalized response operator induces the local observable seminorm

$$\|\Delta\theta\|_{\mathrm{obs},t}^2 = \Delta\theta^\top \bar{J}_t^\top \bar{J}_t \Delta\theta = \|\bar{J}_t \Delta\theta\|_2^2. \tag{4}$$

It measures parameter changes by their observational effects. Perturbations that differ by an element of $\ker \bar{J}_t$ are locally indistinguishable, so $\mathbb{R}^P / \ker \bar{J}_t$ forms the effective parameter space, removing variation invisible to the observations. Let $\bar{J}_t = U_{t,q} \Sigma_{t,q} V_{t,q}^\top$ be its SVD, with $q = \mathrm{rank}(\bar{J}_t)$. Then

$$\bar{J}_t v = U_{t,q} \Sigma_{t,q} z, \qquad z = V_{t,q}^\top v. \tag{5}$$

so the observable effects of v are represented by the q coordinates in z. When the singular spectrum is concentrated, the leading r directions provide a compact representation independent of dimension P. Low response rank alone, however, does not imply a useful optimization representation: the retained space must also express the current residual correction and remain relevant as the anchor moves. Appendix C gives the exact factorization, optimal truncation, and nonlinear remainder.

3.3 The observable-geometry hypothesis

We therefore characterize a useful local representation by three complementary properties: whether the measured response is compressible, whether the retained space can express the current correction, and whether that space persists along the optimization path.

Let $\sigma_i(t)$ denote the nonzero singular values of $\bar{J}_t$, $B_{t,r}$ its leading r right-singular basis, $e_t = e(a_t) \neq 0$, r_t the selected dimension, and $m_t = \min(r_t, r_{t+1})$. At the operator level, define

$$C_r(t) = \frac{\sum_{i=1}^{r} \sigma_i(t)^2}{\sum_i \sigma_i(t)^2}, \qquad \text{observable compactness,} \tag{6}$$

$$R_r(t) = \frac{\|\Pi_{\mathcal{R}(\bar{J}_t B_{t,r})} e_t\|_2^2}{\|e_t\|_2^2}, \qquad \text{residual reachability,} \tag{7}$$

$$D_t = \left(1 - \frac{\|B_{t,r_t}^\top B_{t+1,r_{t+1}}\|_F^2}{m_t}\right)^{1/2}, \qquad \text{geometry drift.} \tag{8}$$

Here $\Pi_{\mathcal{S}}$ projects onto $\mathcal{S}$. C_r is the fraction of squared response energy retained, R_r the reachable fraction of residual energy, and D_t the loss of overlap between consecutive coordinate spaces.

Let $g_t = \bar{J}_t^\top e_t$ and $P_{t,r} = B_{t,r} B_{t,r}^\top$. For equal consecutive ranks, the diagnostics directly control the retained optimization geometry:

$$\|(I - P_{t,r}) g_t\|_2 \leq \sqrt{1 - C_r(t)}\, \|\bar{J}_t\|_F \|e_t\|_2, \tag{9}$$

$$\min_z \|e_t - \bar{J}_t B_{t,r} z\|_2^2 = [1 - R_r(t)] \|e_t\|_2^2, \tag{10}$$

$$\|(P_{t+1,r} - P_{t,r}) v\|_2 \leq \sqrt{r}\, D_t \|v\|_2. \tag{11}$$

These relations connect compactness to retained gradient geometry, reachability to attainable residual correction, and drift to coordinate reuse. Appendix C.2 gives proofs and the unequal-rank case.

These diagnostics characterize the local representation; a finite update also requires its response model to remain predictive. For $\theta = a_t + A_t z$, let $\widehat{J}_{z,t}$ denote the reduced response operator. If $\|\widehat{J}_{z,t} - \bar{J}_t A_t\|_2 \leq \varepsilon_{\text{sec}}$ and the normalized response Jacobian is locally L_t-Lipschitz, then

$$\|e(a_t + A_t z)\|_2 \leq \|e_t - \widehat{J}_{z,t} z\|_2 + \varepsilon_{\text{sec}} \|z\|_2 + \frac{L_t}{2} \|A_t z\|_2^2. \tag{12}$$

The terms capture predicted residual, secant error, and nonlinear curvature. Favorable geometry therefore does not ensure finite-step progress unless the local response model remains accurate. Fresh evaluations test this validity and trigger coordinate reuse or reconstruction. Figure 2 connects this principle to the OM-EKI update cycle.

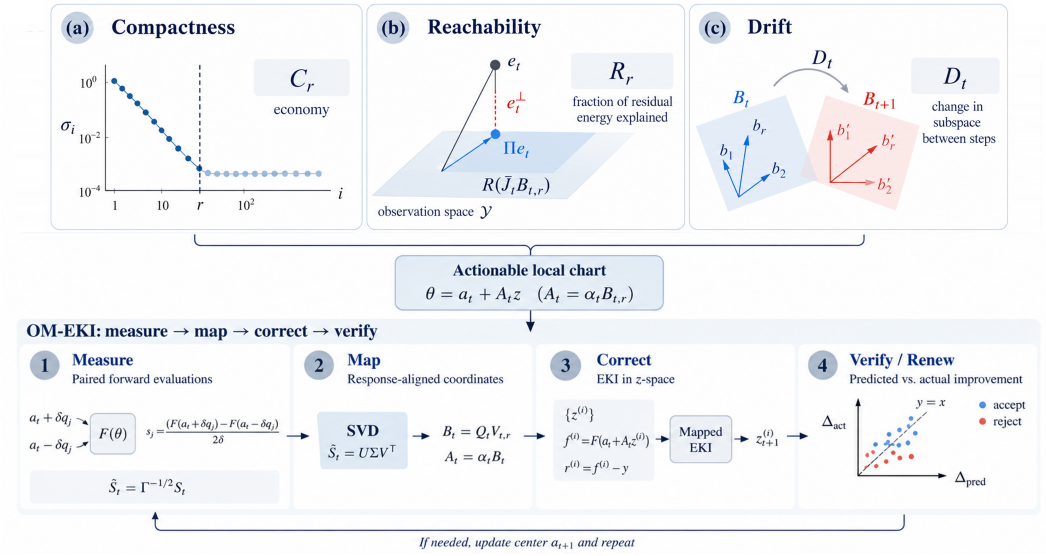

Figure 2: **From observable geometry to OM-EKI.** Compactness, reachability, and drift characterize coordinate quality and persistence; OM-EKI operationalizes this geometry through measure, map, correct, and verify/renew.

4 OM-EKI: Optimization Through a Measured Response Operator

OM-EKI makes observable geometry actionable: it constructs response-aligned coordinates, selects sufficient capacity for the current residual, verifies finite proposals, and renews the map as the geometry changes (Figure 2).

4.1 Measure and construct the observable map

OM does not optimize directly in the probe span. Instead, it uses Q_t to measure a probe-restricted realization of the operator-level observable geometry, then rotates and truncates this measured span into response-aligned optimization coordinates. At each anchor, orthonormal probes $Q_t = [q_1, \dots, q_K]$ generate centered secants $s_j = \Delta\mathcal{F}_j/(2\delta)$, the finite-scale responses highlighted in the Introduction (Figure 2, steps 1–2):

$$\begin{aligned} s_j &= \frac{\mathcal{F}(a_t + \delta q_j) - \mathcal{F}(a_t - \delta q_j)}{2\delta}, \\ S_t &= [s_1, \dots, s_K], \qquad \widetilde{S}_t = \Gamma^{-1/2} S_t. \end{aligned} \tag{13}$$

The normalized secant matrix $\widetilde{S}_t = \bar{J}_t Q_t + O(\delta^2)$ measures the probe-restricted response operator. Its factorization $\widetilde{S}_t = U_t \Sigma_t V_t^\top$ identifies response-aligned combinations of the sampled directions. Let q_t denote its numerical rank and $\hat{s}_{t,j}$ its singular values. OM chooses the retained capacity from two requirements: preserving dominant response variation and avoiding truncation before the reachable component of the current residual is represented. With $e_t = \Gamma^{-1/2}[y - \mathcal{F}(a_t)]$, define

$$\begin{aligned} r_E &= \min\left\{ k : \frac{\sum_{j=1}^{k} \hat{s}_{t,j}^2}{\sum_{j=1}^{q_t} \hat{s}_{t,j}^2} \geq \tau_E \right\}, \\ r_y &= \min\left\{ k : \frac{\sum_{j=1}^{k} (u_{t,j}^\top e_t)^2}{\sum_{j=1}^{q_t} (u_{t,j}^\top e_t)^2} \geq \tau_y \right\}, \\ r_t &= \min\{r_{\max}, q_t, \max(r_{\min}, r_E, r_y)\}. \end{aligned} \tag{14}$$

This operationalizes the two local requirements introduced in Section 3: response compactness controls how economically the measured operator can be represented, while residual-aware capacity prevents that representation from discarding the current reachable correction. Full-residual reachability is recorded separately by the empirical counterpart of R_r.

The resulting observable map is

$$B_t = Q_t V_{t,r_t}, \qquad A_t = \alpha_t B_t, \qquad \mathcal{M}_t(z) = a_t + A_t z. \tag{15}$$

Here B_t contains response-aligned probe combinations, while α_t stabilizes the aggregate response scale across reconstructions, preventing coordinate units from drifting with local sensitivity. The same factorization yields the reduced response model

$$\widehat{J}_{z,t} = \alpha_t \widetilde{S}_t V_{t,r_t}. \tag{16}$$

Appendices C.2 and B.1 give the scaling rule, probe reuse, and full construction.

4.2 Correct the vector residual with mapped EKI

Once OM defines the coordinates, mapped EKI converts observation discrepancies into residual corrections without scalarization. For $\theta^{(i)} = a_t + A_t z^{(i)}$ and $f^{(i)} = \mathcal{F}(\theta^{(i)})$,

$$z_{\text{prop}}^{(i)} = z^{(i)} - C^{zf}\big(C^{ff} + \lambda\Omega_t\big)^{-1}\big(f^{(i)} - y\big), \tag{17}$$

where C^{zf} and C^{ff} are the empirical coordinate–response and response covariances. Here, Ω_t sets the observation metric for the EKI correction, while λ controls regularization. This differs from Γ above, which normalizes responses when constructing the observable geometry

Under an affine mapped response with isotropic coordinate covariance $\sigma^2 I$, defining $J_{\Omega,z,t} = \Omega_t^{-1/2} D\mathcal{F}(a_t) A_t$ and $e_{\Omega,t} = \Omega_t^{-1/2}[y - \mathcal{F}(a_t)]$ gives

$$\Delta z = (J_{\Omega,z,t}^\top J_{\Omega,z,t} + \lambda\sigma^{-2} I)^{-1} J_{\Omega,z,t}^\top e_{\Omega,t}. \tag{18}$$

OM selects the optimization directions; under the affine mapped model, EKI reduces to a regularized Gauss–Newton residual correction in these coordinates. Appendix C.4 gives the derivation.

4.3 Verify finite proposals and renew the map

OM-EKI carries local response geometry to finite updates by verifying every proposal with fresh rollouts (Figure 2, step 4). Let $\widehat{\bar{\ell}}_{\text{prop}}$ denote the ensemble loss predicted by the reduced response model and $\bar{\ell}_{\text{prop}}$ the loss measured by fresh rollouts. Relative to the current verified loss $\bar{\ell}_t$, define

$$\Delta_{\text{pred}} = \bar{\ell}_t - \widehat{\bar{\ell}}_{\text{prop}}, \qquad \Delta_{\text{act}} = \bar{\ell}_t - \bar{\ell}_{\text{prop}}, \qquad \rho_t = \frac{\Delta_{\text{act}}}{\Delta_{\text{pred}}} \quad \text{when } \Delta_{\text{pred}} > 0. \tag{19}$$

The ratio ρ_t compares predicted and realized progress (Conn et al., 2000). OM-EKI accepts a proposal only when the predicted improvement is realized with sufficient agreement; otherwise regularization is increased. Fresh rollouts therefore prevent locally predicted improvements from being accepted when they fail under the simulator.

A map is reused locally and renewed when its lifetime expires or repeated rejection indicates loss of predictive validity. Verification safeguards finite nonlinear updates, while renewal keeps the observable coordinates relevant as the response geometry evolves. Appendix B gives the full acceptance and reconstruction protocol.

5 Experiments

We evaluate OM-EKI at two levels. Three tasks test its utility across parameter scales and faithful, costly, and model-mismatched backward interfaces; controlled mechanism experiments test why and when observable coordinates work. Primary results are four-seed medians under shared protocols. Appendices E–G provide protocols and individual runs, while Appendix D details the controls.

5.1 Practical utility across backward interfaces

We study three structured-response dynamical systems; Appendix E.4 gives their full definitions. The deterministic Neural ODE is a two-state trajectory-regression problem (Chen et al., 2018). POLLU is a stiff 20-state chemical-kinetics benchmark with a neural closure (SciMLBenchmarks.jl Contributors, n.d.). Spiking Heidelberg Digits (SHD) is an event-based spoken-digit benchmark (Cramer et al., 2022) modeled by a hard-reset spiking recurrence. Figure 3 gives the cross-task view.

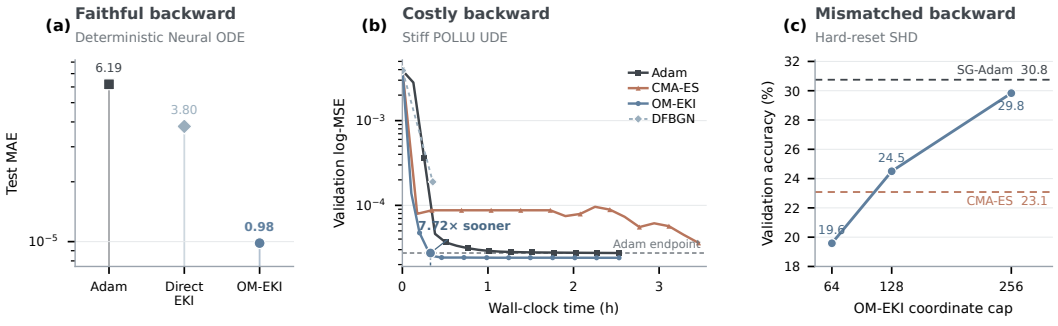

Figure 3: **Practical utility across gradient-access regimes.** The three panels isolate complementary roles of the method: (a) observable mapping beyond the EKI update, (b) response-defined coordinates beyond scalar or residual-subspace optimization, and (c) finite-scale forward responses under a surrogate-dependent backward interface.

5.1.1 When Exact Gradients Are Available: Deterministic Neural ODE

The deterministic Neural ODE provides a controlled setting with standard reverse-mode gradient access. Its $P = 252$ parameters make direct full-space EKI feasible, separating the benefit of EKI from the additional value of observable mapping. OM-EKI uses 64 coordinates; Table 1 reports validation-selected performance on disjoint held-out trajectory windows.

Figure 3(a) shows the test-MAE progression from Adam through direct EKI to OM-EKI. Table 1 adds that OM-EKI achieves $10.25\times$ lower validation MAE while running $2.48\times$ faster than Adam. Direct EKI already outperforms the gradient baselines on this relatively simple problem; OM-EKI improves further in held-out error and runtime. Thus EKI supplies an effective forward update, while 64 measured coordinates improve its search representation even when exact gradients are available.

Table 1: Validation-selected performance on the deterministic Neural ODE.

Method	Val. MAE ($\times 10^{-5}$) ↓	Test MAE ($\times 10^{-5}$) ↓	Time (s) ↓
Adam	7.4811	6.1868	383.58
AdamW	6.5218	6.6356	383.35
Adam + adjoint	3.5998	6.7874	611.62
Direct EKI (full)	<u>2.9513</u>	<u>3.7986</u>	<u>203.00</u>
OM-EKI	**0.72985**	**0.98288**	**154.71**

5.1.2 When Exact Gradients Are Costly: Stiff POLLU UDE

The stiff POLLU UDE tests time-to-quality when exact backward solves are costly. Its neural closure has $P = 19{,}587$ parameters; OM-EKI uses 128 coordinates (0.65%).

Table 2: High-dimensional POLLU UDE. Neural-ODE ES follows Hazelden et al. (2023); CMA-ES follows Hansen & Ostermeier (2001); DFBGN follows Cartis & Roberts (2023). Optimizer configurations are detailed in Appendix E.6; dashes denote unreached time-to-quality values.

Method	Optimization information	Validation log-MSE ↓	Endpoint time (s) ↓	Time to paired Adam quality (s) ↓
Adam	reverse-mode gradients	2.7382×10^{-5}	9,119.29	<u>9,119.29</u>
AdamW	reverse-mode gradients	<u>2.7380×10^{-5}</u>	9,019.96	—
Neural-ODE ES	scalar forward loss	2.1533×10^{-3}	9,266.65	—
CMA-ES	scalar forward loss	3.6101×10^{-5}	12,456.37	—
DFBGN	vector trajectory residuals	1.8926×10^{-4}	1,280.94	—
OM-EKI	vector trajectory observations	$\mathbf{2.4054\times10^{-5}}$	9,115.15	**1,182.56**

Figure 3(b) shows OM-EKI reducing median time-to-Adam-quality by $7.72\times$, while DFBGN and CMA-ES terminate at substantially higher errors. Since DFBGN also uses vector trajectory residuals in a 128-dimensional block, Table 2 suggests that residual access and dimensionality reduction alone do not explain OM-EKI's advantage. OM-EKI also exploits batched rollout parallelism: 100,385 forward evaluations require 9,057 s, versus 221 evaluations and 1,160 s for DFBGN, showing that its much larger evaluation budget does not translate proportionally into wall-clock cost (Appendix E.6).

5.1.3 When Exact Gradients Are Unavailable: Hard-Reset SHD

The hard-reset SHD task tests whether forward-defined search coordinates remain useful when exact gradients are unavailable. OM-EKI leaves the 50,176-parameter recurrence unchanged and constructs its coordinates from finite-scale forward responses on SHD.

Figure 3(c) shows a clear capacity trend. At cap 128, OM-EKI surpasses CMA-ES in both accuracy and Brier error; at cap 256, it comes within 0.92 percentage points of SG-Adam while running $1.27\times$ faster. Table 3 also shows the relationship between search dimension and runtime of OM-EKI. These results suggest that the unchanged hard-reset system exposes actionable search structure through its forward responses, even when the backward interface is surrogate-defined.

Beyond these primary tasks, Appendix D.7 provides an external transfer test in closed-loop PDE policy optimization: in one prespecified PDEControlGym run, OM-EKI reaches Adam's endpoint validation cost $24.1\times$ sooner.

Table 3: Hard-reset SHD validation metrics. SG-Adam replaces the backward derivative with a fast-sigmoid surrogate (Neftci et al., 2019).

Method	Optimizer dim.	Brier ↓	Accuracy ↑	Time (s) ↓
Surrogate backward				
SG-Adam	50,176	**0.041492**	**30.75%**	1,776.01
SG-AdamW	50,176	0.042815	25.58%	3,452.03
Scalar forward loss				
SPSA-Adam	50,176	0.047710	5.25%	2,151.01
Neural-ODE ES	50,176	0.046933	9.00%	1,716.10
CMA-ES	50,176	0.044018	23.08%	1,158.71
Vector forward observations				
OM-EKI (dim. 64)	64	0.045523	19.58%	**456.82**
OM-EKI (cap 128)	124–128	0.043569	24.50%	778.07
OM-EKI (cap 256)	251–256	0.042218	29.83%	1,401.56

5.2 Mechanism tests: observable geometry

We next examine what makes a response-derived search representation useful in practice, from its construction and corrective capacity to its persistence under finite optimization steps.

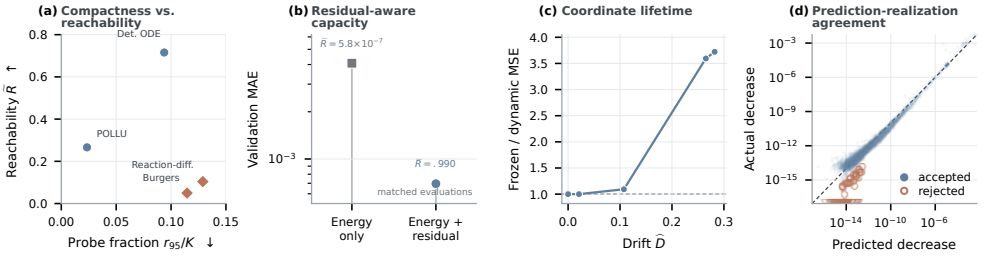

Figure 4: **Mechanism tests for observable geometry.** (a) Compactness r_{95}/K versus residual reachability $\widetilde{R}$. (b) Energy-only versus residual-aware capacity at matched evaluations. (c) Frozen/dynamic terminal MSE ratio versus drift; the dashed line marks parity. (d) Predicted versus realized decrease for accepted and rejected proposals; the dashed line marks agreement.

5.2.1 Response alignment matters beyond dimension reduction

Forced Duffing compares 32 measured and random coordinates under affine ($\kappa = 0$) and nonlinear ($\kappa = 4$) responses. OM-EKI lowers MSE $9.08\times$ and $1.45\times$, respectively (Table 4; Appendix D.2). At $\kappa = 4$, the random span remains full-rank over all 16 forcing dimensions, explaining its smaller gap to OM-EKI.

Table 4: Budget-matched forced-Duffing control. OM-EKI and EKI (rand) use 32 measured and random coordinates; EKI (full) uses all 128 parameters.

Method	Dim.	Val. MSE ↓
Affine response ($\kappa = 0$)		
OM-EKI	32	$\mathbf{6.073\times10^{-13}}$
EKI (rand)	32	5.517×10^{-12}
EKI (full)	128	6.131×10^{-9}
Nonlinear response ($\kappa = 4$)		
OM-EKI	32	$\mathbf{3.337\times10^{-11}}$
EKI (rand)	32	4.852×10^{-11}
EKI (full)	128	1.917×10^{-8}

OM-EKI leads even when random coordinates have sufficient forcing capacity. Matched dimension isolates alignment beyond random reduction; worse full-space EKI shows that ambient capacity does not recover the measured geometry.

5.2.2 Compactness is insufficient without residual reachability

The deterministic and POLLU runs combine compact responses with higher residual reachability (Figure 4(a)). Reaction–diffusion and Burgers remain locally predictive but expose only 10.4% and 5.0% of the full residual. Compactness alone therefore does not ensure actionable coordinates.

The deterministic Neural ODE control in Figure 4(b) and Table 5 compares an energy-only rank rule with a residual-aware rule that expands rank until the current correction is represented. The latter reduces validation MAE by $5.86\times$ with only a 0.012% difference in forward evaluations. This separates their roles: compactness makes coordinates economical, while reachability makes the current correction actionable without covering the ambient space. Additional Burgers results are provided in Appendix D.6, further illustrating this boundary case.

Table 5: Budget-matched residual-capacity control.

Capacity rule	Residual capture ↑	Val. MAE ↓
Energy only	5.76×10^{-7}	4.079×10^{-3}
Energy + residual	**0.990**	$\mathbf{6.961\times10^{-4}}$

5.2.3 Drift limits coordinate lifetime

Figure 4(c) isolates coordinate lifetime under approximately matched compactness, new-map reachability, and forward budget. At matched training-residual endpoints, dynamic and frozen maps are effectively tied at estimated drift $D = 0$–0.021; at $D = 0.265$–0.282, the frozen-map penalty reaches 3.59–$3.72\times$. Drift therefore determines coordinate lifetime: renewal adds no measurable value while the space remains stable but preserves an effective search space as the geometry changes.

5.2.4 Nonlinear verification safeguards finite-step updates

Aligned, reachable, and stable coordinates still define only a local response model. Figure 4(d) tests whether predicted improvements survive finite simulator steps. Static mapping provides the operational ablation, while the controlled drift experiment above identifies when renewal matters.

Table 6: Deterministic Neural ODE component ablation.

Variant	Dim.	Val. MAE ($\times\mathbf{10^{-5}}$) ↓	Time (s) ↓
OM-EKI	64	**1.3224**	48.84
Static mapping	46	22.107	36.91
Unconditional acceptance	61	49.286	44.64

Unconditional acceptance raises validation error $37.27\times$ for only 8.6% less runtime (Table 6). Fresh rollouts retain predicted improvements only when they materialize under the simulator, supplying finite-step reliability.

Together, these controls show that observable geometry is more than a low-dimensional span. Response alignment provides an informative orientation, compactness makes the coordinates economical, reachability makes the current residual correctable, and drift determines their useful lifetime. Nonlinear verification converts this local geometry into reliable finite updates.

6 Conclusion

This work reframes derivative-free optimization of dynamical systems around search representations constructed from structured simulator responses. OM turns measured response structure into residual-conditioned coordinates, while OM-EKI operationalizes this geometry as a general derivative-free optimizer for structured-response dynamical-system problems. Across tasks and mechanism-focused controls, the gains are not explained by avoiding scalarization, retaining vector residuals, or reducing dimension alone: response alignment, corrective reachability, coordinate persistence, and finite-step verification jointly determine when the representation is useful.

These mechanisms also define the method's scope. Observable geometry is local and probe-restricted: finite measurements can miss relevant directions, limited reachability constrains correction, and rapid drift makes coordinate reuse ineffective. Future work should extend observable mapping to noisy or stochastic settings, larger scales, and broader scientific and control systems, while developing adaptive probing and renewal, update rules beyond EKI, and deeper links to identifiability and parameter symmetries. More broadly, forward evaluations can help determine not only how candidates are scored, but how the search itself is organized.

REPRODUCIBILITY STATEMENT

Appendix A links each main-text claim to its formal basis and empirical record. Appendices B and C specify the four-stage OM-EKI procedure and its local, probe-restricted theory. Appendix D documents the matched coordinate-choice, residual-capacity, drift, and nonlinear-verification controls, including the variables held fixed and their realized evaluation budgets. Appendices E–G report task equations, observation construction, data partitions, solvers, budgets, checkpoint rules, hardware, seed-wise primary results, and the corresponding empirical traces.

AI USE STATEMENT

Generative AI tools assisted with language refinement, figure design and production, exploratory brainstorming during the development of the central idea, and the exploration of intermediate mathematical derivations and proof arguments. The authors independently examined, refined, and verified all mathematical derivations and proofs, and determined the final statements and conclusions. The authors also developed and validated the methodology, designed and executed the experiments, checked the references and numerical evidence, and take full responsibility for the final manuscript.

REFERENCES

Luke Bhan, Yuexin Bian, Miroslav Krstic, and Yuanyuan Shi. PDE control gym: A benchmark for data-driven boundary control of partial differential equations. In *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 1083–1095, 2024.

Lucas Böttcher. Gradient-free training of neural ODEs for system identification and control using ensemble kalman inversion. *arXiv preprint arXiv:2307.07882*, 2023. URL https://arxiv.org/abs/2307.07882.

Coralia Cartis and Lindon Roberts. A derivative-free Gauss–Newton method. *Mathematical Programming Computation*, 11(4):631–674, 2019. doi: 10.1007/s12532-019-00161-7.

Coralia Cartis and Lindon Roberts. Scalable subspace methods for derivative-free nonlinear least-squares optimization. *Mathematical Programming*, 199:461–524, 2023. doi: 10.1007/s10107-022-01836-1.

Coralia Cartis, Jan Fiala, Benjamin Marteau, and Lindon Roberts. Improving the flexibility and robustness of model-based derivative-free optimization solvers. *ACM Transactions on Mathematical Software*, 45(3):32:1–32:41, 2019. doi: 10.1145/3338517.

Coralia Cartis, Jaroslav Fowkes, and Zhen Shao. A randomised subspace Gauss–Newton method for nonlinear least-squares. *arXiv preprint arXiv:2211.05727*, 2022. doi: 10.48550/arXiv.2211.05727.

Neil K. Chada, Marco A. Iglesias, Lassi Roininen, and Andrew M. Stuart. Parameterizations for ensemble kalman inversion. *Inverse Problems*, 34(5):055009, 2018. doi: 10.1088/1361-6420/aab6d9.

Neil K. Chada, Andrew M. Stuart, and Xin T. Tong. Tikhonov regularization within ensemble kalman inversion. *SIAM Journal on Numerical Analysis*, 58(2):1263–1294, 2020. doi: 10.1137/19M1242331.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Krzysztof M. Choromanski, Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, and Vikas Sindhwani. From complexity to simplicity: Adaptive ES-active subspaces for blackbox optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/88bade49e98db8790df275fcebb37a13-Abstract.html.

Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust Region Methods*. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000. doi: 10.1137/1.9780898719857.

Paul G. Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014. doi: 10.1137/130916138.

Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke. The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2744–2757, 2022. doi: 10.1109/TNNLS.2020.3044364.

Tiangang Cui, James Martin, Youssef M. Marzouk, Antti Solonen, and Alessio Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11):114015, 2014. doi: 10.1088/0266-5611/30/11/114015.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Andrzej Dulny, Andreas Hotho, and Anna Krause. NeuralPDE: Modelling dynamical systems from data. In *KI 2022: Advances in Artificial Intelligence*, volume 13404 of *Lecture Notes in Computer Science*, pp. 75–89. Springer, 2022. doi: 10.1007/978-3-031-15791-2_8.

Andrzej Dulny, Andreas Hotho, and Anna Krause. Dynabench: A benchmark dataset for learning dynamical systems from low-resolution data. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pp. 438–455. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-43412-9_26.

Amir Gholami, Kurt Keutzer, and George Biros. ANODE: Unconditionally accurate memory-efficient gradients for neural ODEs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 730–736, 2019. doi: 10.24963/ijcai.2019/103.

James C. Gross and Geoffrey T. Parks. Optimization by moving ridge functions: Derivative-free optimization for computationally intensive functions. *Engineering Optimization*, 54(4):553–575, 2022. doi: 10.1080/0305215X.2021.1886286.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001. doi: 10.1162/106365601750190398.

Ruben Harris and Claudia Schillings. Accuracy improvement in ensemble kalman inversion through data-informed ensemble selection. *Inverse Problems*, 41(10):105010, 2025. doi: 10.1088/1361-6420/ae115f.

James Hazelden, Yuhan Helena Liu, Eli Shlizerman, and Eric Shea-Brown. Evolutionary algorithms as an alternative to backpropagation for supervised training of biophysical neural networks and neural ODEs. *arXiv preprint arXiv:2311.10869*, 2023. URL `https://arxiv.org/abs/2311.10869`.

Marco A. Iglesias, Kody J. H. Law, and Andrew M. Stuart. Ensemble kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, 2013. doi: 10.1088/0266-5611/29/4/045001.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL `https://arxiv.org/abs/1412.6980`.

Nikola B. Kovachki and Andrew M. Stuart. Ensemble kalman inversion: A derivative-free technique for machine learning tasks. *Inverse Problems*, 35(9):095005, 2019. doi: 10.1088/1361-6420/ab1c3a.

Ignacio Lopez-Gomez, Costa Christopoulos, H. L. Langeland Ervik, Oliver R. A. Dunbar, Yair Cohen, and Tapio Schneider. Training physics-based machine-learning parameterizations with gradient-free ensemble kalman methods. *Journal of Advances in Modeling Earth Systems*, 14(8):e2022MS003105, 2022. doi: 10.1029/2022MS003105.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019. doi: 10.1109/MSP.2019.2931595.

Andrew Pensoneault and Xueyu Zhu. Efficient bayesian physics informed neural networks for inverse problems via ensemble kalman inversion. *Journal of Computational Physics*, 508:113006, 2024. doi: 10.1016/j.jcp.2024.113006.

Raymond Ros and Nikolaus Hansen. A simple modification in CMA-ES achieving linear time and space complexity. In *Parallel Problem Solving from Nature*, pp. 296–305, 2008. doi: 10.1007/978-3-540-87700-4_30.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017. URL https://arxiv.org/abs/1703.03864.

SciMLBenchmarks.jl Contributors. POLLU work-precision diagrams, n.d. URL https://docs.sciml.ai/SciMLBenchmarksOutput/stable/StiffODE/Pollution/. Accessed September 10, 2026.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992. doi: 10.1109/9.119632.

Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*, 83(3):036701, 2011. doi: 10.1103/PhysRevE.83.036701.

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. doi: 10.1137/18M1221837.

Juntang Zhuang, Nicha Dvornek, Xiaoxiao Li, Sekhar Tatikonda, Xenophon Papademetris, and James Duncan. Adaptive checkpoint adjoint method for gradient estimation in neural ODE. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11639–11649, 2020.

A CLAIM–THEORY–EVIDENCE INDEX

The appendix follows the paper's claim chain. Appendix B specifies the complete procedure, Appendix C develops its mathematical basis, and Appendix D provides diagnostic controls, transfer evidence, and an external PDE-control benchmark. Appendices E–G document protocols, individual runs, and empirical traces.

Table 7 maps each central claim to its formal status and empirical consequence, distinguishing exact identities, conditional local consequences, and finite-step verification.

Table 7: Navigation from the paper's central claims to their formal basis and empirical consequences.

Claim	**Formal status**	**Empirical consequence**
Forward secants expose projected differential geometry	Exact in the linearized probe span; centered finite-difference error is bounded under local smoothness (Props. 4 and 5)	Matched coordinate control tests the optimization value of response alignment, Section D.2
OM turns response geometry into an economical, residual-conditioned local representation	Probe-restricted optimality and residual-reachability identities (Props. 7 and 8)	Cross-task compactness/reachability diagnostics and rank-rule control, Sections D.4 and D.3
Mapped EKI implements a regularized residual correction	Exact linear EKI/Tikhonov identity and local Gauss–Newton interpretation (Props. 12 and 13)	Full-space EKI comparison and primary effectiveness results, Appendix F
OM constructs response-derived coordinates without a P-spanning ensemble	The algorithm first constructs the response representation from the unchanged $\mathcal{F}$; rollout counts and reduced solves then use K and r	Primary runs, rollout batching, and six-figure transfer, Appendix F and Sections D.1 and D.6
Map renewal is useful when observable geometry drifts	Conditional subspace-stability results (Props. 10 and 11)	Matched renewed/frozen intervention, Section D.5 and Figure 12
Fresh rollouts provide finite-step reliability	Verified-state descent invariant with an explicit prediction–realization agreement test	Component ablation and prediction–realization calibration, Table 6 and Figure 10

B ALGORITHMIC SPECIFICATION

Algorithm 1 specifies the OM-EKI loop developed in Section 4. Every evaluation queries the unchanged forward operator $\mathcal{F}$, making its structured output the optimization interface without requiring derivatives, adjoints, or a surrogate backward model. Geometry uses $e = \Gamma^{-1/2}[y - \mathcal{F}(\theta)]$, while verification uses the equivalent sign convention $\varepsilon_\Gamma = -e$.

Algorithm 1 OM-EKI with nonlinear verification and adaptive observable mappings

Require: initial parameters a, observation operator $\mathcal{F}$, target y, fixed geometry weighting Γ
Require: acceptance loss ℓ, EKI covariance schedule Ω_t, update budget T, task configuration

$(\theta^\star, \ell^\star) \leftarrow (a, \ell(a))$; $\lambda \leftarrow \lambda_0$; $t \leftarrow 0$; $Q_{\mathrm{mem}} \leftarrow \varnothing$
while $t < T$ **do**
  form orthonormal probes Q from retained directions Q_{mem} and random completion
  evaluate equation 13 and the normalized residual e
  if $\|e\|_2 = 0$ **then**
    return $\theta^\star$
  end if
  if the secant matrix has no admissible nonzero direction **then**
    terminate: no nonzero measured direction is available
  end if
  select realized coordinate dimension $r \le r_{\max}$ and construct A by equation 15
  generate a rotated regular simplex $Z \in \mathbb{R}^{(r+1)\times r}$ satisfying equation 42
  evaluate $f^{(i)} \leftarrow \mathcal{F}(a + Az^{(i)})$; store this verified state
  update $(\theta^\star, \ell^\star)$ with the lowest-loss verified member if it improves $\ell^\star$
  retain the configured leading mapping directions in Q_{mem}
  $h \leftarrow 0$; $c \leftarrow 0$
  while $t < T$ and $h < H$ **do**
    compute d_i by equation 46 and clip mapped-coordinate norms
    predict proposal observations with the reduced secant model
    evaluate every proposed member with the unchanged forward simulator
    compute ρ_t by equation 19
    if $\Delta_{\mathrm{pred}} > 0$, $\Delta_{\mathrm{act}} > 0$, finite ρ_t, and $\rho_t > \eta_{\mathrm{accept}}$ **then**
      accept proposal; replace verified state; $c \leftarrow 0$
      if $\rho_t \ge \eta_{\mathrm{strong}}$ **then**
        $\lambda \leftarrow \max(\lambda/2, \lambda_{\min})$
      end if
    else
      restore verified state; $c \leftarrow c + 1$; $\lambda \leftarrow \min(4\lambda, \lambda_{\max})$
    end if
    update $(\theta^\star, \ell^\star)$ with the lowest-loss verified member if it improves $\ell^\star$
    $t \leftarrow t + 1$; $h \leftarrow h + 1$
    if $c \ge c_{\max}$ **then**
      break
    end if
  end while
  $a \leftarrow \theta^\star$
end while
return $\theta^\star$

B.1 MAPPING CONSTRUCTION AND COORDINATE-CAPACITY RULE

All map geometry uses normalized observations: $\widetilde{S} = \Gamma^{-1/2} S$ and $e = \Gamma^{-1/2}[y - \mathcal{F}(a)]$. Let $\widetilde{S} = U\widehat{\Sigma}V^\top$ have numerical rank $q > 0$, and suppose $e \neq 0$. The sensitivity- and reachable-residual-based candidate dimensions are

$$r_E = \min\left\{k : \frac{\sum_{j=1}^{k} \hat{s}_j^2}{\sum_{j=1}^{q} \hat{s}_j^2} \geq \tau_E\right\}, \tag{20}$$

$$r_y = \min\left\{k : \frac{\sum_{j=1}^{k} (u_j^\top e)^2}{\sum_{j=1}^{q} (u_j^\top e)^2} \geq \tau_y\right\}, \qquad r = \min\{r_{\max}, q, \max(r_{\min}, r_E, r_y)\}. \tag{21}$$

The realized dimension r satisfies both response-energy and reachable-residual thresholds unless limited by $r_{\max}$. If the residual has no component in the measured span, we set $r_y = 1$ and record zero reachability; $q = 0$ reports a degenerate map, while $e = 0$ terminates optimization.

The r_y criterion allocates capacity within the residual component reached by all q measured directions. Its coverage of the full residual is recorded separately by

$$\widehat{R}_r = \frac{\sum_{j=1}^{r} (u_j^\top e)^2}{\|e\|_2^2}, \tag{22}$$

the empirical counterpart of R_r. Response energy controls economy, while r_y prevents truncation before the measured span represents the current correction.

At reconstruction c, define the retained secant scale

$$\nu_c = \|\widetilde{S}_c V_{c,r}\|_F = \left(\sum_{j=1}^{r} \hat{s}_{c,j}^2\right)^{1/2}. \tag{23}$$

At the first reconstruction, the reference observable radius and probe radius are fixed as

$$\rho_\star = c_\theta \nu_0, \qquad \delta = c_\theta \sigma, \tag{24}$$

where $c_\theta > 0$ caps parameter-space scale and $\sigma > 0$ is the simplex scale. Later mappings reuse $\rho_\star$, preventing coordinate units from changing with aggregate local sensitivity. The scale in equation 15 is

$$\alpha_c = \min\left\{c_\theta, \frac{\rho_\star}{\max(\nu_c, \epsilon_{\text{trace}})}\right\}. \tag{25}$$

Here $\epsilon_{\text{trace}} > 0$ is a numerical floor; a zero-rank secant matrix triggers the termination case above.

For reconstruction $c > 0$, the first m probe columns retain the previous leading unscaled basis directions. The retained columns are orthonormalized and rank-deficient columns removed; projected Gaussian completion and QR supply the remainder. Simplex and Kalman anomalies are reinitialized. After H local updates or $c_{\max}$ consecutive rejections, the map is reconstructed at the best verified parameter vector. Probe and simplex rotations use independent reproducible random streams.

B.2 Computational structure and fixed settings

Table 8 separates fixed OM-EKI rules from task-specific capacity and numerical settings; Appendix E supplies solvers, data splits, and random seeds.

Table 8: Fixed OM-EKI rules and task-specific numerical settings.

Quantity	Symbol	Setting
Fixed OM-EKI logic		
Secant-energy fraction	τ_E	0.99
Reachable-residual fraction	τ_y	0.99
Relative numerical-rank tolerance	ϵ_{rank}	10^{-6}
Retained-scale numerical floor	$\epsilon_{\mathrm{trace}}$	10^{-12}
Nonlinear verification interval	k	every update ($k = 1$)
Acceptance threshold	η_{accept}	0.05
Strong-agreement threshold	η_{strong}	0.75
Accepted/rejected λ multiplier		$1/2$ (strong) / 4 (rejected)
Early-rebuild rejection count	$c_{\max}$	2
Task-specific capacity and numerical settings		
Task-specific fields		$K, r_{\min}, r_{\max}, H, m, \sigma, c_\theta,$ $\lambda_0, \lambda_{\min}, \lambda_{\max}, \Omega_t$ schedule, and chunk sizes

Each reconstruction evaluates one anchor and $2K$ paired probes, then initializes an $(r+1)$-member ensemble; each local update evaluates $N = r+1$ fresh proposals. Response factorization uses a $K \times K$ Gram matrix, and each Kalman update solves an $N \times N$ positive-definite system. Computation therefore remains in measured response and ensemble spaces, avoiding dense linear algebra on the P-dimensional parameter axis. Observation chunks accumulate the same sufficient statistics while controlling peak memory. The forward-evaluation counts and reduced linear-algebra dimensions are thus governed by the measured capacities K and r, rather than an ensemble spanning the ambient parameterization.

B.3 Nonlinear verification and the verified-state invariant

The acceptance rule in Algorithm 1 preserves a verified-state invariant: ensemble-mean loss decreases on acceptance and is restored on rejection, while the best verified loss is nonincreasing. Fresh simulator evaluations therefore connect each local prediction to finite progress under the actual model.

C Mathematical Basis

The mathematical development follows the method: the observation operator induces an observable quotient and optimal local reduction, centered secants recover probe-restricted differential geometry, the geometry diagnostics control the retained space, and ensemble-space EKI implements regularized residual correction. Throughout this subsection, define the normalized response map $\mathcal{G}(\theta) = \Gamma^{-1/2}\mathcal{F}(\theta)$ and $J = D\mathcal{G}(a) = \bar{J}(a)$. Geometry uses Q for this normalized map, whereas the EKI identities whiten observations with Ω_t. Positive-definite normalization preserves the Jacobian kernel, and finite forward responses realize these local constructions.

C.1 Observable quotient and optimal local reduction

Proposition 1 (Exact observable quotient of the linearized problem)**.** *Let $\mathcal{G} : \mathbb{R}^P \to \mathbb{R}^M$ be differentiable at a, let $J = D\mathcal{G}(a)$ have rank q, and let $B_q \in \mathbb{R}^{P\times q}$ have orthonormal columns spanning $\mathcal{R}(J^\top)$. Then, for every parameter perturbation v,*

$$Jv = JB_qB_q^\top v. \tag{26}$$

Consequently the linearized parameter-to-normalized-response map factors through a q-dimensional coordinate $z = B_q^\top v$, and $\mathbb{R}^P/\ker J$ is isomorphic to $\mathcal{R}(J^\top)$.

Proof. The fundamental theorem of linear algebra gives the orthogonal decomposition $\mathbb{R}^P = \mathcal{R}(J^\top) \oplus \ker J$. Hence $v = B_qB_q^\top v + v_0$ for some $v_0 \in \ker J$. Applying J eliminates v_0 and proves equation 26. The quotient removes exactly $\ker J$ and therefore has dimension $\mathrm{rank}(J) = q$.

Proposition 2 (Optimal rank-r observable approximation)**.** *Let $J = U\Sigma V^\top$ have nonzero singular values $\sigma_1 \geq \cdots \geq \sigma_q > 0$. For $r < q$,*

$$\min_{B^\top B = I_r} \|J(I - BB^\top)\|_F^2 = \sum_{i=r+1}^{q} \sigma_i^2, \tag{27}$$

and the minimum is attained by $B = V_r$, the leading r right singular vectors. Equivalently, V_r maximizes the aggregate first-order normalized-response sensitivity $\|JB\|_F^2$ among rank-r orthonormal parameter mappings.

Proof. Because $\|J(I - BB^\top)\|_F^2 = \|J\|_F^2 - \|JB\|_F^2$, minimizing the omitted observation energy is equivalent to maximizing $\mathrm{tr}(B^\top J^\top JB)$. The Rayleigh–Ritz principle selects the leading eigenvectors of $J^\top J$, giving the stated tail sum.

Proposition 3 (Second-order discrepancy after observable projection)**.** *Under the assumptions of Proposition 1, suppose $D\mathcal{G}$ is L_J-Lipschitz on the segments from a to $a + v$ and from a to $a + w$, where $w = B_q B_q^\top v$. Then*

$$\|\mathcal{G}(a + v) - \mathcal{G}(a + w)\|_2 \leq \frac{L_J}{2}\big(\|v\|_2^2 + \|w\|_2^2\big) \leq L_J\|v\|_2^2. \tag{28}$$

Proof. Taylor expansion at a gives $\mathcal{G}(a+v) = \mathcal{G}(a) + Jv + R_v$ and $\mathcal{G}(a+w) = \mathcal{G}(a) + Jw + R_w$, with $\|R_v\|_2 \leq L_J\|v\|_2^2/2$ and the analogous bound for R_w. Proposition 1 gives $Jv = Jw$, so only the two remainders remain. Finally, orthogonal projection implies $\|w\|_2 \leq \|v\|_2$.

These propositions establish the ideal construction behind OM: the linearized normalized response operator factors exactly through its observable quotient, its leading singular directions give the best rank-r response approximation, and the corresponding nonlinear response discrepancy is second order. The next subsection connects this ideal geometry to the finite probes used by the algorithm.

C.2 Forward secants and probe-restricted geometry

For a smooth Neural ODE flow, parameter sensitivity $S_\theta(t) = \partial\Phi_{0:t}^\theta(x_0)/\partial\theta$ satisfies

$$\dot{S}_\theta(t) = D_x f_\theta(x(t), t)S_\theta(t) + D_\theta f_\theta(x(t), t), \qquad S_\theta(0) = 0. \tag{29}$$

After composition with the observation map and normalization this gives $J = D\mathcal{G}(a) = \bar{J}(a)$. The equation identifies the classical sensitivity represented by the secants; OM-EKI measures its action through forward rollouts and never solves the sensitivity equation. The propositions below apply directly to the resulting normalized response operator $\mathcal{G}$, including non-ODE simulators wherever the stated local regularity holds.

Proposition 4 (Centered secant error)**.** *Let $\mathcal{G} : \mathbb{R}^P \to \mathbb{R}^M$ be three times continuously differentiable on a neighborhood of $\{a + sq : |s| \leq \delta\}$, with $\|q\|_2 = 1$ and $\|D^3\mathcal{G}(a + sq)[q, q, q]\|_2 \leq L_3$. Then*

$$\left\| \frac{\mathcal{G}(a + \delta q) - \mathcal{G}(a - \delta q)}{2\delta} - D\mathcal{G}(a)q \right\|_2 \leq \frac{L_3}{6}\delta^2. \tag{30}$$

Proof. Apply the one-dimensional Taylor formula to $g(s) = \mathcal{G}(a + sq)$ at $s = \delta$ and $s = -\delta$. Subtraction cancels even-order terms and the bound follows after division by 2δ.

If each normalized rollout response has deterministic error at most ε_F, the centered quotient adds at most ε_F/δ to the smooth truncation bound. This separates finite-difference bias from evaluation error and makes the probe-radius tradeoff explicit.

Proposition 5 (Projected first- and second-order recovery)**.** *Let $Q^\top Q = I$, $P_Q = QQ^\top$, and $H = JQ + E$ with $\|E\|_2 \leq \varepsilon$. Then, for every $u \in \mathbb{R}^M$,*

$$\|QH^\top u - P_Q J^\top u\|_2 \leq \varepsilon\|u\|_2, \tag{31}$$

$$\|H^\top H - Q^\top J^\top JQ\|_2 \leq 2\|JQ\|_2\varepsilon + \varepsilon^2. \tag{32}$$

For the normalized centered secant matrix under Proposition 4, one may take $\varepsilon \leq \sqrt{K}L_3\delta^2/6$.

Proof. The first difference is $QE^\top u$ and $\|Q\|_2 = 1$. For the second, expand $(JQ + E)^\top (JQ + E) - Q^\top J^\top JQ$ and apply submultiplicativity. The final bound follows from the Frobenius norm of the K column-wise secant errors.

Proposition 6 (Error of the deployed reduced observation model)**.** *Let $S = JQ+E_\delta$, let $A = \alpha QV_r$, and assume $D\mathcal{G}$ is L_J-Lipschitz on the segment from a to $a + Ad$. Then the normalized response model used by OM-EKI satisfies*

$$\|\mathcal{G}(a + Ad) - \mathcal{G}(a) - \alpha SV_r d\|_2 \le \frac{L_J}{2}\|Ad\|_2^2 + \alpha\|E_\delta\|_2\|d\|_2. \tag{33}$$

Under the assumptions of Proposition 4 for all K probes, $\|E_\delta\|_2 \le \sqrt{K}L_3\delta^2/6$.

Proof. Add and subtract $JAd = \alpha JQV_r d$. The nonlinear Taylor remainder is at most $L_J\|Ad\|_2^2/2$, and the remaining linear discrepancy is $-\alpha E_\delta V_r d$. Since V_r has orthonormal columns, the displayed bound follows. The final claim follows by bounding the Frobenius norm of the matrix whose K columns obey Proposition 4.

For the normalized residual $e(\theta) = \Gamma^{-1/2}[y - \mathcal{F}(\theta)]$, define $\widehat{H} = \alpha SV_r$, and $\varepsilon_{\mathrm{sec}} = \alpha\|E_\delta\|_2$. The same result and the triangle inequality give

$$\|e(a + Ad)\|_2 \le \|e(a) - \widehat{H}d\|_2 + \varepsilon_{\mathrm{sec}}\|d\|_2 + \frac{L_J}{2}\|Ad\|_2^2,$$

which is equation 12. Hence a predicted reduction that exceeds the secant and nonlinear remainder is a sufficient local descent condition. The mapped step scale controls the nonlinear term, the centered probe radius controls the secant term, and fresh rollout evaluation tests the condition without knowing the constants.

Proposition 7 (Optimality inside the probe span)**.** *Let $Q^\top Q = I$, $H = JQ$, and let V_r contain the leading r right singular vectors of H. Then*

$$V_r \in \arg\max_{V^\top V = I_r} \|HV\|_F^2, \qquad B = QV_r, \tag{34}$$

and B maximizes $\|J\widetilde{B}\|_F^2$ over orthonormal rank-r mappings in $\mathrm{col}(Q)$.

Proof. $\|HV\|_F^2 = \mathrm{tr}(V^\top H^\top HV)$, so Rayleigh–Ritz gives the result. Substitution of $\widetilde{B} = QV$ gives the second statement.

Thus the response-aligned basis is optimal inside the sampled parameter span for the normalized response. For measured secants, Proposition 4 controls the normalized operator perturbation, while spectral separation controls the stability of individual leading directions.

Proposition 8 (Linearized residual-reachability identity)**.** *Let $e = \Gamma^{-1/2}[y - \mathcal{F}(a)]$, $\bar{J} = \Gamma^{-1/2}D\mathcal{F}(a)$, and let $\Pi(B)$ be the orthogonal projector onto $\mathcal{R}(\bar{J}B)$. For $e \neq 0$, define $c(B) = \|\Pi(B)e\|_2^2/\|e\|_2^2$. Then*

$$\min_z \|e - \bar{J}Bz\|_2^2 = \|e\|_2^2[1 - c(B)]. \tag{35}$$

For two mappings B_0 and B_1 evaluated at the same anchor,

$$\min_z \|e - \bar{J}B_0 z\|_2^2 - \min_z \|e - \bar{J}B_1 z\|_2^2 = \|e\|_2^2[c(B_1) - c(B_0)]. \tag{36}$$

Thus, at a fixed anchor, increased residual reachability (residual capture) produces an exactly quantified reduction in the best attainable linearized residual; nonlinear verification determines whether that gain persists for the finite simulator step.

Proof. Least squares decomposes e orthogonally into $\Pi(B)e + (I - \Pi(B))e$. The fitted component is $\Pi(B)e$ and the minimum residual is $(I - \Pi(B))e$. Pythagoras gives equation 35; subtracting the identity for two mappings gives equation 36.

Joint interpretation of the geometry diagnostics.

Proposition 9 (Local implications of compactness, reachability, and drift)**.** *Let $\bar{J} = U\Sigma V^\top$, let $B_r = V_r$, let $P_r = B_r B_r^\top$, and let $e \neq 0$ and $g = \bar{J}^\top e$. Then*

$$\|(I - P_r)g\|_2 \le \sigma_{r+1}\|e\|_2 \le \sqrt{1 - C_r}\,\|\bar{J}\|_F\|e\|_2,$$

where $\sigma_{r+1} = 0$ if $r \ge \operatorname{rank}(\bar{J})$. Moreover,

$$\min_z \|e - \bar{J}B_r z\|_2^2 = (1 - R_r)\|e\|_2^2.$$

Finally, if B and B' have the same rank r, with projectors P and P' and drift $D = [1 - \|B^\top B'\|_F^2/r]^{1/2}$, then for every v,

$$\|(P' - P)v\|_2 \le \sqrt{r}\,D\|v\|_2.$$

Proof. The omitted gradient is $V_{>r}\Sigma_{>r}U_{>r}^\top e$, whose norm is at most $\sigma_{r+1}\|e\|_2$. Since $\sigma_{r+1}^2 \le \sum_{i>r}\sigma_i^2 = (1 - C_r)\|\bar{J}\|_F^2$, the first bound follows. The second identity is Proposition 8. For the last statement, the principal angles $\{\vartheta_i\}_{i=1}^r$ satisfy $rD^2 = \sum_i \sin^2\vartheta_i$ and $\|P' - P\|_2 = \max_i \sin\vartheta_i \le \sqrt{r}\,D$.

The definition in equation 8 also permits the selected coordinate dimension to change. The following decomposition separates this change from rotation of the common part of the two spaces.

Proposition 10 (Unequal-dimension drift decomposition)**.** *Let $B \in \mathbb{R}^{P\times r}$ and $B' \in \mathbb{R}^{P\times s}$ have orthonormal columns, let $P = BB^\top$, $P' = B'B'^\top$, and let $m = \min(r, s) > 0$. With*

$$D = \left[1 - \frac{\|B^\top B'\|_F^2}{m}\right]^{1/2},$$

the projector discrepancy satisfies

$$\|P' - P\|_F^2 = |r - s| + 2mD^2, \qquad \|(P' - P)v\|_2 \le \sqrt{|r - s| + 2mD^2}\,\|v\|_2. \tag{37}$$

Thus D measures angular disagreement across the m comparable directions, while $|r - s|$ records a separate change in represented capacity. When $r = s$, the principal-angle argument in Proposition 9 gives the sharper operator bound used in equation 11.

Proof. Using $P^2 = P$, $P'^2 = P'$, and cyclicity of the trace,

$$\|P' - P\|_F^2 = \operatorname{tr}(P') + \operatorname{tr}(P) - 2\operatorname{tr}(PP') = r + s - 2\|B^\top B'\|_F^2.$$

Substituting $\|B^\top B'\|_F^2 = m(1 - D^2)$ gives the identity because $r + s - 2m = |r - s|$. The vector bound follows from $\|P' - P\|_2 \le \|P' - P\|_F$.

The normalization in equation 15 has a direct algebraic consequence. For the normalized secant matrix $\widetilde{S} = U\widehat{\Sigma}V^\top$, let $\nu = \|\widetilde{S}V_r\|_F$, $B = QV_r$, and $A = \alpha B$. Then

$$\|\alpha\widetilde{S}V_r\|_F = \alpha\nu \le \rho_\star, \qquad \|Ad\|_2 = \alpha\|d\|_2 \le c_\theta\|d\|_2. \tag{38}$$

If neither the cap nor the numerical floor is active, $\alpha = \rho_\star/\nu$ and hence

$$\|\alpha\widetilde{S}V_r\|_F^2 = \rho_\star^2. \tag{39}$$

These identities follow from equation 25 and $B^\top B = I$. They quantify the change predicted by the secant model, while nonlinear verification measures the corresponding finite observation change. The scalar scaling preserves the relative retained singular spectrum without individually amplifying weak directions.

Proposition 11 (Local stability under anchor motion)**.** *Write $\bar{J}(a) = \Gamma^{-1/2}D\mathcal{F}(a)$ and $G(a) = \bar{J}(a)^\top\bar{J}(a)$. Suppose $\|\bar{J}(\theta)\|_2 \le \beta$ and $\|\bar{J}(a') - \bar{J}(a)\|_2 \le L_J\|a' - a\|_2$ on a neighborhood. Then*

$$\|G(a') - G(a)\|_2 \le 2\beta L_J\|a' - a\|_2. \tag{40}$$

If the rth and $(r + 1)$st eigenvalues of $G(a)$ are separated by $\gamma > 0$ and the right-hand side is at most $\gamma/2$, the leading ideal observable subspaces obey

$$\|\sin\Theta(B_r(a'), B_r(a))\|_2 \le \frac{4\beta L_J}{\gamma}\|a' - a\|_2. \tag{41}$$

Proof. Expand $G(a') - G(a)$ into $\bar{J}(a')^\top[\bar{J}(a') - \bar{J}(a)] + [\bar{J}(a') - \bar{J}(a)]^\top \bar{J}(a)$ and apply submultiplicativity to obtain equation 40. equation 41 is the Davis–Kahan invariant-subspace perturbation bound under the stated gap condition (Davis & Kahan, 1970).

Propositions 9–11 connect the operator-to-path criterion to optimization: compactness controls economical representation, reachability gives the attainable linearized correction, and drift bounds the loss of persistence along the path. The stability result separates reusable from rapidly changing observable spaces.

C.3 Minimal full-rank simplex

A centered N-member ensemble has anomaly rank at most $N-1$, so full covariance rank in $\mathbb{R}^r$ requires $N \geq r+1$. OM-EKI takes the minimal choice $N = r+1$. If $U \in \mathbb{R}^{(r+1)\times r}$ has orthonormal columns perpendicular to $\mathbf{1}$, then $Z = \sigma\sqrt{r}\,UR$ for any orthogonal R satisfies

$$\mathbf{1}^\top Z = 0, \qquad \frac{1}{N-1} Z^\top Z = \sigma^2 I_r. \tag{42}$$

Random rotation changes the ensemble orientation without changing these moments. The minimal $r+1$ construction therefore gives full isotropic covariance in the mapped space at the lowest possible ensemble size.

C.4 Ensemble-space Kalman identity

For a fixed EKI update, let $\Omega \succ 0$ denote its observation covariance. Let $X \in \mathbb{R}^{N\times r}$ and $Y \in \mathbb{R}^{N\times M}$ be row-oriented coordinate and whitened-observation anomalies, with

$$X_{i:} = \frac{(z^{(i)} - \bar{z})^\top}{\sqrt{N-1}}, \qquad Y_{i:} = \frac{[\Omega^{-1/2}(f^{(i)} - \bar{f})]^\top}{\sqrt{N-1}}. \tag{43}$$

With $r_i = \Omega^{-1/2}(f^{(i)} - y)$, the covariance-form update is

$$d_i = -X^\top Y (Y^\top Y + \lambda I_M)^{-1} r_i. \tag{44}$$

The push-through identity

$$Y(Y^\top Y + \lambda I)^{-1} = (YY^\top + \lambda I)^{-1} Y \tag{45}$$

gives the ensemble-space update

$$d_i = -X^\top (YY^\top + \lambda I_N)^{-1} Y r_i, \qquad z^{(i)}_{\text{prop}} = z^{(i)} + d_i. \tag{46}$$

The forms are exactly equal for $\lambda > 0$; ensemble space changes the linear algebra, not the optimizer. It stores no dense covariance on the P-dimensional parameter axis. Observation residuals may be accumulated in chunks because the sufficient statistics are sums of $YY^\top$ and Yr_i terms.

Proposition 12 (EKI as regularized residual fitting)**.** *For $\lambda > 0$, define*

$$w_i^\star = \arg\min_{w\in\mathbb{R}^N} \left\{ \frac{1}{2}\|Y^\top w + r_i\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \right\}. \tag{47}$$

Then $w_i^\star = -(YY^\top + \lambda I)^{-1} Y r_i$ and the OM-EKI direction is $d_i = X^\top w_i^\star$. Consequently, the linearized residual is fit only by directions represented in the mapped coordinate-anomaly span, with λ controlling the coefficient norm.

Proof. The objective in equation 47 is strictly convex. Its normal equation is $(YY^\top + \lambda I)w = -Yr_i$. Solving it and multiplying by $X^\top$ gives equation 46.

This makes precise why EKI is useful here: it solves a vector-residual ridge problem from empirical forward covariation instead of estimating a gradient from scalar losses. It also exposes the classical EKI subspace property and the role of regularization (Chada et al., 2018; 2020).

Proposition 13 (Local Gauss–Newton form of isotropic mapped EKI)**.** *Let $g(z) = \Omega^{-1/2}\mathcal{F}(a+Az)$ be affine over the ensemble, $g(z) = g(0)+J_z z$, and let the coordinate ensemble have mean zero and covariance $C^{zz} = \sigma^2 I_r$ with $\sigma > 0$. For the normalized target $\widetilde{y} = \Omega^{-1/2}y$, residual $e = \widetilde{y} - g(0)$, and $\lambda > 0$, the covariance-form EKI correction of the ensemble mean is*

$$\Delta z = \sigma^2 J_z^\top (\sigma^2 J_z J_z^\top + \lambda I)^{-1} e = (J_z^\top J_z + \lambda\sigma^{-2} I)^{-1} J_z^\top e. \tag{48}$$

It is therefore the unique minimizer of

$$\min_d \left\{ \frac{1}{2}\|J_z d - e\|_2^2 + \frac{\lambda}{2\sigma^2}\|d\|_2^2 \right\}. \tag{49}$$

Proof. For the affine operator, $C^{zf} = C^{zz}J_z^\top = \sigma^2 J_z^\top$ and $C^{ff} = J_z C^{zz} J_z^\top = \sigma^2 J_z J_z^\top$. Substitution into the covariance-form EKI update gives the first expression. The push-through identity gives the second. It is also the normal-equation solution of equation 49.

Proposition 13 gives the exact correspondence in equation 18 for an affine mapped response with isotropic coordinate covariance, which the regular simplex provides at reconstruction. Subsequent nonlinear steps use evolving empirical ensemble covariances: they retain the measured coordinate system but are not exact Gauss–Newton iterations. Fresh nonlinear evaluation tests whether the local residual model remains accurate away from the anchor.

The resulting chain is constructive: forward secants measure the probe-restricted operator, its factorization selects an economical response-aligned representation, reachability determines whether that representation can correct the current residual, and mapped EKI solves the corresponding regularized residual problem. Reconstruction maintains its validity along the path, while nonlinear verification retains only progress realized by the unchanged simulator.

D Controlled Mechanism and Extended Evidence

This section separates the three levels of the mechanism: response alignment is the construction principle; compactness, residual reachability, and geometry drift are representation properties; and nonlinear verification is the finite-step safeguard. Rollout batching supplies computational evidence, controlled and transfer experiments examine the mechanism, and PDEControlGym tests its practical consequence in an external closed-loop PDE-control benchmark.

D.1 Single-GPU rollout batching

OM-EKI's independent rollouts admit single-GPU batching without changing the objective, ensemble, coordinate dimension, or evaluation budget; unlike data minibatching, batching only chunks a fixed parameter population. The POLLU control uses the hardware in Appendix E, coordinate dimension 128, a 129-member population, five updates, and 1,032 evaluations. Replaying identical populations across batch sizes reproduces observations within 4.70×10^{-8} relative error.

Table 9: Single-GPU rollout batching for POLLU. Speedup is relative to batch 8 for the same end-to-end OM-EKI workload; validation differs by at most 1.11% from the reference.

Batch	Time (s) ↓	Sec./update ↓	Speedup ↑	Validation ↓	Peak CUDA (MiB)
8	724.24	144.85	1.00×	0.001799	72.57
16	334.49	66.90	2.17×	0.001819	72.65
32	167.31	33.46	4.33×	0.001819	72.65
64	**104.51**	**20.90**	**6.93×**	0.001814	72.65

The fixed workload becomes $6.93\times$ faster from batch 8 to 64, showing how independent forward rollouts translate into usable single-device parallelism.

D.2 MATCHED COORDINATE-CHOICE CONTROL

This control separates response alignment from dimension reduction by matching the reduced dimension, parameter anchor, perturbation scale, update logic, and forward budget. Let $u = (x, v)$ and let $\varphi(t) \in \mathbb{R}^{16}$ collect normalized constant and Fourier time features. With a fixed orthonormal mixing $U \in \mathbb{R}^{128\times16}$, the controlled dynamics are

$$\dot{x} = v, \qquad \dot{v} = -0.2v - x - \kappa x^3 + \theta^\top U\varphi(t). \tag{50}$$

Thus 128 parameters act through 16 forcing dimensions; a generic random-32 span remains full-rank through U. The trajectory is affine in θ at $\kappa = 0$; increasing κ adds nonlinearity without changing the parameterization, forcing basis, or observations.

Problem seed 0 uses three training and two disjoint validation trajectories, sampled 61 times on $[0, 8]$ with fixed-step RK4. OM-EKI uses 64 probes to construct 32 observable coordinates; Random-coordinate EKI (the main text's EKI (rand)) searches a fixed 32-dimensional random orthonormal span generated independently of OM, and Full-space EKI (the main text's Direct EKI (full)) searches all 128 parameters. All variants share the target, parameter scale, and nonlinear acceptance, and their evaluation totals differ by at most 1.68% within each κ. Table 4 reports $\kappa = 0$ and 4, so the dimension-matched comparison isolates response alignment.

D.3 RESIDUAL-AWARE COORDINATE-CAPACITY CONTROL

This control changes only the retained-capacity rule on deterministic Neural ODE seed 5. Both variants use 64 probes, cap 64, identical probe construction, nonlinear verification, and nearly identical forward budgets (16,270 versus 16,268 evaluations; 0.012%). Energy-only capacity follows r_E in equation 21; adding r_y yields final dimensions 6 and 60. Table 5 thereby isolates the value of allocating measured capacity to the current residual.

Finite-step component protocol. The component intervention in Table 6 uses problem seed 3, 3,000 attempted updates, 64 probes, and cap 64. The full method renews its map every 50 local updates and verifies every proposal; the static arm retains its initial map, and the unconditional arm accepts every proposal. Their realized dimensions are 64, 46, and 61, with 205,642, 141,177, and 183,385 evaluations, so the comparison includes each component's effect on dimension and forward work.

D.4 CROSS-TASK REACHABILITY DIAGNOSTICS

This diagnostic asks whether compact, locally predictive responses suffice when the coordinates reach little of the current residual. Figure 4(a) shows one run per listed task, and Table 10 gives the complete values. For nonzero singular values $\hat{s}_1 \geq \cdots \geq \hat{s}_q$, let $r_{95} = \min\{r : \widehat{C}_r \geq 0.95\}$, where $\widehat{C}_r = \sum_{j=1}^{\min(r,q)} \hat{s}_j^2 / \sum_{j=1}^{q} \hat{s}_j^2$. We report median r_{95}/K and $\widetilde{R} = \mathrm{median}_c\, \widehat{R}_{r_c}$ across map anchors. The two primary dynamical-system runs use prespecified representative seeds (deterministic 4 and POLLU 0) and exclude SHD; the DynaBench transfer rows use prespecified seed 21 and extend the comparison toward lower reachability.

Local-model fidelity is measured by $\chi_{\mathrm{nl}} = \|F_+ + F_- - 2\mathbf{1}\mathcal{F}(a)^\top\|_F / \|F_+ - F_-\|_F$, where the rows of $F_\pm$ are the paired responses $\mathcal{F}(a \pm \sigma q_k)^\top$, and by the agreement ratio $\rho = \Delta_{\mathrm{act}}/\Delta_{\mathrm{pred}}$. Table 10 reports low χ_{nl} and agreement ratios near one, so the local models remain predictive despite capturing only 10.4% and 5.0% of the full residual.

Table 10: Cross-task diagnostics underlying Figure 4(a). Here $\widetilde{\chi}_{\mathrm{nl}}$ is the median centered nonlinearity ratio and $\widetilde{\rho}$ the median realized-to-predicted agreement ratio; lower $\widetilde{\chi}_{\mathrm{nl}}$ and $\widetilde{\rho}$ near one indicate an accurate local response model.

Task	P	r_{95}/K	$\widetilde{R}$	$\widetilde{\chi}_{\mathrm{nl}}$	$\widetilde{\rho}$
Primary tasks					
Deterministic ODE	252	0.094	0.716	0.014	0.629
POLLU UDE	19,587	0.023	0.266	0.010	0.969
Lower-reachability transfer tasks					
DynaBench reaction–diffusion	28,930	0.129	0.104	0.033	1.015
DynaBench Burgers	28,930	0.115	0.050	0.089	0.964

In these representative runs, the two primary dynamical-system runs shown here combine compact responses with substantially higher reachability. This separates local model quality and representation economy from the fraction of the current correction that the coordinates can express. For the DynaBench tasks, reachability is measured on the one-step training residual, while a 16-step held-out rollout selects checkpoints.

D.5 Stationary-geometry reconstruction control

This control tests whether map renewal is useful when observable geometry remains stationary. Unlike the component result in Table 6, a common-probe Duffing comparison holds the probe span, simplex resets, 32 coordinates, and update protocol fixed, so recomputing the observable basis is the only difference.

Table 11: Strict common-probe reconstruction control. Both variants use the same update protocol at every continuation value.

κ	Renewed-map MSE ↓	Frozen-map MSE ↓
0	8.652×10^{-13}	7.744×10^{-13}
0.25	5.206×10^{-13}	4.936×10^{-13}
1	1.168×10^{-12}	1.079×10^{-12}
2	3.314×10^{-12}	5.077×10^{-12}
4	2.368×10^{-11}	2.477×10^{-11}

The variants are effectively tied. With forcing dimension 16 and map dimension 32, the leading observable subspace remains nearly stationary under the shared probe span even at $\kappa = 4$. Hence $\mathcal{R}(\bar{J}_t B_0) \approx \mathcal{R}(\bar{J}_t B_t)$, giving nearly identical attainable local residuals by Proposition 8. Proposition 11 identifies this regime when anchor-induced Jacobian variation is small relative to the observable eigengap. Together with the main ablation, the control separates useful map renewal from unnecessary renewal.

Controlled moving-geometry complement. The complementary intervention isolates when renewal becomes valuable by increasing geometric drift while preserving compactness and new-map reachability. Here $\kappa = 0$, $\varphi(t) \in \mathbb{R}^{32}$, and $U_0, U_1 \in \mathbb{R}^{128\times32}$ span mutually orthogonal parameter subspaces. The additive term in equation 50 is replaced by

$$b_{\gamma,\theta}(t,x) = \cos(\gamma x)\,\theta^\top U_0\varphi(t) + \sin(\gamma x)\,\theta^\top U_1\varphi(t), \tag{51}$$

so γ rotates state-dependent parameter sensitivity while leaving ambient dimension and rollout protocol fixed.

Geometry is evaluated at five anchors $a_j = (1-\tau_j)a_0 + \tau_j\theta^\star$, $\tau_j \in \{0, 0.25, 0.5, 0.75, 1\}$, using 128 probes and rank 16. For basis $B_{\gamma,j}$, drift is the median transition value of $[1 - \|B_{\gamma,j}^\top B_{\gamma,j+1}\|_F^2/16]^{1/2}$; two independent same-anchor reconstructions estimate its noise floor. The

displayed $\widehat{C}_{16}$ and $\widehat{R}_{16}$ are also anchor medians. Table 12 reports terminal training MSE under a fixed attempted-update budget to isolate matched-budget residual correction.

Table 12: Controlled drift-condition closure. Every renewed/frozen pair uses exactly 9,319 objective evaluations.

γ	$\widehat{C}_{16}$	$\widehat{R}_{16}$	D	Renewed MSE ↓	Frozen MSE ↓	Frozen/ renewed
0	0.990	0.979	0.000	0.002468	0.002468	1.00×
1.5	0.969	0.956	0.021	0.006121	0.006121	1.00×
3	0.968	0.943	0.107	0.004500	0.004920	1.09×
5	0.967	0.984	0.265	0.004872	0.017514	3.59×
6	0.964	0.986	0.282	0.004119	0.015330	3.72×

Drift increases monotonically with γ (rank correlation 1.000) and, for nonzero rotation, lies well above the largest same-anchor noise estimate of 4.34×10^{-4}. Both arms compute the same candidate reconstructions, which the frozen arm discards, preserving identical evaluation budgets.

D.6 High-dimensional NeuralPDE transfer

This transfer tests fixed measured-coordinate capacity as ambient parameter dimension grows. We use the official low-resolution DynaBench data (Dulny et al., 2023) and circular-CNN NeuralPDE vector field (Dulny et al., 2022). At problem seed 21, 64 windows define the one-step training observation and 32 held-out windows select checkpoints by 16-step rollout. Burgers widths 32 and 64 give $P = 28{,}930$ and $P = 113{,}154$; Adam receives 3,200 steps and OM-EKI 1,200 updates with coordinate dimension 192.

Table 13: Extended DynaBench Burgers rollout evaluated on validation trajectories.

Width	Method	P	Optimizer dim.	16-step MSE ↓	MSE / persistence ↓
32	Adam	28,930	28,930	0.12838	0.6685
32	**OM-EKI**	28,930	192	0.16926	0.8814
64	Adam	113,154	113,154	0.12176	0.6340
64	**OM-EKI**	113,154	192	0.13537	0.7049

Both methods beat persistence. With coordinate dimension fixed at 192 as P nearly quadruples, the OM-EKI gap to Adam narrows from 31.85% to 11.18%. These runs support the use of a fixed 192-coordinate capacity at six-figure P; the lower reachability in Figure 4(a) is consistent with the remaining quality gap.

D.7 External PDE-control transfer

This experiment tests external transfer from trajectory fitting to closed-loop PDE policy optimization on the standard PDEControlGym reaction–diffusion boundary-control environment (Bhan et al., 2024).

Benchmark and protocol. The controlled plant is

$$\partial_t u = \partial_{xx} u + \beta(x)u, \qquad u(0,t) = 0, \quad u(1,t) = U(t), \qquad \beta(x) = 50\cos(8\arccos x)\,. \tag{52}$$

For episode e and control step k, let $u_{e,k}(\theta) \in \mathbb{R}^{201}$ be the state field and $U_{e,k}(\theta)$ the applied boundary action. We define

$$s_{e,k}(\theta) = \left[\max\left\{201^{-1}\|u_{e,k}(\theta)\|_2^2, 10^{-20}\right\}\right]^{1/2}, \qquad c_{e,k}(\theta) = \sqrt{\lambda_u}\,\frac{U_{e,k}(\theta)}{U_{\max}}, \tag{53}$$

where $\lambda_u = 10^{-3}$ and $U_{\max} = 20$. With $E_{\mathrm{tr}} = 4$ training episodes and $N_c = 1000$ control decisions, the structured training output is

$$\mathcal{F}_{\mathrm{tr}}(\theta) = \mathrm{vec}_{e=1}^{E_{\mathrm{tr}}}\left[s_{e,1:N_c}(\theta), c_{e,1:N_c}(\theta)\right] \in \mathbb{R}^M, \qquad M = 2E_{\mathrm{tr}}N_c = 8000. \tag{54}$$

The target is $y = 0$, and the channel scaling is embedded in $\mathcal{F}_{\mathrm{tr}}$, so $\Gamma = I_M$ and $e(\theta) = -\mathcal{F}_{\mathrm{tr}}(\theta)$. The shared scalar objective is $\ell(\theta) = M^{-1}\|\mathcal{F}_{\mathrm{tr}}(\theta)\|_2^2$. Thus the response retains episode, time, and state-versus-control structure before scalarization.

We retain the standard configuration: $T = 1$, $\Delta t = 10^{-5}$, $\Delta x = 5 \times 10^{-3}$, full-state sensing, $U(t) \in [-20, 20]$, control interval 10^{-3}, and reset amplitude sampled uniformly from $[1, 10]$. The environment has 201 spatial states and 1,000 control decisions per episode. Batched evaluation collapses the 100 Euler steps under each held action into the algebraically equivalent affine transition; against the reference Gym implementation, a complete deterministic audit trajectory has maximum relative state error 7.35×10^{-6}, and its episodic return agrees at the reported precision.

A prespecified seed fixes the shared 201–64–64–1 feedback policy ($P = 17{,}153$) and episode draws: four training episodes and disjoint sets of ten validation and ten test amplitudes. Both optimizers use ℓ and the same validation checkpoint rule. Adam differentiates this scalar objective through the equivalently batched dynamics, whereas OM-EKI constructs its coordinates from $\mathcal{F}_{\mathrm{tr}}$. Validation and test use the same observation construction over ten episodes each; the native episodic return is reported separately and does not enter $\mathcal{F}_{\mathrm{tr}}$. Adam receives 600 gradient steps, and OM-EKI receives 600 nonlinearly verified update attempts. OM-EKI uses $K = 256$, coordinate range $[32, 192]$, map lifetime $H = 10$, perturbation scale $\sigma = 0.5$, parameter scale $c_\theta = 0.08$, and the same nonlinear-verification rule as the primary experiments.

Table 14: Optimizer-only comparison on the standard PDEControlGym reaction–diffusion environment. Time to Adam quality is the first registered validation checkpoint reaching Adam's endpoint cost; full-budget time includes optimization, validation, and OM geometry construction.

Method	Search dimension	Validation cost ↓	Test cost ↓	Test final L_2 ↓	Official return ↑	Time to Adam quality (s) ↓	Full-budget time (s) ↓
Adam	17,153	0.46237	0.68153	0.10031	302.534	1,525.20	**1,525.88**
OM-EKI	32–192	**0.45762**	**0.67352**	**0.05149**	**302.568**	**63.39**	2,080.93

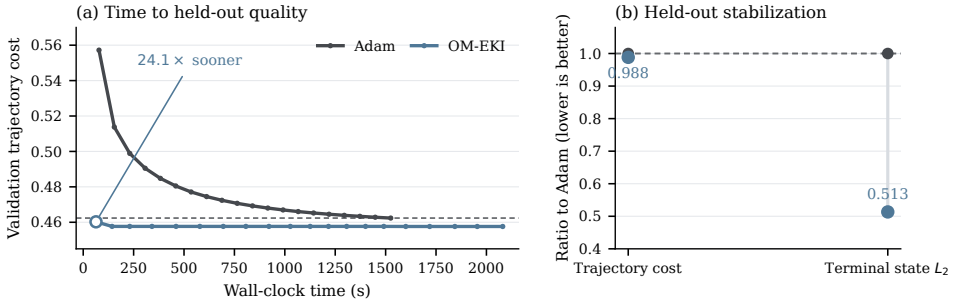

Figure 5: PDEControlGym reaction–diffusion boundary stabilization. (a) Post-initial registered validation costs against end-to-end wall time; the horizontal line is Adam's endpoint quality and the marked OM-EKI checkpoint is the first to cross it. (b) Held-out test metrics normalized by Adam. Raw values are given in Table 14.

In this prespecified run, OM-EKI reaches Adam's endpoint validation quality at its first registered checkpoint (update 30), $24.1\times$ sooner in wall time. Its held-out validation and test costs and reported return are comparable to Adam's selected endpoint, while the held-out terminal-state norm is reduced by 48.7%. At the full 600-update budget, OM-EKI takes longer wall time than Adam because it continues reconstructing and verifying coordinates after its quality has plateaued; the result separates time to a common quality from full-budget runtime.

At the initial policy, a single direction captures 95% of the measured response energy, and the selected coordinates attain residual reachability of 0.999. The initial observable geometry is therefore both compact and corrective, consistent with the rapid early improvement. This extends the evidence for response-derived coordinates to a neural feedback controller in a standard PDE-control benchmark.

E EXPERIMENTAL SETUP AND REPORTING PROTOCOL

This section records the protocols, hardware, task definitions, solvers, and optimizer settings used by the three primary experiments.

E.1 COMPARISON AND REPORTING

Within each primary comparison, methods share the model, data partition, forward solver, initialization policy, and checkpoint rule; explicitly labeled solver variants are reported separately. We call observations excluded from parameter updates *held out*: validation selects checkpoints, whereas test data are evaluated only after selection. The deterministic task reports validation and frozen-test MAE; POLLU and SHD report validation metrics. Hyperparameters and budgets are fixed before evaluation, and primary aggregates are medians over seeds 1–4. Each quality and time comes from the same selected record; Appendix F.1 gives the underlying values. POLLU measures transfer from two training initial conditions to one held-out validation initial condition.

We report held-out quality, end-to-end wall time, and time-to-quality at common evaluation gates. Wall time includes optimization, validation, and geometry construction; a dash means that the threshold was not reached or the quantity is not applicable.

E.2 HARDWARE AND SOFTWARE

All reported wall times were measured on one NVIDIA GeForce RTX 4070 Laptop GPU. The reference environment uses Python 3.10.20, PyTorch 2.5.1 with CUDA 12.4, NumPy 2.2.5, and SciPy 1.15.2. Population methods evaluate conditionally independent forward rollouts in predeclared memory-safe chunks. Appendix D.1 reports the corresponding single-GPU scaling control.

E.3 PRIMARY TASK CONFIGURATIONS

Table 15 collects the task-specific choices needed to interpret the main comparisons.

Table 15: Primary experimental configurations. "Budget" denotes optimizer steps for gradient methods and nonlinearly verified update attempts for OM-EKI.

Task	Model, data, and forward dynamics	Held-out metrics	Primary budget and OM capacity
Deterministic ODE	2–50–2, $P = 252$; 64/32/32 windows from sequential train/validation/test regions with disjoint start indices; fixed-step RK4	validation and frozen-test MAE	12,000; dimension 64; rollout chunk 32
POLLU UDE	20 states, 25 reactions, 22 times; 2/1 train/validation trajectories; overparameterized 20–128–128–3 closure, $P = 19{,}587$; stiff TR-BDF2 solve	validation log-MSE	600; dimension 128; secant chunk 64
Hard-reset SHD	600/600 class-balanced train/validation panels; 64 hidden units and 100 time bins; $P = 50{,}176$; exact Heaviside spike and reset	validation Brier score	1,200 SG steps / 120 OM updates; cap 256, dimension 251–256; chunk 32

E.4 TASK EQUATIONS AND OBSERVABLE OUTPUTS

Deterministic Neural ODE. The data-generating dynamics follow the torchdiffeq demonstration, $\dot{u} = u^{\odot 3}A$, with $A = \left[\begin{smallmatrix} -0.1 & 2 \\ -2 & -0.1 \end{smallmatrix}\right]$ and $u(0) = (2, 0)$. The learned vector field is $f_\theta(u) = W_2 \tanh(W_1 u^{\odot 3} + b_1) + b_2$, with widths 2–50–2. For window starts s_i, the structured output is $\mathcal{F}(\theta) = \mathrm{vec}\{\widehat{u}_\theta(t_{s_i:s_i+9})\}_{i=1}^{64} \in \mathbb{R}^{1280}$; validation and test use disjoint start-index regions.

POLLU UDE. The state satisfies $\dot{x} = S\widetilde{r}_\theta(x)$, where $S \in \mathbb{R}^{20\times 25}$ and $r(x)$ are the standard POLLU stoichiometry and mass-action rates (SciMLBenchmarks.jl Contributors, n.d.). For state component ℓ, let $s_\ell = \max\{10^{-7}, \max_{i,t} x^{\mathrm{ref}}_{i,\ell}(t)\}$ over the two reference training trajectories and define the componentwise input $\xi_\ell(x) = \log(1+\max\{x_\ell, 0\}/s_\ell)$. Reaction channels $q \in \{1, 2, 15\}$ use $\widetilde{r}_{\theta,q}(x) = r_q(x)m_{\theta,q}(x)$, where $m_\theta(x) = \mathrm{softplus}(\mathrm{clip}_{[-8,8]}\, g_\theta(\xi(x)))/(2\log 2) > 0$; the remaining rates retain their standard form. The closure g_θ has widths 20–128–128–3. Setting all weights to zero and the three output biases to $b_\star = \log(\exp(2\log 2) - 1)$ gives $m_\theta(x) \equiv 1$, recovering the standard POLLU system. With two training initial conditions and 21 post-initial observation times, $\mathcal{F}(\theta) = \mathrm{vec}\{\log_{10}\max(\widehat{x}_{\theta,i,\ell}(t_k), 10^{-9})\} \in \mathbb{R}^{840}$.

Hard-reset SHD recurrence. For input drive d_t, the recurrent hidden state emits $s_t = H(v_t - 1)$, updates $q_{t+1} = \alpha q_t + d_t + s_{t-1}W_{\mathrm{rec}}$, and resets $v_{t+1} = (\beta v_t + q_t) \odot (1 - s_t)$, with $s_{-1} = 0$, $\alpha = e^{-1/5}$, and $\beta = e^{-1/10}$ for 1-ms discretization and time constants 5 and 10 ms. A two-stage linear output filter is max-pooled over 100 time bins and passed through softmax. Thus $\mathcal{F}(\theta)$ concatenates 20 probabilities for each of 600 training examples, giving $M = 12{,}000$. OM-EKI evaluates the exact Heaviside/reset recurrence; the SG reference changes only its backward derivative surrogate.

E.5 Observation, acceptance, and solver conventions

OM-EKI receives the structured training outputs defined in Appendix E.4; POLLU DFBGN receives their residual vector, whereas scalar baselines receive only the corresponding loss. Validation observations select checkpoints but do not construct mappings, Kalman updates, or trust decisions. All primary mappings use $\Gamma = I$. The Kalman update separately uses $\Omega_t = \gamma_t I$, with $(\gamma_0, \mathrm{decay}, \mathrm{floor}) = (0.1, 0.995, 10^{-4})$ for the deterministic ODE, $(10^{-3}, 0.95, 10^{-4})$ for POLLU, and $(0.02, 0.995, 10^{-4})$ for SHD. This schedule changes the EKI residual scale and damping without rotating observation directions or changing their relative weighting.

The deterministic model uses one RK4 substep per observation interval, while its targets use four. POLLU training uses fixed-step TR-BDF2 with maximum step 0.5 and two quasi-Newton iterations per stage; reference trajectories use SciPy BDF with relative and absolute tolerances 10^{-10} and 10^{-12}. The SHD recurrence advances over 100 time bins with the exact Heaviside spike and hard reset used by all derivative methods.

E.6 Optimizer settings

Table 16 gives the task-specific quantities left symbolic in Algorithm 1. We write $\gamma_t = \max(\gamma_{\min}, \gamma_0 d^t)$ so that $\Omega_t = \gamma_t I$. The coordinate ensemble contains $r_t + 1$ members; K is the number of secant probes, H the maximum map lifetime in attempted local updates, and m the number of leading mapping directions carried into the next probe span.

Table 16: Primary task-specific OM-EKI settings. The interval reports the allowed realized coordinate-dimension range.

(a) Coordinate capacity and map reuse

Task	T	K	$[r_{\min}, r_{\max}]$	H	m	Secant/ensemble chunk
Deterministic ODE	12,000	64	[64, 64]	100	16	32/32
POLLU UDE	600	128	[128, 128]	10	32	64/64
Hard-reset SHD	120	257	[4, 256]	8	16	32/32

(b) Coordinate scale and EKI regularization

Task	σ	c_θ	λ_0	$(\gamma_0, d, \gamma_{\min})$
Deterministic ODE	0.5	0.05	1	$(0.1, 0.995, 10^{-4})$
POLLU UDE	0.7	0.10	1	$(10^{-3}, 0.95, 10^{-4})$
Hard-reset SHD	0.7	0.35	16	$(0.02, 0.995, 10^{-4})$

All three tasks use trust radius 1, verify every proposal, rebuild after two consecutive rejections, and bound λ to $[10^{-5}, 10^{6}]$. Deterministic ODE and POLLU use the direct ensemble-space solve; SHD uses its equivalent Cholesky form and accumulates observation statistics in chunks of 800 entries. The selected SHD runs realize 251–256 coordinates under the cap of 256 (Table 20). Total objective evaluations across seeds 1–4 are 794,200–811,466, 100,385–106,852, and 46,582–48,652 for deterministic ODE, POLLU, and SHD, respectively, including all mapping and verification rollouts.

Comparator definitions follow Adam and AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019), SPSA (Spall, 1992), and separable CMA-ES (Hansen & Ostermeier, 2001; Ros & Hansen, 2008); Table 17 gives their task-specific settings.

Table 17: Settings for the comparison optimizers in the main and seed-wise tables. "CMA-ES" denotes the separable variant; "Gen." denotes generations and "pop." population size.

Task	Gradient / surrogate-gradient comparators	Derivative-free / forward-based comparators
Deterministic ODE	Adam and discrete AdamW: 12,000 steps, learning rate 10^{-3}; AdamW weight decay 10^{-2}. Continuous-adjoint Adam uses the same step count and learning rate.	Full-space EKI: 12,000 updates, $\sigma = 0.5$, $\lambda = 1$, and the same Ω_t schedule as OM-EKI.
POLLU UDE	Adam and AdamW: 600 steps, learning rate 3×10^{-4}; AdamW weight decay 10^{-2}.	Neural-ODE ES: 775 gen., pop. 128, $\sigma = 0.01$, learning rate 3×10^{-4}; separable CMA-ES: 781 gen., pop. 128, $\sigma = 0.1$; DFBGN: block 128, $\rho_{\mathrm{beg}} = 0.2$, $\rho_{\mathrm{end}} = 10^{-8}$, $\Delta_{\max} = 1.6$.
Hard-reset SHD	SG-Adam and SG-AdamW: 1,200 steps, learning rate 10^{-3}; SG-AdamW weight decay 10^{-2}.	SPSA-Adam: 1,200 steps, learning rate 2.5×10^{-4}, perturbation 0.01; Neural-ODE ES: 645 gen., pop. 64, $\sigma = 0.01$, learning rate 10^{-3}; separable CMA-ES: 654 gen., pop. 64, $\sigma = 0.01$.

POLLU subspace Gauss–Newton comparison. DFBGN provides a vector-residual subspace comparator to the scalar-loss baselines. We use the official implementation of Cartis & Roberts (2023) (revision `82da43b`) with the same POLLU initial parameters, training residual vector, forward simulator, and validation metric; its subspace block is 128. An independent pilot fixed $\rho_{\mathrm{beg}} = 0.2$; a second (seed 1903) selected $\Delta_{\max} = 1.6$ from $\{0.8, 1.6, 3.2\}$, after which seeds 1–4 shared these settings. Each seed received the corresponding OM-EKI forward-evaluation cap (100,385–106,852), but DFBGN reached its native stopping radius $\rho_{\mathrm{end}} = 10^{-8}$ after 221–278 evaluations. Wall time includes training rollouts and validation; validation only selects checkpoints. The shorter DFBGN trajectory thus reflects native solver termination rather than budget exhaustion.

F Individual Runs

This appendix reports the individual runs underlying Section 5.1; the main tables give their aggregate summaries. Values are taken from validation-selected checkpoints, and a dash denotes an unreached threshold or an inapplicable quantity. Figures 6–9 provide the corresponding traces.

F.1 Seed-wise primary results

Tables 18, 19, and 20 report the individual runs for the three primary tasks.

Table 18: Individual deterministic Neural ODE runs underlying Table 1. All methods use 12,000 optimizer steps or update attempts under their respective update rules. Wall time is reported separately; equal update counts do not imply equal forward-evaluation cost.

Method	Seed	Budget	Validation MAE ↓	Test MAE ↓	Time (s) ↓
Adam	1	12,000	9.0967×10^{-5}	8.9675×10^{-5}	383.71
Adam	2	12,000	3.3273×10^{-5}	3.6705×10^{-5}	383.44
Adam	3	12,000	8.8463×10^{-5}	6.4393×10^{-5}	381.33
Adam	4	12,000	6.1159×10^{-5}	5.9343×10^{-5}	386.01
OM-EKI	1	12,000	3.0237×10^{-6}	1.5849×10^{-6}	154.86
OM-EKI	2	12,000	5.9430×10^{-6}	6.6536×10^{-6}	154.55
OM-EKI	3	12,000	8.6539×10^{-6}	1.3004×10^{-5}	153.78
OM-EKI	4	12,000	2.6570×10^{-5}	3.5715×10^{-5}	158.64
AdamW	1	12,000	6.8686×10^{-5}	8.8948×10^{-5}	387.46
AdamW	2	12,000	5.0788×10^{-5}	5.3001×10^{-5}	382.17
AdamW	3	12,000	1.1129×10^{-4}	7.9302×10^{-5}	383.64
AdamW	4	12,000	6.1751×10^{-5}	5.3411×10^{-5}	383.06
Adam + adjoint	1	12,000	2.6914×10^{-5}	7.0466×10^{-5}	610.37
Adam + adjoint	2	12,000	3.2580×10^{-5}	1.1994×10^{-5}	612.25
Adam + adjoint	3	12,000	7.1945×10^{-5}	8.0131×10^{-5}	612.93
Adam + adjoint	4	12,000	3.9417×10^{-5}	6.5282×10^{-5}	611.00
Full-space EKI	1	12,000	2.4214×10^{-5}	1.9648×10^{-5}	202.63
Full-space EKI	2	12,000	2.7478×10^{-5}	2.0661×10^{-5}	204.09
Full-space EKI	3	12,000	3.1549×10^{-5}	5.5310×10^{-5}	203.38
Full-space EKI	4	12,000	1.0283×10^{-4}	1.0142×10^{-4}	200.02

Table 19: Individual POLLU UDE runs underlying Table 2. Time to paired Adam quality is the first evaluation gate reaching that seed's Adam endpoint score.

Method	Seed	Validation log-MSE ↓	Endpoint time (s) ↓	Time to paired Adam quality (s) ↓	Speedup ↑
Adam	1	2.7460×10^{-5}	9,169.90	9,169.90	1.00×
Adam	2	2.7293×10^{-5}	9,212.45	9,212.45	1.00×
Adam	3	2.7304×10^{-5}	9,068.68	9,068.68	1.00×
Adam	4	2.7464×10^{-5}	9,042.65	9,042.65	1.00×
OM-EKI	1	2.4300×10^{-5}	9,057.77	1,710.34	5.36×
OM-EKI	2	2.3867×10^{-5}	9,155.04	1,216.63	7.57×
OM-EKI	3	2.4113×10^{-5}	9,148.55	1,095.96	8.27×
OM-EKI	4	2.3994×10^{-5}	9,081.74	1,148.48	7.87×
AdamW	1	2.7461×10^{-5}	8,918.26	—	—
AdamW	2	2.7293×10^{-5}	9,054.38	—	—
AdamW	3	2.7302×10^{-5}	8,985.54	—	—
AdamW	4	2.7458×10^{-5}	9,109.81	—	—
Neural-ODE ES	1	2.1663×10^{-3}	9,215.60	—	—
Neural-ODE ES	2	2.1033×10^{-3}	9,254.55	—	—
Neural-ODE ES	3	2.1403×10^{-3}	9,278.75	—	—
Neural-ODE ES	4	2.1774×10^{-3}	9,285.17	—	—
CMA-ES	1	3.6139×10^{-5}	12,537.65	—	—
CMA-ES	2	3.5998×10^{-5}	12,375.09	—	—
CMA-ES	3	6.3565×10^{-5}	12,325.11	—	—
CMA-ES	4	3.6063×10^{-5}	12,910.21	—	—
DFBGN	1	1.3856×10^{-3}	1,452.88	—	—
DFBGN	2	2.9355×10^{-4}	1,160.41	—	—
DFBGN	3	4.4325×10^{-5}	1,261.86	—	—
DFBGN	4	8.4978×10^{-5}	1,300.02	—	—

Table 20: Individual hard-reset SHD validation runs underlying Table 3, including the OM-EKI capacity variants.

Method	Seed	Brier ↓	Accuracy ↑	Time (s) ↓
Surrogate backward				
SG-Adam	1	0.040049	31.00%	1,852.98
SG-Adam	2	0.041214	35.33%	1,794.35
SG-Adam	3	0.043953	23.00%	1,757.66
SG-Adam	4	0.041771	30.50%	1,756.74
SG-AdamW	1	0.045105	19.67%	3,443.44
SG-AdamW	2	0.039232	38.50%	3,460.62
SG-AdamW	3	0.041391	31.50%	3,646.57
SG-AdamW	4	0.044240	19.67%	3,378.99
Scalar forward loss				
SPSA-Adam	1	0.047841	6.17%	2,146.14
SPSA-Adam	2	0.047579	4.33%	2,155.87
SPSA-Adam	3	0.048015	4.17%	2,235.57
SPSA-Adam	4	0.047544	6.83%	2,145.85
Neural-ODE ES	1	0.046539	12.17%	1,692.38
Neural-ODE ES	2	0.047086	8.67%	1,838.70
Neural-ODE ES	3	0.046780	9.33%	1,728.57
Neural-ODE ES	4	0.047459	5.67%	1,703.63
CMA-ES	1	0.042942	25.50%	1,168.69
CMA-ES	2	0.044175	23.83%	1,269.10
CMA-ES	3	0.043861	22.33%	1,148.73
CMA-ES	4	0.044209	21.83%	1,141.66
Vector forward observations (OM-EKI)				
OM-EKI (dim. 64)	1	0.045003	20.17%	450.43
OM-EKI (dim. 64)	2	0.045808	19.33%	446.87
OM-EKI (dim. 64)	3	0.045452	19.83%	463.21
OM-EKI (dim. 64)	4	0.045595	19.33%	473.55
OM-EKI (cap 128)	1	0.042822	26.00%	744.13
OM-EKI (cap 128)	2	0.044889	22.83%	770.80
OM-EKI (cap 128)	3	0.042902	25.67%	831.49
OM-EKI (cap 128)	4	0.044237	23.33%	785.34
OM-EKI (cap 256)	1	0.042225	29.00%	1,406.88
OM-EKI (cap 256)	2	0.043615	26.83%	1,382.21
OM-EKI (cap 256)	3	0.041965	33.17%	1,406.26
OM-EKI (cap 256)	4	0.042211	30.67%	1,396.87

The matched rank-cap-128 and rank-cap-256 runs retain 124–128 and 251–256 coordinates, respectively.

G Detailed Empirical Figures

The main paper summarizes the empirical evidence in Figures 3–4. The following plots provide the seed-wise traces and task-specific diagnostics behind those summaries.

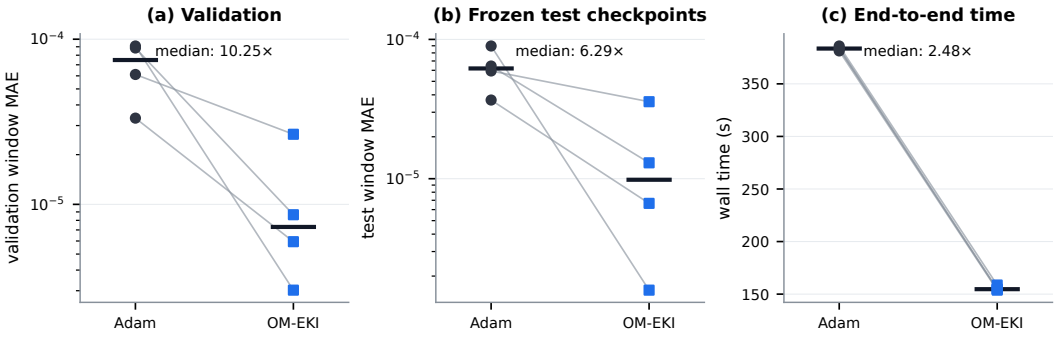

Figure 6: Seed-wise deterministic Neural ODE validation error, frozen test error, and end-to-end time. Lines pair methods within each of four seeds; horizontal bars mark medians.

OM-EKI improves held-out error and wall time in every paired seed.

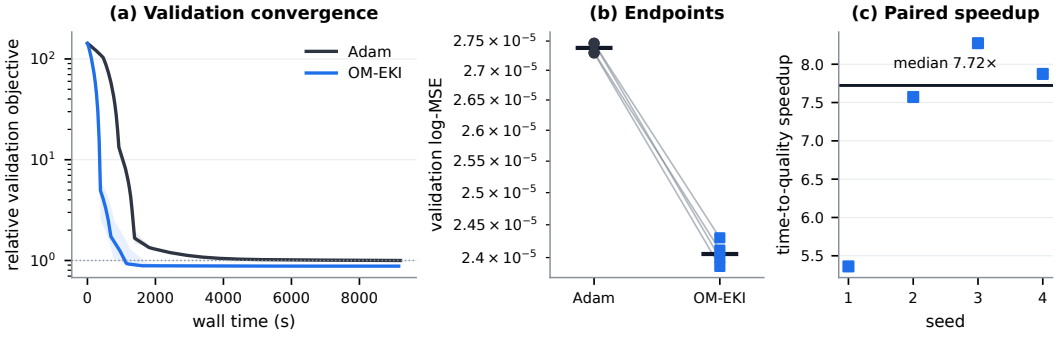

Figure 7: Four-seed POLLU validation convergence, validation-selected endpoints, and paired time-to-Adam-quality speedups. In (a), shading spans the seed-wise trajectories; horizontal bars mark medians.

All four OM-EKI runs reach and finish below paired Adam quality; DFBGN terminates above it in every seed (Table 19).

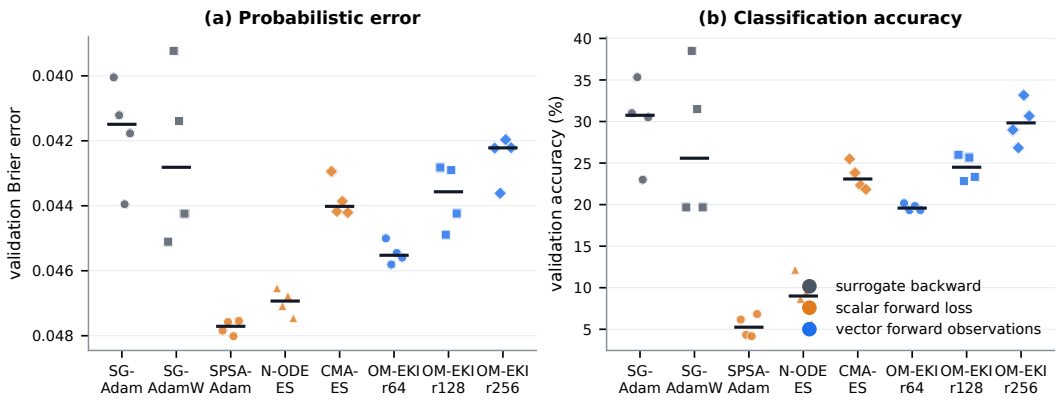

Figure 8: Seed-wise variability for the hard-reset SHD comparison. Points are individual runs and horizontal segments are medians. Colors distinguish surrogate-backward, scalar-forward-loss, and vector-forward-observation methods.

The capacity trend holds for Brier error and accuracy, with the cap-256 range overlapping the SG-Adam range.

Convergence traces. Within each seed, training traces are normalized by each method's initial cumulative-best objective, whereas paired validation traces share Adam's initial validation metric as the reference. The traces are interpolated on a common wall-time grid before computing pointwise medians and ranges.

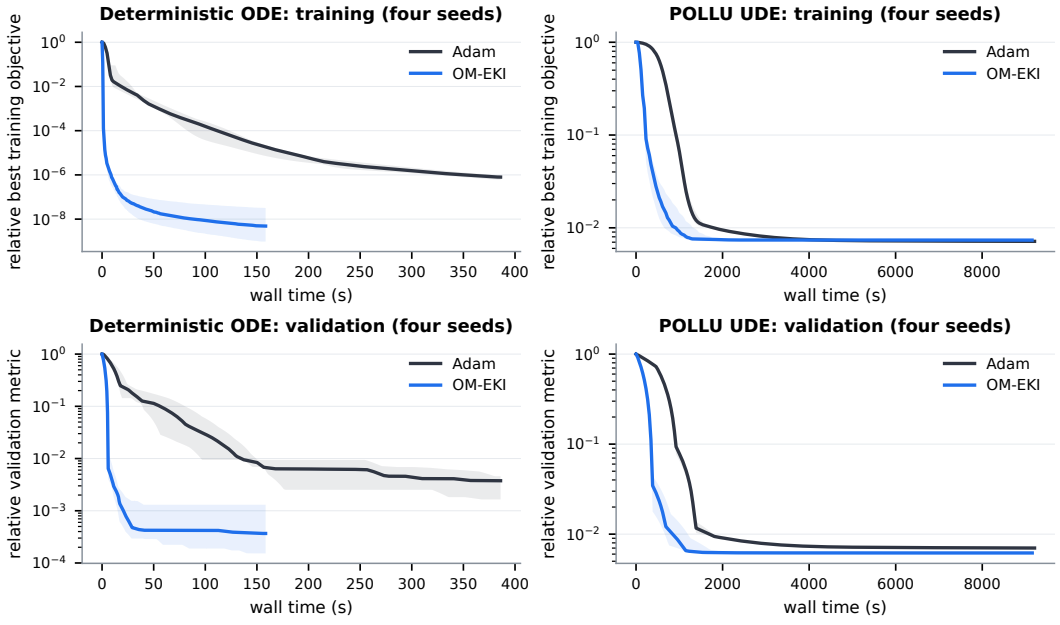

Figure 9: Median cumulative-best training and validation traces for the deterministic ODE and POLLU UDE. Training uses method-specific initial normalization; validation uses the paired Adam initial value. Shading spans the four-seed range.

The traces recover the deterministic held-out ordering and POLLU's earlier time-to-quality crossing under the stated checkpoint rules.

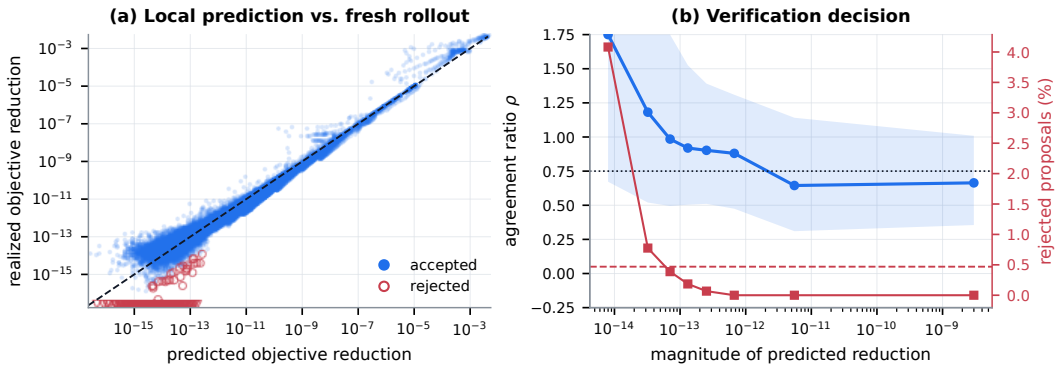

Figure 10: Calibration of nonlinear verification on the deterministic Neural ODE. (a) Predicted versus realized objective reduction; the dashed line is exact agreement, hollow red points are rejected proposals, and downward triangles place nonpositive realized reductions at the plotting floor. (b) Median agreement ratio ρ with 10–90 percentile band across eight equal-count bins of predicted-reduction magnitude; red squares give the rejection rate, the red dashed line is the acceptance threshold, and the black dotted line is the strong-agreement threshold.

Nonlinear-verification calibration. Large predicted reductions lie close to the agreement diagonal, while discrepancies concentrate near the numerical floor. Fresh rollouts therefore calibrate the local model at finite step sizes, complementing the endpoint intervention in Table 6.

Capacity and drift diagnostics. Figures 11 and 12 expand the capacity and coordinate-lifetime conclusions in Section 5.2.

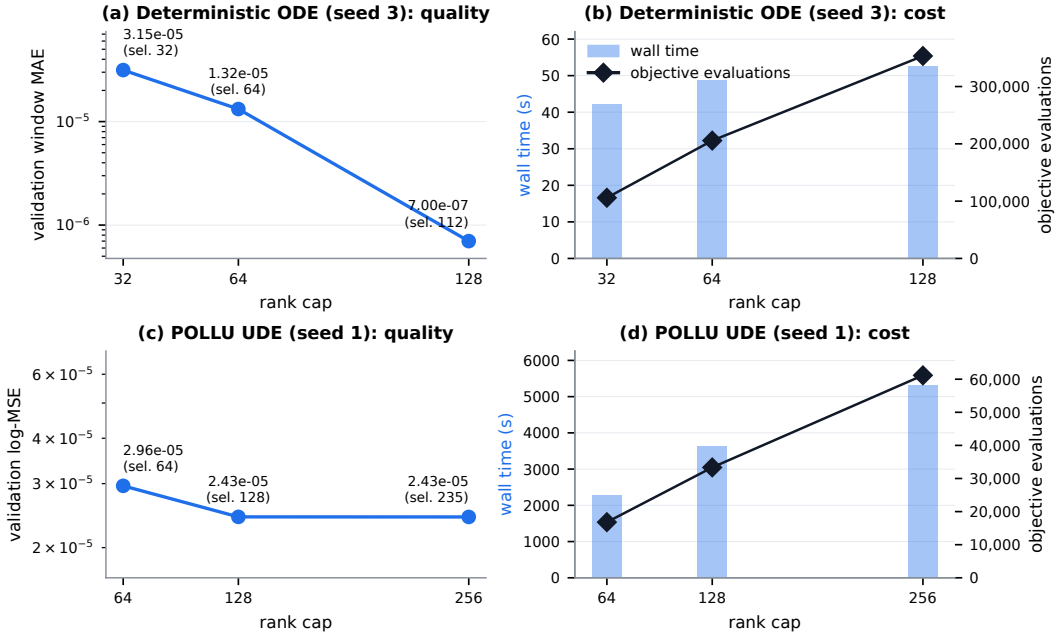

Figure 11: Rank-cap sensitivity on the deterministic ODE and POLLU UDE. Each point reports its realized coordinate dimension; the adjacent panels show the associated wall time and objective evaluations.

Increasing the cap exposes a task-dependent tradeoff. The deterministic run continues to improve through realized dimension 112, whereas POLLU changes little from dimension 128 to 235 despite

additional forward work. This complements Table 5: sufficient capacity is necessary, but useful capacity is determined by the observed task rather than the cap alone.

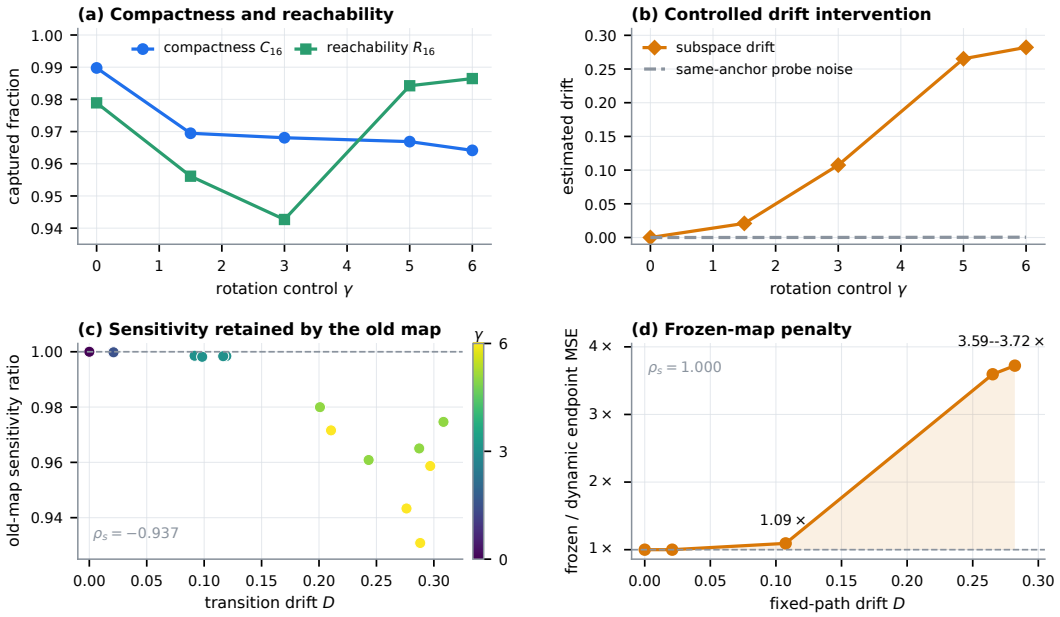

Figure 12: Controlled-drift audit underlying Figure 4(c). Panels (a,b) control compactness and new-map reachability; panels (c,d) link lost old-map sensitivity to the endpoint penalty, identifying drift as a limit on coordinate lifetime.