INTERPRETABLE TEST-TIME ADAPTATION OF LATENT DYNAMICS FROM SPARSE SENSORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-driven models of spatiotemporal dynamics are trained on a limited range of physical parameters and routinely deployed on regimes outside it, where a few point sensors may be the only measurements available. Under such parameter shift, the learned latent dynamics no longer describe the system, and the full-state data needed for retraining the model does not exist. We propose TTA-SINDy-SHRED for sparse identification of nonlinear dynamics with shallow recurrent decoder networks (SINDy-SHRED), enabling efficient test-time adaptation to unseen parameter regimes. The joint training objective shapes the SINDy-SHRED latent representation so that a mechanistic governing equation captures the dynamics across regimes, with all regime variation concentrated in the coefficient matrix of the latent space physics model. At test time, all network weights stay fixed and only two context variables are inferred. Adaptation requires no backpropagation, keeps the latent dynamics interpretable, and updates the decoder from point measurements alone. We evaluate on shock-forming advection–diffusion, reaction–diffusion pattern formation, and decaying two-dimensional turbulence. In every held-out regime, adaptation from a short calibration window reduces long-horizon forecast error relative to the unadapted model. We further test transferability to both neural-operator and attention-based architectures paired with learned latent dynamics, and TTA-SINDy-SHRED attains the lowest error across forecast horizons.

1 INTRODUCTION

Data-driven modeling has become a dominant paradigm for spatiotemporal forecasting across the physical and biological sciences (Karniadakis et al., 2021; Brunton & Kutz, 2024; Li et al., 2020; Lu et al., 2021; McCabe et al., 2025). Their deployment, however, often differs substantially from the conditions under which training data are collected. Full spatial states could be available from simulations or controlled experiments in selected parameter regimes, but are expensive or infeasible to acquire in most practical settings. For instance, ocean monitoring relies on sparse buoy networks, turbulence diagnostics use a handful of pressure taps, and reactor monitoring samples only a small fraction of the spatial domain. Moreover, changes in physical parameters—viscosity, diffusion rates, or reaction constants—can move a system outside the training envelope. These conditions motivate efficient adaptation from data-rich source regimes to unseen target regimes where only sparse sensor measurements are available.

Existing modeling approaches address different aspects of this challenge. Projection-based reduced-order models construct low-dimensional approximations from state snapshots (Benner et al., 2015), while neural operators learn solution maps across families of input functions and physical parameters (Li et al., 2020; Lu et al., 2021; Kovachki et al., 2023). Physics-informed neural networks incorporate governing equations into the training objective (Raissi et al., 2019), and learned reconstruction models recover spatial fields from sparse measurements (Erichson et al., 2020; Santos et al., 2023). For adaptation across dynamical environments, context-conditioned models separate shared structure from environment-specific variation (Yin et al., 2021; Kirchmeyer et al., 2022). However, learning across parameter regimes does not by itself ensure reliable extrapolation beyond them, and reconstructing from sparse observations does not by itself provide a dynamical model for robust forecasting beyond the observation window. The deployment problem considered here requires both capabilities together:

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.

adaptation to unseen parameters from a short, spatially sparse observation window, followed by long-horizon full-state forecasting without network retraining.

Shallow recurrent decoder networks (SHRED) (Williams et al., 2024) reconstruct high-dimensional spatiotemporal fields from time histories of sparse sensor measurements, motivated by delay-coordinate embedding (Takens, 1981). SINDy-SHRED (Gao et al., 2026b) extends this framework by jointly training a recurrent temporal unit, a shallow decoder, and a sparse governing equation in the SINDy class (Brunton et al., 2016), factorizing sparse-sensor prediction into an observation map and an interpretable latent vector field. Under parameter shift, however, the learned governing equation may no longer accurately describe the latent evolution. The factorization itself provides a natural route to adaptation: the dynamics can be corrected while the learned coordinates and the interpretable form of the governing equation are preserved.

We introduce TTA-SINDy-SHRED, a test-time adaptation method for frozen SINDy-SHRED models under parameter shift. All network weights remain fixed and two context variables are inferred from a short calibration window of sensor data. These contexts adjust the latent governing coefficients and the decoder output layer through affine families fitted across source regimes, requiring no backpropagation and no full-state target observations. We evaluate on the viscous Burgers equation, the Gray–Scott reaction–diffusion system, and 2D Navier–Stokes system. These systems span canonical challenges in computational physics—shock formation, pattern-generating reaction kinetics, and turbulent decay—and represent settings where sparse sensing is the operational norm. The ability to adapt to unseen conditions from point measurements alone opens a practical route to deployment beyond the training envelope. Our contributions are the following:

- We propose an efficient test-time adaptation framework that adapts latent dynamical models to unseen parameter regimes for complex dynamics under extremely sparse sensing conditions. The adapted model retains mechanistic interpretability through the discovery of latent dynamics and enables robust long-horizon online forecast.
- We show that the dominant source of longer-horizon forecast error under parameter shift is the latent temporal dynamics rather than spatial reconstruction. The decoder context typically provides additional adaptation that is more useful at shorter horizons.
- We demonstrate that the same adaptation mechanism transfers to alternative architectures (FNO, DeepONet, Senseiver) paired with learned latent dynamics, showing that the mechanism is not architecture-specific, while TTA-SINDy-SHRED, whose latent space is trained jointly with its governing dynamics equation, attains the lowest long-horizon forecast error.

2 Related Works

Sparse sensing methods. Reconstructing high-dimensional spatiotemporal fields from limited measurements is a long-standing problem. Classical approaches recover fields from linear projections under sparsity or low-rank assumptions (Everson & Sirovich, 1995; Candès et al., 2006; Donoho, 2006). Neural network methods have relaxed these assumptions, with shallow decoders mapping sensor readings to full fields (Erichson et al., 2020), Voronoi tessellation handling arbitrary and moving sensor configurations (Fukami et al., 2021), and set-valued cross-attention aggregating sparse inputs (Santos et al., 2023). The Shallow Recurrent Decoder (SHRED) (Williams et al., 2024) introduced a temporal strategy grounded in delay-coordinate embedding (Takens, 1981), trading instantaneous spatial coverage for a time-lagged sensor window, and has since been extended to mobile sensors (Ebers et al., 2024), parametric reduced-order modeling (Tomasetto et al., 2025), uncertainty quantification (Gao et al., 2026a), and multi-fidelity data assimilation (Bao et al., 2026; Zhang et al., 2026). These methods reconstruct and forecast within the training distribution or with access to a forward simulator; none adapts the identified dynamics to an unseen parameter regime from sensor data alone.

Scientific machine learning for parametric dynamics. Learning surrogate models of parametric PDEs has been approached from several directions. Operator learning methods (Li et al., 2020; Lu et al., 2021) learn input-to-output solution maps and have been scaled through multi-physics pretraining (McCabe et al., 2023; Herde et al., 2024), but generalize to new coefficients primarily through fine-tuning on target-regime data. On the structured side, the Sparse Identification of Nonlinear Dynamics (SINDy) (Brunton et al., 2016; Rudy et al., 2017) and its variants (Fasel et al.,

2022; Messenger & Bortz, 2021; Myhr et al., 2026) discover interpretable governing equations from data, and several frameworks combine SINDy with learned latent coordinates (Champion et al., 2019; Conti et al., 2023; Mars Gao & Nathan Kutz, 2024). This line of work has been extended to systems with adjustable control parameters (Nicolaou et al., 2023). Projection-based approaches learn polynomial reduced operators from snapshot data (Peherstorfer & Willcox, 2016), and affine-parametric extensions embed the parameter dependence directly into the regression (McQuarrie et al., 2023). Latent-space dynamics identification methods take a similar approach inside an autoencoder (Fries et al., 2022; Bonneville et al., 2024).

Test-time learning methods. Test-time adaptation methods adapt a pretrained model to distribution shift at inference, e.g., via entropy minimization or self-supervised objectives (Wang et al., 2020; Sun et al., 2020; Wang et al., 2026a;b). For dynamical systems, a growing body of work conditions a learned dynamics model on a low-dimensional context refit to each new environment: gradient-based approaches adapt a context vector or environment-specific component (Yin et al., 2021; Kirchmeyer et al., 2022; Nzoyem et al., 2025), amortized methods infer it directly from trajectories (Wang et al., 2022), and affine context structures admit closed-form adaptation with identifiability guarantees (Blanke & Lelarge, 2024). For PDE learning, context parameters of neural solvers have been adapted by first-order optimization (Kassaï Koupaï et al., 2024), and in-context conditioning on example trajectories enables gradient-free generalization to new coefficients (Yang et al., 2023; Serrano et al., 2025). All of these methods operate on full-state observations, and none yields an interpretable sparse governing equation for the adapted dynamics.

3 Context-Conditioned Test-Time Adaptation

Consider a high-dimensional spatiotemporal system whose dynamics depend on a scalar physical parameter μ. In the target regime, the full state $\mathbf{x}_t \in \mathbb{R}^n$ is not observed. Instead, we record $s \ll n$ sensor measurements $\mathbf{y}_t = S\mathbf{x}_t$, where $S \in \mathbb{R}^{s\times n}$ is a fixed row-selection operator. The task is to reconstruct $\mathbf{x}_t$ and forecast $\mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+H}$ from the sensor history alone, when the test parameter μ^* lies outside the set of values seen in training.

Sensing with shallow recurrent decoder networks. We build on the SINDy-SHRED (Gao et al., 2026b) architecture, which reconstructs the full state from sparse temporal measurements. A GRU encoder maps a window of lagged sensor observations to a low-dimensional latent state $\mathbf{z}_t \in \mathbb{R}^{1\times d_z}$, a shallow decoder maps this latent state back to the full spatial field, and the latent trajectory is simultaneously regularized toward a SINDy model (Brunton et al., 2016):

$$\mathbf{z}_t = f_\theta(\mathbf{y}_{t-L:t}), \qquad \hat{\mathbf{x}}_t = W\phi(\mathbf{z}_t) + \mathbf{b}, \qquad \dot{\mathbf{z}} = \Theta(\mathbf{z})\,\Xi. \tag{1}$$

Here $\phi : \mathbb{R}^{d_z} \to \mathbb{R}^p$ collects the decoder's hidden features up to the penultimate layer, and $(W, \mathbf{b})$ are the weights and bias of its output layer. The polynomial library $\Theta(\mathbf{z}) \in \mathbb{R}^{1\times \ell}$ contains ℓ terms evaluated on the latent state after per-dimension normalization, and $\Xi \in \mathbb{R}^{\ell\times d_z}$ is the corresponding sparse coefficient matrix. After training, latent forecasts are produced by forward Euler integration of the latent dynamics and decoded through the shallow decoder. Our method freezes the encoder and decoder and adapts two scalars from a short calibration window of target-regime sensor data: one indexes the latent dynamics, and the other indexes an affine correction to the decoder output layer learned from source data. No gradient is taken with respect to any network parameter.

3.1 Identifying Latent Space Physics Across Regimes

Given source parameter values $\mathcal{M}_{\text{train}} = \{\mu_1, \ldots, \mu_I\}$, we train one SINDy-SHRED model on trajectories pooled across all regimes, jointly optimizing the network parameters and the latent SINDy coefficients Ξ_{pooled}. The encoder and decoder therefore share one latent coordinate system across regimes, while Ξ_{pooled} is a single average over the source dynamics and serves as the unadapted baseline.

Latent space parametric coefficient family. The source and test regimes are chosen close enough that the system undergoes no bifurcation or pattern-type transition between them, so that the latent dynamics vary smoothly with μ. By Taylor's theorem, a smooth parameter dependence is locally well approximated by a linear model, so we represent the latent SINDy coefficient matrix as an affine

function of a scalar context variable c,

$$\Xi(c) = \Xi_0 + c\,\Xi_1. \tag{2}$$

For each source regime we define the standardized context coordinate $c_i = (\mu_i - \bar{\mu})/\sigma_\mu$, where $\bar{\mu}$ and σ_μ are the sample mean and standard deviation of the source parameter values, respectively. This centers the family so that $c = 0$ corresponds to the source centroid, where the pooled coefficient matrix is structurally close to Ξ_0, and makes the regularization scale-invariant and the context values comparable across physical parameters. Substituting (2) into the latent dynamics gives $\dot{\mathbf{z}} = \Theta(\mathbf{z})\,\Xi_0 + c\,\Theta(\mathbf{z})\,\Xi_1$, where $\Theta(\mathbf{z})$ is computed from the frozen encoder's latent trajectory, $\dot{\mathbf{z}}$ is its first-order forward difference, and c_i is known for each regime. The unknowns Ξ_0 and Ξ_1 therefore enter linearly, and stacking the augmented design rows $[\Theta(\mathbf{z}_t^{(i)}),\ c_i\,\Theta(\mathbf{z}_t^{(i)})]$ over all source time steps produces a single overdetermined linear system, solved by ridge regression:

$$(\Xi_0,\,\Xi_1) = \arg\min_{\Xi_0,\,\Xi_1} \sum_{i,t} \big\|\dot{\mathbf{z}}_t^{(i)} - \Theta(\mathbf{z}_t^{(i)})\,\Xi(c_i)\big\|_2^2 + \lambda_{\Xi,0}\|\Xi_0\|_F^2 + \lambda_{\Xi,1}\|\Xi_1\|_F^2. \tag{3}$$

Fitting Ξ_0 and Ξ_1 jointly avoids the alternative of fitting independent per-regime matrices and extracting the family by principal component analysis, which is contaminated by arbitrary null-space components whenever the per-regime libraries are rank-deficient.

A decoder correction family is constructed analogously, now against the frozen decoder's reconstruction residual on the full source snapshots rather than the latent derivative. We fit an affine output-layer correction $\Delta W(c) = \Delta W_0 + c\,\Delta W_1$, $\Delta\mathbf{b}(c) = \Delta\mathbf{b}_0 + c\,\Delta\mathbf{b}_1$ by the same joint ridge regression,

$$(\Delta W_k, \Delta\mathbf{b}_k) = \arg\min \sum_{i,t} \big\|\mathbf{x}_t^{(i)} - \big(W + \Delta W(c_i)\big)\phi(\mathbf{z}_t^{(i)}) - \mathbf{b} - \Delta\mathbf{b}(c_i)\big\|_2^2 + \lambda_{W,0}\,\Omega_0 + \lambda_{W,1}\,\Omega_1, \tag{4}$$

where $(\Delta W_k, \Delta\mathbf{b}_k)$ denotes the four blocks $\Delta W_0, \Delta W_1, \Delta\mathbf{b}_0, \Delta\mathbf{b}_1$, and $\Omega_k = \|\Delta W_k\|_F^2 + \|\Delta\mathbf{b}_k\|_2^2$ penalizes the centroid ($k = 0$) and the slope ($k = 1$) of the correction separately, with $\lambda_{W,1} > \lambda_{W,0}$ so that the parameter-dependent slope, which must extrapolate at test time, is shrunk more aggressively than the centroid.

3.2 Test-Time Inference of the Context

At test time we observe a calibration window $\{\mathbf{y}_t\}_{t=1}^{C}$ from an unseen target regime μ^*. The encoder, decoder, and both families are frozen, and the unknowns are the dynamics context c_{dyn}, constrained to $|c_{\mathrm{dyn}}| \le c_{\max}$, and the decoder context c_{dec}. We solve for them sequentially, because c_{dyn} enters the sensor predictions nonlinearly through a rollout while c_{dec} enters linearly once that rollout is fixed. The encoder is called once, to form the anchor $\mathbf{z}_1 = f_\theta(\mathbf{y}_{1-L:1})$ from the lags preceding the window; every later latent state is generated by the adapted dynamics, $\mathbf{z}_{t+1}(c) = \mathbf{z}_t(c) + \Delta t\,\Theta(\mathbf{z}_t(c))\,\Xi(c)$ for $t = 1, \ldots, C-1$. Each rolled-out state is decoded at the sensor rows through the $c_{\mathrm{dec}} = 0$ member of the decoder family, and c_{dyn} minimizes the resulting residual,

$$\begin{aligned} \mathbf{r}_t(c) &= \mathbf{y}_t - S\big[(W + \Delta W_0)\,\phi(\mathbf{z}_t(c)) + \mathbf{b} + \Delta\mathbf{b}_0\big], \\ \mathcal{J}(c) &= \sum_{t=2}^{C} \|\mathbf{r}_t(c)\|_2^2 + \lambda_{\mathrm{dyn}}\,c^2, \qquad c_{\mathrm{dyn}}^\star = \arg\min_{|c| \le c_{\max}} \mathcal{J}(c). \end{aligned} \tag{5}$$

The sum starts at $t = 2$ because $\mathbf{z}_1$ is produced by the encoder, so $\mathbf{r}_1$ does not depend on c. Decoding through $(W + \Delta W_0,\ \mathbf{b} + \Delta\mathbf{b}_0)$ rather than the bare output layer introduces no test-time freedom, since ΔW_0 and $\Delta\mathbf{b}_0$ are fixed source-fit quantities, and removes the frozen decoder's parameter-independent reconstruction bias from $\mathcal{J}$, so that solving for c_{dyn} at $c_{\mathrm{dec}} = 0$ attributes the residual to the dynamics first (Appendix A.3). Both (5) and (7) are therefore restrictions of a single objective in $(c_{\mathrm{dyn}}, c_{\mathrm{dec}})$, up to terms constant in the free variable, so the two solves descend one criterion.

Because $\mathcal{J}$ is reached through $C - 1$ compositions of the nonlinear rollout, its minimizer admits no closed form, and we compute it by Levenberg–Marquardt. At iterate c_k we re-roll and re-decode at $c_k \pm \epsilon$ with $\epsilon = 10^{-3}$, form the central-difference Jacobian $\mathbf{j}_t = [\mathbf{r}_t(c_k + \epsilon) - \mathbf{r}_t(c_k - \epsilon)]/(2\epsilon)$, and take the damped Gauss–Newton step

$$\Delta_k = -\frac{\sum_t \mathbf{j}_t^\top \mathbf{r}_t(c_k) + \lambda_{\mathrm{dyn}}\,c_k}{\sum_t \|\mathbf{j}_t\|_2^2 + \lambda_{\mathrm{dyn}} + \tau_k}, \tag{6}$$

with $c_k + \Delta_k$ clipped to $[-c_{\max}, c_{\max}]$. The step is accepted only if it decreases $\mathcal{J}$, after which the damping τ_k is reduced; otherwise τ_k is increased and a shorter step is retried from c_k, so the loss is non-increasing by construction. The iteration is initialized from the minimizer of $\mathcal{J}$ on a coarse grid over $|c| \le c_{\max}$ rather than from $c = 0$, which prevents it from stalling on the plateau that $\mathcal{J}$ exhibits when μ^* lies outside the source hull.

Decoder context in closed form. Once $c^\star_{\mathrm{dyn}}$ is fixed, the rollout $\hat{\mathbf{z}}_t = \mathbf{z}_t(c^\star_{\mathrm{dyn}})$ no longer depends on c, and because the source-fit correction enters the sensor prediction affinely in c, the decoder context is a scalar ridge regression,

$$c^\star_{\mathrm{dec}} = \arg\min_{c} \sum_{t=1}^{C} \left\| \mathbf{y}_t - S\big[\big(W + \Delta W(c)\big)\,\phi(\hat{\mathbf{z}}_t) + \mathbf{b} + \Delta\mathbf{b}(c)\big] \right\|_2^2 + \lambda_{\mathrm{dec}}\, c^2, \tag{7}$$

whose minimizer is available in closed form in Equation (14) of Appendix D. The sum runs from $t = 1$ because c_{dec} does affect the decoded anchor. The base point of this regression, $(W + \Delta W_0,\ \mathbf{b} + \Delta\mathbf{b}_0)$, is exactly the decoder through which $c^\star_{\mathrm{dyn}}$ was fit, so (7) solves for the increment that the $c_{\mathrm{dec}} = 0$ member left unexplained.

Deployment. Both contexts are now fixed and nothing is refit beyond the calibration window. Forecasting integrates $\Xi(c^\star_{\mathrm{dyn}})$ forward from the encoder latent at the end of the window over the horizon H and decodes each state through the corrected output layer, $\hat{\mathbf{x}}_t = \big(W + \Delta W(c^\star_{\mathrm{dec}})\big)\phi(\hat{\mathbf{z}}_t) + \mathbf{b} + \Delta\mathbf{b}(c^\star_{\mathrm{dec}})$.

4 Experiments

TTA-SINDy-SHRED provides an effective mechanism for adapting a pretrained reduced-order model to parameter regimes outside the training set, as we demonstrate on three systems in which the parameter enters the governing equations through the measured field, through an unobserved field, or through a reciprocal dependence for which the affine family is only a first-order approximation. We compare four arms, namely the pooled model (No-adapt), decoder correction alone (Decoder only), dynamics adaptation alone (Dynamics only), and joint adaptation (Full), which separates error originating in the latent rollout from error introduced at reconstruction. Table 1 reports relative ℓ^2 errors in latent and physical space at two held-out regimes per system, and Appendices A and B give the experimental settings, the remaining regimes, per-regime horizon curves, and a sweep over the calibration window.

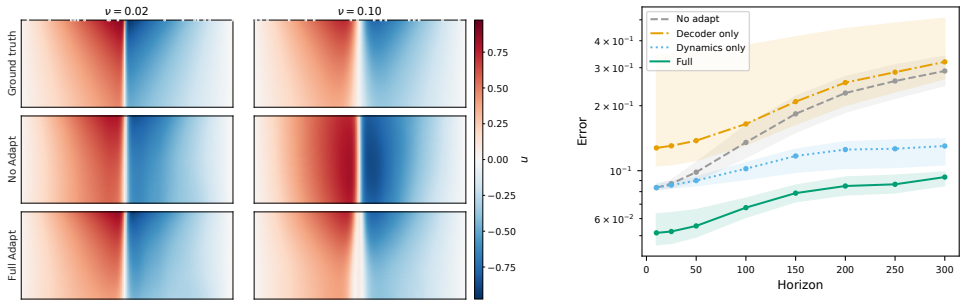
4.1 One-Dimensional Viscous Burgers Equation

The first system we consider is the one-dimensional viscous Burgers equation,

$$\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \nu\,\frac{\partial^2 u}{\partial x^2}, \tag{8}$$

governing a velocity field $u(x,t)$ on the periodic domain $x \in [0, 2\pi)$. Introduced by Bateman (Bateman, 1915) and studied extensively by Burgers (Burgers, 1948) as a reduced model of turbulence, Equation (8) is the canonical setting in which nonlinear advective steepening competes against viscous diffusion. As ν decreases the shock sharpens and energy concentrates at finer spatial scales, while larger ν smooths the solution across the domain. Burgers is the structurally cleanest test case for the affine family of Section 3.1. Writing $\nu = \nu^{(\mathrm{src})} + \delta$, where δ is the deviation from a source reference viscosity, adds the term $\delta\,\partial^2 u/\partial x^2$ to Equation (8), a scaled Laplacian acting on the observed field that is exactly affine in δ, so the coefficient family $\Xi(c) = \Xi_0 + c\,\Xi_1$ follows directly from how viscosity enters the governing operator. Below the training range the shock is sharper than in any training trajectory, and above it the solution becomes progressively more diffuse, so in both directions Ξ_{pooled} encodes a mismatched shock structure that accumulates as systematic rollout error over long horizons.

The Burgers rows of Table 1 separate the contributions of the two contexts. Decoder-only adaptation changes the reconstruction map but not the rollout, so its correction is fitted against a drifting latent



(a) Ground truth, No-adapt, and Full adaptation at the extrapolation regimes.

(b) Decoded error against forecast horizon.

Figure 1: **Out-of-distribution prediction on the Burgers system.** (a) Space-time evolution of u at $\nu = 0.02$ and $\nu = 0.10$ over the forecast horizon, with time increasing downward and sensor locations marked along the top of the ground-truth panels. (b) Decoded error against horizon, aggregated as the mean of per-regime medians.

Table 1: Latent and decoded relative rollout errors of the four arms at two held-out regimes per system, one below and one above the source range, with $H = 300$ for Burgers and Gray–Scott and $H = 550$ for Navier–Stokes. Values are medians over seeds for Burgers and Gray–Scott and means for Navier–Stokes. Bold and underline mark the lowest and second-lowest decoded error in each row, ranked before rounding.

		No-adapt		Decoder only		Dynamics only		Full	
System	Regime	Latent	Decoded	Latent	Decoded	Latent	Decoded	Latent	Decoded
Burgers	$\nu = 0.020$	0.566	0.295	0.566	0.344	0.232	<u>0.206</u>	0.232	**0.102**
	$\nu = 0.100$	0.556	0.297	0.556	0.296	0.160	<u>0.120</u>	0.160	**0.090**
Gray–Scott	$D_v = 0.064$	0.847	0.237	0.847	0.239	0.582	<u>0.132</u>	0.582	**0.126**
	$D_v = 0.070$	0.927	0.223	0.927	0.222	0.985	<u>0.147</u>	0.985	**0.147**
Gray–Scott	$D_u = 0.154$	0.514	0.119	0.514	0.119	0.135	<u>0.048</u>	0.135	**0.044**
	$D_u = 0.166$	0.947	0.245	0.947	0.246	0.371	**0.065**	0.371	<u>0.065</u>
Navier–Stokes	$\mathrm{Re} = 800$	0.697	0.239	0.697	0.228	0.245	**0.072**	0.245	<u>0.079</u>
	$\mathrm{Re} = 2000$	1.266	0.163	1.266	0.164	0.979	**0.030**	0.979	<u>0.035</u>

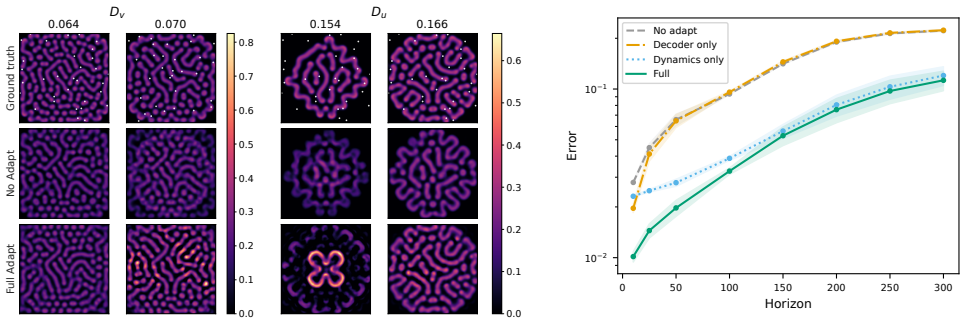
trajectory and makes the decoded error worse in most regimes (Appendix B.1). Inferring c_{dyn} removes the source of the drift, so the latent error falls sharply and the decoded error follows. Full adaptation then fits the decoder correction on the adapted trajectory, where it no longer has to absorb dynamics error, and gives the best reconstruction in every regime. The two contexts therefore act in sequence, with c_{dyn} first correcting the latent evolution and c_{dec} becoming useful only once that evolution is corrected.

4.2 Two-Dimensional Reaction-Diffusion System

The second system we consider is the two-dimensional Gray–Scott reaction-diffusion model, originally introduced as a simplified model of autocatalytic reactions in chemical reactors (Gray & Scott, 1984). It has since become a standard model for studying nonlinear pattern formation. The model describes the interaction of two chemical species, a substrate U and an autocatalyst V, undergoing local reaction and diffusive transport on a doubly-periodic square domain. The governing equations take the form

$$\frac{\partial U}{\partial t} = D_u \nabla^2 U - UV^2 + F(1-U), \qquad \frac{\partial V}{\partial t} = D_v \nabla^2 V + UV^2 - (F+k)V. \tag{9}$$

Depending on (F, k), Equation (9) produces spots, stripes, labyrinths, and oscillatory states (Pearson, 1993), which has made it a common benchmark for numerical methods, data assimilation, and reduced-order modeling (Bao & Kutz, 2025). Both diffusion coefficients multiply a Laplacian, so each admits an affine family as in Section 4.1. We sweep them separately, holding (F, k) inside one



(a) Reconstructed V field at the extrapolation regimes.

(b) Decoded error against forecast horizon, D_v sweep.

Figure 2: **Out-of-distribution prediction on the Gray–Scott system.** (a) The V field at $H = 300$ at the extrapolation regimes of both sweeps, with sensor locations marked on the ground-truth panels. (b) Decoded error against horizon for the D_v sweep, aggregated as the mean of per-regime medians.

Pearson phase region so that no sweep crosses a change in pattern type, and in both sweeps only V is measured. A shift in D_v acts directly on the measured field and changes the characteristic length scale of the pattern, whereas a shift in D_u acts on the unobserved substrate and reaches the sensors only through the coupling UV^2. The two sweeps thus test direct and indirect parameter dependence under the same observation model.

The Gray–Scott rows of Table 1 show that the dynamics context again carries most of the gain, including in the D_u sweep, where the parameter never acts on the measured field. Latent and decoded errors need not move together. At $D_v = 0.070$, adaptation leaves the latent error slightly higher than the unadapted model, yet the adapted model reconstructs the full field better, because the encoder's latent state is not the only trajectory that decodes well. Unlike Burgers, decoder-only adaptation is close to neutral at long horizons rather than harmful. Its value appears early in the rollout, before the latent trajectory has drifted (Figure 2b). As the horizon grows, c_{dyn} dominates, and full adaptation keeps the early reconstruction benefit while matching or improving on dynamics-only adaptation in every regime.

4.3 Two-Dimensional Incompressible Flow

The third system we consider is the two-dimensional incompressible Navier–Stokes equations in vorticity form on a doubly periodic square domain,

$$\frac{\partial \omega}{\partial t} + \mathbf{u} \cdot \nabla \omega = \nu \, \Delta \omega, \qquad \mathbf{u} = (\partial_y \psi, \, -\partial_x \psi) \, , \qquad -\Delta \psi = \omega, \tag{10}$$

where $\omega(x, y, t)$ is the vorticity, ψ the streamfunction, and $\nu = \nu_\star / \mathrm{Re}$ for a fixed constant $\nu_\star$. The system is a standard benchmark for the generalization of learned surrogates to unseen coefficients (Li et al., 2020; Serrano et al., 2026) and an idealization of geophysical flow (Boffetta & Ecke, 2012), in which enstrophy cascades to small scales and is dissipated (Kraichnan, 1967) while coherent vortices slowly decay and merge (McWilliams, 1984). Every regime starts from the same random vorticity field, so trajectories differ only through ν, and the model observes ω at 32 fixed sensor locations, fewer than 0.2% of the grid. The Reynolds number enters only through the viscous term, and writing $\mathrm{Re} = \mathrm{Re}^{(\mathrm{src})} + \delta$ gives $\nu = \nu_\star / (\mathrm{Re}^{(\mathrm{src})} + \delta)$, which is not affine in δ. Its first-order change is still a scaled Laplacian, so $\Xi(c) = \Xi_0 + c\,\Xi_1$ is a local approximation whose accuracy degrades away from the source range. We train on $\mathrm{Re} \in \{1000, 1100, \ldots, 1800\}$ and test on both sides of this range over a horizon long enough that the pooled rollout need not stay bounded.

Above the source range the pooled rollout diverges for several seeds, and decoder-only adaptation cannot prevent this because it decodes the same latent trajectory. Adapting the dynamics keeps more of these rollouts bounded and gives the lowest decoded error in every regime, with full adaptation slightly behind, as a decoder correction fitted during calibration describes the early state of a decaying flow better than its late state. Because every regime starts from the same vorticity field, Re acts mainly on how fast the vortices lose amplitude, and the pooled model fails exactly there, forecasting

vortex cores that are too intense, which full adaptation corrects (Appendix B.4). Appendix E extends the evaluation to Reynolds numbers between 300 and 2500.

4.4 Sensitivity Analysis and Baseline Comparison

Sensitivity. We conduct a sensitivity study on the two-dimensional Navier–Stokes system, varying the latent dimension d_z, calibration window C, sensor noise σ, lag window L, and number of sensors s from the default configuration of Section 4. Performance is robust in all five hyperparameters: errors converge as d_z increases; results remain largely stable for $L \geq 20$; smaller calibration windows ($C = 25$, $C = 50$) yield higher and more variable error at longer horizons; and sensor noise up to $\sigma = 0.2$ degrades accuracy only marginally. The four-arm ablation further demonstrates that, as the forecast horizon grows, the dynamics-only arm surpasses decoder-only; full adaptation, which combines both contexts, attains a balance at all target forecast horizons by applying the corrected latent dynamics before decoder adjustment. This crossover—decoder-only effective at short horizons, dynamics-only increasingly important at longer horizons—is consistent with the complementary roles of c_{dyn} and c_{dec}: the dynamics context corrects the temporal latent evolution over the forecast window, while the decoder context refines the spatial reconstruction at each time step. Figure 13 summarizes the results; more details are provided in Appendix C.

Baseline Comparison. To test whether the observed test-time adaptation gains generalize beyond the proposed architectural choices, we construct four baseline methods that replace both the temporal unit and the latent dynamics model while retaining the decoder structure and the two-scalar adaptation mechanism. The four baselines consist of a Fourier Neural Operator (FNO) operating on full-field snapshots (Li et al., 2020), a masked variant of the FNO that receives only sparse sensor observations (FNO-masked), a DeepONet-style branch network (Lu et al., 2021), and the Senseiver (Santos et al., 2023). Table 12 summarizes the architectural components of each method. Figure 3 plots the full-adaptation error against forecast horizon for all five methods, lines report the mean of per-regime medians over five seeds and shaded bands show the interquartile range; Table 2 reports the numerical values. While FNO and DeepONet achieve the lowest single-step error (H=1), TTA-SINDy-SHRED attains the lowest error at every subsequent horizon, with the baseline errors growing several times faster beyond $H = 50$. By $H = 550$ the error of TTA-SINDy-SHRED remains below 0.04, while the best-performing baseline (DeepONet) exceeds 0.10. These results indicate that the jointly trained SINDy dynamics provide a stronger and more stable prior for long-range latent rollouts than the ConditionalMLP used by the baselines. Detailed baseline architectures, the shared decoder and dynamics formulations, and per-regime ablation curves are provided in Appendix D.

Multi-parameter regime shifts. The present formulation parameterizes dynamical regime variation through a single context variable c, corresponding to a one-dimensional family $\Xi(c) = \Xi_0 + c\,\Xi_1$. While this suffices when the dominant source of distribution shift is a single physical parameter, real-world systems often vary along several axes simultaneously—for instance, Reynolds and Prandtl numbers may shift together in a convective flow. The affine family extends naturally to k context dimensions:

$$\Xi(\mathbf{c}) = \Xi_0 + c_1\,\Xi_1 + \cdots + c_k\,\Xi_k, \tag{11}$$

where each Ξ_j is fitted jointly by the same augmented ridge regression (Equation 3), and the decoder family generalizes analogously. Stage 1 inference becomes a k-dimensional optimization, initialized from a modest set of candidate contexts (e.g., a Sobol sequence) and refined by multi-start Levenberg–Marquardt, avoiding the G^k scaling of a Cartesian grid; Stage 2 remains a closed-form multivariate ridge solve. Because the algebraic structure is unchanged—only the dimension of the family grows—the extension requires no architectural modification, only additional source regimes spanning the multi-parameter space and a proportional increase in calibration cost. Validating this extension on coupled parameter shifts is a natural next step.

5 Conclusion

We have introduced TTA-SINDy-SHRED, a method that adapts a pretrained reduced-order model to unseen parameter regimes by inferring two context variables from a short calibration window of sparse sensor measurements. The dynamics context acts on the latent trajectory and shifts the SINDy coefficient matrix along a learned affine family, and the decoder context applies a corresponding

Table 2: Decoded relative L^2 error (median over 5 seeds, averaged across four test regimes) under full adaptation. Medians are reported for robustness to outlier seed divergence. **Bold** and underline denote best and second-best per column.

	Forecast horizon H						
Method	1	25	50	100	200	400	550
TTA-SINDy-SHRED	0.011	**0.011**	**0.013**	**0.015**	**0.020**	**0.030**	**0.037**
FNO	**0.003**	0.012	0.021	0.039	0.063	0.106	0.133
FNO-masked	0.111	0.104	0.097	0.088	0.090	0.147	0.208
DeepONet	0.004	0.013	0.018	0.029	0.044	0.061	0.103
Senseiver	0.007	0.018	0.024	0.037	0.050	0.084	0.100

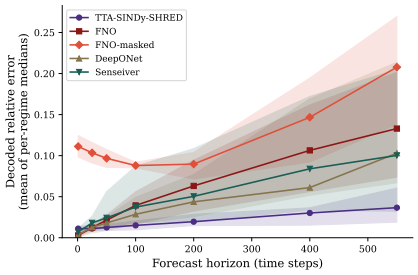


Figure 3: Decoded relative ℓ^2 error vs. forecast horizon for full-adaptation.

affine correction to the output layer; the entire adaptation reduces to a Levenberg–Marquardt solve over a single scalar followed by a closed-form ridge regression.

Experiments on viscous Burgers, Gray–Scott reaction-diffusion, and Navier–Stokes flow demonstrated that full adaptation substantially reduces the relative ℓ^2 error compared to the unadapted pooled model across structurally distinct forms of parameter dependence. The four-arm ablation consistently showed that the dynamics context accounts for most of the long-horizon improvement, while the decoder context provides a complementary reconstruction correction most pronounced at shorter forecast horizons. Baseline experiments confirmed that the context variable mechanism transfers across architectures, improving out-of-distribution predictions when applied to FNO, DeepONet, and Senseiver paired with post-hoc dynamics models. TTA-SINDy-SHRED nevertheless attained the lowest error at every forecast horizon, indicating that the jointly trained SINDy dynamics provide a stronger prior for long-range latent rollout than post-hoc models fitted to unconstrained latent spaces. The joint training objective shapes the latent geometry so that the dynamics are well approximated by a sparse polynomial library and vary smoothly with the context coordinate, a structural alignment that the post-hoc approach cannot enforce and that enables reliable extrapolation without retraining. Sensitivity analysis further showed that adaptation remains robust and effective across a range of hyperparameters, including the number of sensors, lag length, latent dimension, calibration window length, noise level, etc.

Several limitations of the current formulation point to natural future extensions. The current formulation infers a fixed context from a single calibration window, which assumes the target regime remains stationary over the forecast horizon; in systems where the governing conditions evolve continuously during deployment, an online adaptation strategy that updates the context over successive observation windows would be required to maintain predictive accuracy. The present formulation also parameterizes regime variation through a single context variable; extending the family to multiple context dimensions would enable adaptation to coupled parameter shifts, such as simultaneous variation of Reynolds and Prandtl numbers in convective flows.

Beyond these methodological extensions, the framework opens several practical directions. Training once on a coarse set of source regimes and adapting at test time to new parameter values could serve as a lightweight surrogate for parametric PDE solvers, replacing repeated full simulations in design optimization and uncertainty quantification loops with a two-scalar calibration from sparse measurements. Storing one trained model and two context variables per regime, rather than full-field snapshot databases, offers a compressed representation of parametric solution manifolds whose storage savings compound as both the spatial and parameter dimensions grow. In monitoring and digital-twin applications where the operating condition evolves over time through material degradation, changing boundary conditions, or varying loadings, an online variant that updates the context as new sensor data arrives would enable continuous recalibration without retraining, provided the drift remains within the regime captured by the learned family. Finally, the context-conditioned adaptation extends naturally to multi-physics systems where parameter shifts act on fields that are not directly instrumented but remain dynamically coupled to those that are. Potential settings include fluid-structure interaction, where structural sensors must inform adaptation of the surrounding flow, and conjugate heat transfer, where surface temperature measurements constrain the adaptation of internal convective dynamics.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing methods, by writing and editing the software code for model training, figure generation, and our model training procedure. We have not used generative AI tools to formulate the method or its mathematical foundations, compose mathematical claims, propose hypotheses, design experiments, or interpret results, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to draft parts of the paper, including formatting the authors' mathematical expressions, tables, and figures to the required conventions, to identify, search for, and summarize related literature and gaps in prior work, to suggest and format references, to edit the manuscript for readability, and to review drafts of the manuscript for clarity, internal consistency, and agreement between the text and our results. The full author team reviewed all AI-assisted work, verified and tested all code, computed all reported numbers from our own experiment outputs, and checked every cited work against its original source. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yuxuan Bao and J. Nathan Kutz. Data assimilation and discrepancy modeling with shallow recurrent decoders. *arXiv preprint arXiv:2512.01170*, 2025.

Yuxuan Bao, Jan Zajac, Megan Powers, Venkat Raman, and J. Nathan Kutz. Cheap2Rich: A multi-fidelity framework for data assimilation and system identification of multiscale physics. *arXiv preprint arXiv:2601.20295*, 2026.

Harry Bateman. Some recent researches on the motion of fluids. *Monthly Weather Review*, 43(4): 163–170, 1915.

Peter Benner, Serkan Gugercin, and Karen Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review*, 57(4):483–531, 2015.

Matthieu Blanke and Marc Lelarge. Interpretable meta-learning of physical systems. In *International Conference on Learning Representations*, 2024.

Guido Boffetta and Robert E. Ecke. Two-dimensional turbulence. *Annual Review of Fluid Mechanics*, 44:427–451, 2012.

Christophe Bonneville, Youngsoo Choi, Debojyoti Ghosh, and Jonathan L. Belof. GPLaSDI: Gaussian process-based interpretable latent space dynamics identification through deep autoencoder. *Computer Methods in Applied Mechanics and Engineering*, 2024.

Steven L Brunton and J Nathan Kutz. Promising directions of machine learning for partial differential equations. *Nature Computational Science*, 4(7):483–494, 2024.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Johannes Martinus Burgers. A mathematical model illustrating the theory of turbulence. *Advances in Applied Mechanics*, 1:171–199, 1948. doi: 10.1016/S0065-2156(08)70100-5.

Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 2006.

Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116 (45):22445–22451, 2019. doi: 10.1073/pnas.1906995116.

Paolo Conti, Giorgio Gobat, Stefania Fresca, Andrea Manzoni, and Attilio Frangi. Reduced order modeling of parametrized systems through autoencoders and SINDy approach: continuation of periodic solutions. *Computer Methods in Applied Mechanics and Engineering*, 2023.

David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 2006.

Megan R. Ebers, Jan P. Williams, Katherine M. Steele, and J. Nathan Kutz. Leveraging arbitrary mobile sensor trajectories with shallow recurrent decoder networks for full-state reconstruction. *IEEE Access*, 2024.

N. Benjamin Erichson, Lionel Mathelin, Zhewei Yao, Steven L. Brunton, Michael W. Mahoney, and J. Nathan Kutz. Shallow neural networks for fluid flow reconstruction with limited sensors. *Proceedings of the Royal Society A*, 2020.

Richard Everson and Lawrence Sirovich. Karhunen–Loève procedure for gappy data. *Journal of the Optical Society of America A*, 1995.

Urban Fasel, J. Nathan Kutz, Bingni W. Brunton, and Steven L. Brunton. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A*, 2022.

William D. Fries, Xiaolong He, and Youngsoo Choi. LaSDI: Parametric latent space dynamics identification. *Computer Methods in Applied Mechanics and Engineering*, 2022.

Kai Fukami, Romit Maulik, Nesar Ramachandra, Koji Fukagata, and Kunihiko Taira. Global field reconstruction from sparse sensors with Voronoi tessellation-assisted deep learning. *Nature Machine Intelligence*, 2021.

Mars Liyao Gao, Yuxuan Bao, Amy S. Rude, Xinwei Shen, and J. Nathan Kutz. UQ-SHRED: Uncertainty quantification of shallow recurrent decoder networks for sparse sensing via engression. *arXiv preprint arXiv:2604.01305*, 2026a.

Mars Liyao Gao, Jan P. Williams, and J. Nathan Kutz. Sparse identification of nonlinear dynamics and Koopman operators with shallow recurrent decoder networks. *Proceedings of the National Academy of Sciences*, 123, 2026b. doi: 10.1073/pnas.2508144123.

Peter Gray and Stephen K. Scott. Autocatalytic reactions in the isothermal, continuous stirred tank reactor: Oscillations and instabilities in the system $a + 2b \rightarrow 3b$; $b \rightarrow c$. *Chemical Engineering Science*, 39(6):1087–1097, 1984.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel De Bezenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. *Advances in Neural Information Processing Systems*, 37:72525–72624, 2024.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Aly-Khan Kassam and Lloyd N. Trefethen. Fourth-order time-stepping for stiff PDEs. *SIAM Journal on Scientific Computing*, 26(4):1214–1233, 2005.

Armand Kassaï Koupaï, Jorge Mifsut Benet, Yuan Yin, Jean-Noël Vittaut, and Patrick Gallinari. Boosting generalization in parametric PDE neural solvers through adaptive conditioning. In *Advances in Neural Information Processing Systems*, 2024.

Matthieu Kirchmeyer, Yuan Yin, Jérémie Donà, Nicolas Baskiotis, Alain Rakotomamonjy, and Patrick Gallinari. Generalizing to new physical systems via context-informed dynamics model. In *International Conference on Machine Learning*, 2022.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

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

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

L Mars Gao and J Nathan Kutz. Bayesian autoencoders for data-driven discovery of coordinates, governing equations and fundamental constants. *Proceedings of the Royal Society A*, 480(2286): 20230506, 2024.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

Michael McCabe, Payel Mukhopadhyay, Tanya Marwah, Bruno Regaldo-Saint Blancard, Francois Rozet, Cristiana Diaconu, Lucas Meyer, Kaze WK Wong, Hadi Sotoudeh, Alberto Bietti, et al. Walrus: A cross-domain foundation model for continuum dynamics. *arXiv preprint arXiv:2511.15684*, 2025.

Shane A. McQuarrie, Parisa Khodabakhshi, and Karen E. Willcox. Nonintrusive reduced-order models for parametric partial differential equations via data-driven operator inference. *SIAM Journal on Scientific Computing*, 2023.

James C. McWilliams. The emergence of isolated coherent vortices in turbulent flow. *Journal of Fluid Mechanics*, 146:21–43, 1984.

Daniel A. Messenger and David M. Bortz. Weak SINDy for partial differential equations. *Journal of Computational Physics*, 2021.

Håkon Noren Myhr, Sølve Eidnes, and J Nathan Kutz. SIPHy: Sparse identification of port-Hamiltonian systems from noisy data. *arXiv preprint arXiv:2609.20399*, 2026.

Zachary G. Nicolaou, Guanyu Huo, Yihui Chen, Steven L. Brunton, and J. Nathan Kutz. Data-driven discovery and extrapolation of parameterized pattern-forming dynamics. *Physical Review Research*, 2023.

Roussel Desmond Nzoyem, David A.W. Barton, and Tom Deakin. Neural context flows for meta-learning of dynamical systems. In *International Conference on Learning Representations*, 2025.

John E. Pearson. Complex patterns in a simple system. *Science*, 261(5118):189–192, 1993. doi: 10.1126/science.261.5118.189.

Benjamin Peherstorfer and Karen Willcox. Data-driven operator inference for nonintrusive projection-based model reduction. *Computer Methods in Applied Mechanics and Engineering*, 2016.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 2017.

Javier E Santos, Zachary R Fox, Arvind Mohan, Daniel O'Malley, Hari Viswanathan, and Nicholas Lubbers. Development of the Senseiver for efficient field reconstruction from sparse observations. *Nature Machine Intelligence*, 5(11):1317–1325, 2023.

Louis Serrano, Armand Kassaï Koupaï, Thomas X Wang, Pierre Erbacher, and Patrick Gallinari. Zebra: In-context generative pretraining for solving parametric PDEs. In *International Conference on Machine Learning*, 2025.

Louis Serrano, Jiequn Han, Edouard Oyallon, Shirley Ho, and Rudy Morel. Test-time generalization for physics through neural operator splitting. In *International Conference on Machine Learning*, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 2020.

Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, pp. 366–381. Springer, 1981. doi: 10.1007/BFb0091924.

Matteo Tomasetto, Jan P. Williams, Francesco Braghin, Andrea Manzoni, and J. Nathan Kutz. Reduced order modeling with shallow recurrent decoder networks. *Nature Communications*, 2025.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Ke Alexander Wang, Jiaxin Shi, and Emily B Fox. Test-time regression: a unifying framework for designing sequence models with associative memory. *Journal of Machine Learning Research*, 27 (169):1–41, 2026a.

Rui Wang, Robin Walters, and Rose Yu. Meta-learning dynamics forecasting using task inference. In *Advances in Neural Information Processing Systems*, 2022.

Zixuan Wang, Xingyu Dang, Rui-Jie Zhu, Zixin Wen, Hengyu Fu, Wenhao Chai, and Jason D Lee. Learning what to remember: Test-time training via context distillation. *arXiv preprint arXiv:2608.01672*, 2026b.

Jan P. Williams, Olivia Zahn, and J. Nathan Kutz. Sensing with shallow recurrent decoder networks. *Proceedings of the Royal Society A*, 480(2298):20240054, 2024. doi: 10.1098/rspa.2024.0054.

Liu Yang, Siting Liu, Tingwei Meng, and Stanley J. Osher. In-context operator learning with data prompts for differential equation problems. *Proceedings of the National Academy of Sciences*, 2023.

Yuan Yin, Ibrahim Ayed, Emmanuel de Bézenac, Nicolas Baskiotis, and Patrick Gallinari. LEADS: Learning dynamical systems that generalize across environments. In *Advances in Neural Information Processing Systems*, 2021.

Xingyue Zhang, Yuxuan Bao, Mars Liyao Gao, and J. Nathan Kutz. SENDAI: A hierarchical sparse-measurement, efficient data assimilation framework. *arXiv preprint arXiv:2601.21664*, 2026.

A Experimental Setup

All three systems follow a common protocol. Trajectories are simulated for a set of source and test parameter values from a shared initial condition, one pooled SINDy-SHRED model is trained per seed on the source regimes, the dynamics and decoder families are fitted once on the source latents, and each test regime is then adapted from its calibration window and forecast over a fixed horizon. Appendices A.1–A.5 specify each step.

A.1 Data Generation

Table 3 summarizes the simulations. Within each system, every regime starts from the same initial condition, so trajectories differ only through the swept parameter. Fields are min–max scaled to $[0, 1]$ per grid point on the pooled training rows, and errors are reported after inverting this scaling.

Viscous Burgers. Equation (8) is integrated with Fourier pseudo-spectral derivatives and an adaptive RK45 scheme, with output every $\Delta t = 0.01$ up to $T = 10$, from the initial condition $u_0(x) = \sin x + 0.5 \sin 2x + 0.25 \cos 3x$.

Gray–Scott. Equation (9) is advanced by Strang splitting, with exact Fourier half-steps for diffusion around one RK4 reaction step at $\delta t = 1$. After 200 burn-in steps, every fifth state is recorded, and the time axis is rescaled by 3×10^{-4}, giving $\Delta t = 1.5 \times 10^{-3}$. The initial condition is $(U, V) = (1, 0)$ outside and $(0.5, 0.25)$ inside a centered square of half-width 12 cells, perturbed by Gaussian noise of standard deviation 0.02 and clipped to $[0, 1]$. The rates $(F, k) = (0.030, 0.059)$ are shared by all regimes. The model state contains both species, while only V is sensed.

Two-dimensional Navier–Stokes. Equation (10) is solved in vorticity form with a Fourier pseudo-spectral discretization, $2/3$-rule dealiasing, and ETDRK4 (Kassam & Trefethen, 2005) with step 2.5×10^{-4}. Every twentieth state is recorded, giving $\Delta t = 5 \times 10^{-3}$, and $\nu = 12/\mathrm{Re}$. The initial vorticity has random phases and spectral amplitude $(k/k_0)^2/\big(1 + (k/k_0)^4\big)^2$ with $k_0 = 4$, and is dealiased, shifted to zero mean, and scaled to $\max|\omega| = 1$.

Table 3: Simulation settings. Held-out regimes are reported in Table 1 and Appendix B.

	Burgers	Gray–Scott (D_v sweep)	Gray–Scott (D_u sweep)	Navier–Stokes
Domain	$[0, 2\pi)$, periodic	128×128 cells, $\Delta x = 1$, doubly periodic		$[0, 2\pi)^2$, doubly periodic
Grid / state dimension n	128 / 128	128^2 / 32,768 (U and V)		128^2 / 16,384
Time integration	RK45 (adaptive)	Strang splitting, RK4 reaction		ETDRK4, $2/3$ dealiasing
Burn-in	none	200 solver steps		none
Snapshots per regime	1001	800		1000
Model time step Δt	1.0×10^{-2}	1.5×10^{-3}		5.0×10^{-3}
Fixed parameters	none	$D_u = 0.16$	$D_v = 0.08$	$\nu = 12/\mathrm{Re}$
Source regimes	$\nu \in \{0.03, 0.04, 0.05, 0.06, 0.07\}$	$D_v \in \{0.065, 0.066, \ldots, 0.069\}$	$D_u \in \{0.156, 0.158, \ldots, 0.164\}$	$\mathrm{Re} \in \{1000, 1100, \ldots, 1800\}$
Held-out regimes	$\nu \in \{0.020, 0.025, 0.035, \ldots, 0.065, 0.080, 0.090, 0.100\}$	$D_v \in \{0.064, 0.0655, \ldots, 0.0685, 0.070\}$	$D_u \in \{0.154, 0.157, \ldots, 0.163, 0.166\}$	$\mathrm{Re} \in \{800, 900, 1900, 2000\}$
Sensed field / sensors s	u / 16	V / 32	V / 32	ω / 32
Sensor placement	uniform random, per seed	uniform random, min. separation 3 cells, per seed		uniform random, fixed across seeds

A.2 Architecture and Training

All models use the SINDy-SHRED architecture and training objective of Gao et al. (2026b). Table 4 lists the per-system settings.

Architecture. A two-layer GRU with hidden size d_z maps the L most recent sensor vectors to the latent state. The latent is rescaled per dimension to $[-1, 1]$ by the training minimum and maximum, and the SINDy library is evaluated in these coordinates. The decoder is $\mathrm{Linear}(d_z, 350) \to \mathrm{ReLU} \to \mathrm{Linear}(350, 400) \to \mathrm{ReLU} \to \mathrm{Linear}(400, n)$, with dropout 0.1 after each hidden layer during training. Its last layer is the output layer (W, b), and ϕ is its 400-dimensional hidden output.

Training objective. The loss is the reconstruction MSE of the scaled field, plus 10 times the MSE between the next GRU latent and a forward-Euler integration of $\Theta(z)\Xi$ over three substeps of $\Delta t/3$, plus 0.1 times the absolute mean of the latent. An ensemble of five SINDy models with thresholds $0.25 \cdot 10^{0.2i-1}$, $i = 0, \ldots, 4$, is hard-thresholded at a fixed epoch interval. Ξ_{pooled} is the ensemble mean, mapped to the normalized latent coordinates by an exact polynomial change of basis.

Optimization. Training uses AdamW (learning rate 10^{-3}, weight decay 10^{-2}), gradient clipping at norm 10^3, no learning-rate schedule, and no early stopping. The nominal batch size is 128, and each epoch draws its effective batch size uniformly from $[64, 128]$. Each source trajectory is split in time into training, validation, and test segments, and windows from all source regimes are pooled.

Table 4: Architecture and training hyperparameters. Parameter counts include the GRU, the decoder, and the five SINDy coefficient matrices.

	Burgers	Gray–Scott	Navier–Stokes
Lags L	20	20	20
Latent dimension d_z	3	12	12
GRU layers	2	2	2
Decoder hidden widths	350, 400	350, 400	350, 400
Library (polynomial order / size ℓ)	1 / 4	2 / 91	2 / 91
Thresholding interval (epochs)	100	100	200
Epochs	200	200	800
Split per source regime (train / val / test snapshots)	813 / 63 / 125	650 / 50 / 100	800 / 100 / 100
Seeds	0, 1, 2	0, 1, 2	42, 52, 62, 72, 82
Parameters	193,449	13,292,970	6,722,986

A.3 Parametric Families and Test-Time Inference

Context coordinate. Contexts are standardized as $c_i = (\mu_i - \bar{\mu})/\sigma_\mu$, with the sample mean and standard deviation ($I - 1$ normalization) of the source values, so $\sigma_c = 1$. Table 5 lists the contexts of the extrapolation regimes, which measure extrapolation distance in source standard deviations.

Family fits. Both families are fitted on the leading latent rows of each source trajectory. The latent derivative is the forward difference $(z_{t+1} - z_t)/\Delta t$, which matches the one-step Euler rollout. The squared residuals of both ridge regressions, Equations (3) and (4), are averaged over all stacked rows. The decoder bias is fitted as the coefficient of a constant feature equal to the root-mean-square of ϕ over the first source regime, and rescaled afterwards.

Stage 1. The objective is evaluated on G equispaced points of $[-c_{\max}, c_{\max}]$, and Levenberg–Marquardt starts from the grid minimizer with damping τ_0. Each iteration uses a central-difference derivative with step $10^{-3}\sigma_c$ and the damped step of Equation (6), clipped to the box. An accepted step divides the damping by 10; a rejected step multiplies it by 10 and is retried, up to 25 times, with the damping kept in $[10^{-12}, 10^{10}]$. Iteration stops after K iterations, when the step falls below $10^{-10}(1 + |c|)$, or when no retry is accepted.

Stage 2. The decoder context minimizes Equation (7) and is computed in closed form by Equation (14) on the Stage 1 rollout, with sums over all calibration steps and sensors.

Interaction between the two contexts. Both contexts reduce the same sensor residual in Equations (5) and (7) and are therefore partially confounded, since a dynamics error that shifts the rolled-out latent and a decoder error that mis-maps a correct latent state both produce residuals that are not mutually orthogonal in general. Solving for c_{dyn} at $c_{\mathrm{dec}} = 0$ does not remove this confounding; it fixes the order in which the residual is attributed. The centroid correction $(\Delta W_0, \Delta \mathbf{b}_0)$ in (5) is what makes this ordering interpretable. Without it, the minimizer of $\mathcal{J}$ would trade latent adaptation against a static decoder bias, and $c^\star_{\mathrm{dyn}}$ would be an uninterpretable mixture of the two mismatches. The cost of the ordering is that $c^\star_{\mathrm{dyn}}$ is biased whenever the true decoder correction is large, since part of that correction is attributed to the dynamics.

Table 5: Context coordinates, family penalties, and inference settings.

	Burgers	Gray–Scott (D_v)	Gray–Scott (D_u)	Navier–Stokes
$\bar{\mu}$ / σ_μ	0.05 / 0.0158	0.067 / 0.00158	0.16 / 0.00316	1400 / 273.9
Extrapolation contexts c	$-1.90,\ -1.58,\ 1.90,\ 2.53,\ 3.16$	$-1.90,\ 1.90$	$-1.90,\ 1.90$	$-2.19,\ -1.83,\ 1.83,\ 2.19$
Family fit rows per source regime	560	600	600	800
$\lambda_{\Xi,0}$ / $\lambda_{\Xi,1}$		10^{-2} / 10^{-2}		
$\lambda_{W,0}$ / $\lambda_{W,1}$		10^{-3} / 10^{-2}		
λ_{dyn}		10^{-3}		
λ_{dec}	10^{-3}	1	1	1
$c_{\max}$	4.0	3.0	3.0	3.0
Grid size G / iteration cap K / initial damping τ_0		61 / 50 / 10^{-3}		
Calibration window C	250	125	150	200
Forecast horizon H	300	300	300	550

A.4 Evaluation Protocol

Windows. Latent row t is the encoder output for the L sensor vectors ending at snapshot $t + L - 1$, so the first row corresponds to snapshot 19. The calibration window is the first C rows of a test trajectory. Stage 1 rolls out from the encoder latent of the first row. The forecast starts from the encoder latent of row C and covers the next H rows, so calibration and evaluation windows do not overlap.

Arms. Table 6 defines the four arms. They share the encoder, the anchor state, and the integrator, and differ only in the coefficient matrix and output layer used for forecasting.

Table 6: Definition of the four evaluation arms.

Arm	Coefficients	Output layer	Decoder context fitted on
No-adapt	Ξ_{pooled}	(W, b)	not used
Decoder only	Ξ_{pooled}	$(W + \Delta W(c), b + \Delta b(c))$	calibration rollout of Ξ_{pooled}
Dynamics only	$\Xi(c^\star_{\text{dyn}})$	(W, b)	not used
Full	$\Xi(c^\star_{\text{dyn}})$	$(W + \Delta W(c^\star_{\text{dec}}), b + \Delta b(c^\star_{\text{dec}}))$	calibration rollout of $\Xi(c^\star_{\text{dyn}})$

Metrics. The decoded error is $\|\hat{X} - X\|_F / \|X\|_F$, where $X \in \mathbb{R}^{H \times n}$ stacks the H forecast snapshots in physical units. For Gray–Scott, X contains both species, so the error includes the unsensed substrate U. The latent error uses the same formula on the normalized latent trajectory, with the encoder latents of the true test sensor sequence as reference. Rollouts that become non-finite are recorded as divergent.

Seeds and aggregation. For Burgers and Gray–Scott, the seed sets the initialization, the batch sizes, the dropout masks, and the sensor locations. For Navier–Stokes, the sensors are drawn once and shared across seeds. Burgers and Gray–Scott results are medians over seeds, and Navier–Stokes results are means. Seeds whose rollout diverges are excluded from these statistics, which leaves two seeds for the D_v sweep and one for the D_u sweep. The horizon curves of Section 4 show the mean over regimes of the per-regime median, and those of Appendix B.2 show the per-regime median with interquartile bands. In Appendix C, noise of standard deviation $\sigma \cdot \text{std}_t(y_j)$ is added to each sensor j of the test trajectories only.

A.5 Computational Cost

Adaptation computes no gradients with respect to network weights. It evaluates G grid points and a few Levenberg–Marquardt iterations of a C-step rollout decoded only at the s sensor rows, then one closed-form solve, so its cost is independent of the grid size n. Table 7 reports wall-clock times on Navier–Stokes. Inferring both contexts for a new regime takes about 3.5 s on a MacBook Pro, more than three orders of magnitude less than training the pooled model on an H100, which would also require full-state data from the target regime. Burgers and Gray–Scott models were trained and adapted on the same MacBook Pro.

Table 7: Wall-clock cost on Navier–Stokes (128^2 grid, $d_z = 12$, $C = 200$, $H = 550$). MacBook Pro timings use the models of Appendix E.

Stage	Frequency	Hardware	Wall-clock
Pooled training, 800 epochs	once per seed	NVIDIA H100, shared by 5 jobs	98–124 min
Pooled training, 200 epochs	once per seed	MacBook Pro (M4)	$\approx$ 20 min
Family fit, incl. encoding 9 source trajectories	once per seed	MacBook Pro (M4)	2.0–2.5 s
Context inference (Stages 1 and 2)	per target regime	MacBook Pro (M4)	3.4–4.1 s
Full adaptation job (4 regimes, 4 arms, evaluation)	once per seed	4 cores of a CPU cluster node	69–76 s

B Additional Experimental Results

This appendix reports the held-out regimes not shown in Table 1, including regimes inside the source range, together with per-regime error-versus-horizon curves, a sweep over the calibration window, and additional qualitative results. All numbers come from the runs of Section 4, with the settings of Appendix A and the aggregation rules of Appendix A.4.

B.1 Per-Regime Errors

Tables 8–11 report latent and decoded errors for the remaining tested regimes of each system. Together with Table 1, they cover every held-out regime of the main experiments.

Adaptation inside and outside the source range. Full adaptation has a lower decoded error than No-adapt in every tested regime of all four sweeps. The gain is not limited to extrapolation. Inside the source range, Full lowers the decoded error by 66–69% on Burgers, 39–63% on the D_v sweep, and 48–72% on the D_u sweep, so adaptation also improves on the pooled model between training values.

Contribution of each context. Across all tested regimes, Full attains the lowest decoded error in every Burgers and D_v regime and in four of the six D_u regimes. In the remaining two D_u regimes, Dynamics only is lower by at most 0.001. On Navier–Stokes, Dynamics only is lowest at all four Reynolds numbers, for the reason discussed in Section 4.3. Decoder only is worse than No-adapt in eight of the nine Burgers regimes and within 0.002 of it on Gray–Scott. On Navier–Stokes it improves on No-adapt below the source range and is marginally worse above it. The dynamics context lowers the latent error in every regime except $D_v = 0.070$. The large latent error of the adapted arms at $\mathrm{Re} = 2000$ in Table 1 comes from a single seed whose latent state grows while its reconstruction stays accurate.

Table 8: Burgers relative errors at the regimes not shown in Table 1, with $C = 250$ and $H = 300$, median over seeds 0, 1, 2. Bold and underline mark the lowest and second-lowest decoded error in each row; † marks viscosities inside the source range $\nu \in [0.03, 0.07]$.

ν	No-adapt		Decoder only		Dynamics only		Full	
	Latent	Decoded	Latent	Decoded	Latent	Decoded	Latent	Decoded
0.025	0.552	0.299	0.552	0.334	0.205	<u>0.185</u>	0.205	**0.099**
0.035†	0.539	0.309	0.539	0.328	0.212	<u>0.152</u>	0.212	**0.104**
0.045†	0.530	0.309	0.530	0.324	0.229	<u>0.126</u>	0.229	**0.104**
0.055†	0.533	0.294	0.533	0.321	0.219	<u>0.104</u>	0.219	**0.098**
0.065†	0.539	0.279	0.539	0.316	0.207	<u>0.089</u>	0.207	**0.086**
0.080	0.547	0.257	0.547	0.308	0.187	<u>0.089</u>	0.187	**0.079**
0.090	0.552	0.268	0.552	0.302	0.173	<u>0.101</u>	0.173	**0.080**

Table 9: Gray–Scott D_v sweep relative errors at the regimes not shown in Table 1, with $C = 125$ and $H = 300$, median over seeds 0 and 1 (Appendix A.4). Bold and underline mark the lowest and second-lowest decoded error in each row; † marks regimes inside the source range $D_v \in [0.065, 0.069]$.

D_v	No-adapt		Decoder only		Dynamics only		Full	
	Latent	Decoded	Latent	Decoded	Latent	Decoded	Latent	Decoded
0.0655†	0.885	0.222	0.885	0.224	0.415	<u>0.090</u>	0.415	**0.082**
0.0665†	0.896	0.216	0.896	0.218	0.423	<u>0.092</u>	0.423	**0.081**
0.0675†	0.932	0.215	0.932	0.216	0.554	<u>0.116</u>	0.554	**0.106**
0.0685†	0.946	0.219	0.946	0.219	0.685	<u>0.145</u>	0.685	**0.134**

Table 10: Gray–Scott D_u sweep relative errors at the regimes not shown in Table 1, with $C = 150$ and $H = 300$, for seed 0 (Appendix A.4). Bold and underline mark the lowest and second-lowest decoded error in each row, ranked before rounding; † marks regimes inside the source range $D_u \in [0.156, 0.164]$.

D_u	No-adapt		Decoder only		Dynamics only		Full	
	Latent	Decoded	Latent	Decoded	Latent	Decoded	Latent	Decoded
0.1570†	0.637	0.148	0.637	0.146	0.562	<u>0.084</u>	0.562	**0.076**
0.1590†	0.733	0.180	0.733	0.178	0.594	<u>0.086</u>	0.594	**0.081**
0.1610†	0.824	0.211	0.824	0.209	0.445	<u>0.085</u>	0.445	**0.085**
0.1630†	0.885	0.233	0.885	0.232	0.331	**0.065**	0.331	<u>0.066</u>

Table 11: Navier–Stokes relative errors at the regimes not shown in Table 1, with $C = 200$ and $H = 550$, mean over seeds 42, 52, 62, 72, 82. Bold and underline mark the lowest and second-lowest decoded error in each row.

Re	No-adapt		Decoder only		Dynamics only		Full	
	Latent	Decoded	Latent	Decoded	Latent	Decoded	Latent	Decoded
900	0.545	0.209	0.545	0.204	0.135	**0.029**	0.135	<u>0.033</u>
1900	1.241	0.159	1.241	0.160	0.179	**0.017**	0.179	<u>0.020</u>

B.2 Error Across the Forecast Horizon

Figures 4–6 show the decoded error of the four arms as a function of the forecast horizon for every tested regime. The main text shows only the average over regimes, for Burgers and the D_v sweep. The corresponding per-regime curves for Navier–Stokes appear in Figure 14 of Appendix D.

Burgers. In the extrapolation regimes, Full lies below No-adapt at every horizon. At $\nu \in \{0.045, 0.055, 0.065\}$, inside the source range, the pooled model matches or outperforms the adapted arms over the first few tens of steps. The source dynamics are already close to the target there, and adaptation separates from the pooled model only as the rollout error accumulates. Decoder only lies above No-adapt at almost every horizon and has the widest interquartile band.

Gray–Scott. Decoder only is slightly more accurate than No-adapt at the shortest horizon in most regimes and indistinguishable from it afterwards. Full is the most accurate arm at nearly every horizon of both sweeps. Its margin over Dynamics only is largest at short horizons, where the decoder correction dominates, and narrows as the latent error grows. In the D_u sweep, the two arms coincide at the longest horizons.

Navier–Stokes. Below the source range, both arms with a decoder correction reduce the short-horizon error. No-adapt and Decoder only then rise steeply beyond $H = 400$, while the adapted arms remain below 0.07. Above the source range, the pooled model is more accurate than Full up to about $H = 300$, and the benefit of adaptation appears only at the longest horizons, where the pooled median reaches 0.12 while Dynamics only remains at 0.009 ($\mathrm{Re} = 1900$) and 0.029 ($\mathrm{Re} = 2000$).

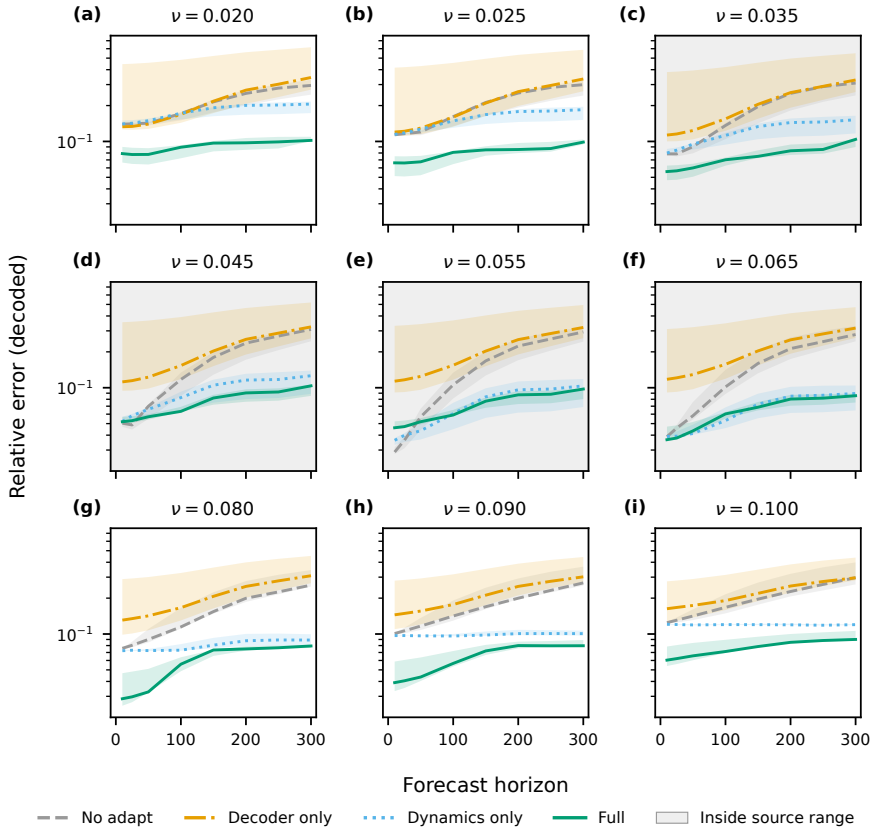


Figure 4: Decoded relative error versus forecast horizon on every tested Burgers viscosity at $C = 250$. Lines show the median over seeds 0, 1, 2 and bands the interquartile range. Gray panels mark viscosities inside the source range.

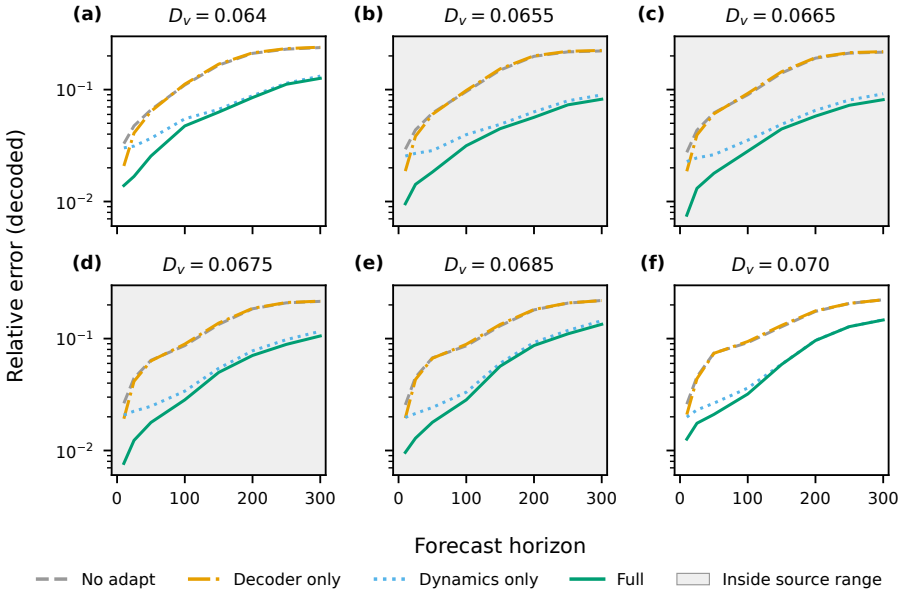


Figure 5: Decoded relative error versus forecast horizon on every tested regime of the Gray–Scott D_v sweep at $C = 125$, median over seeds 0 and 1. Gray panels mark regimes inside the source range.

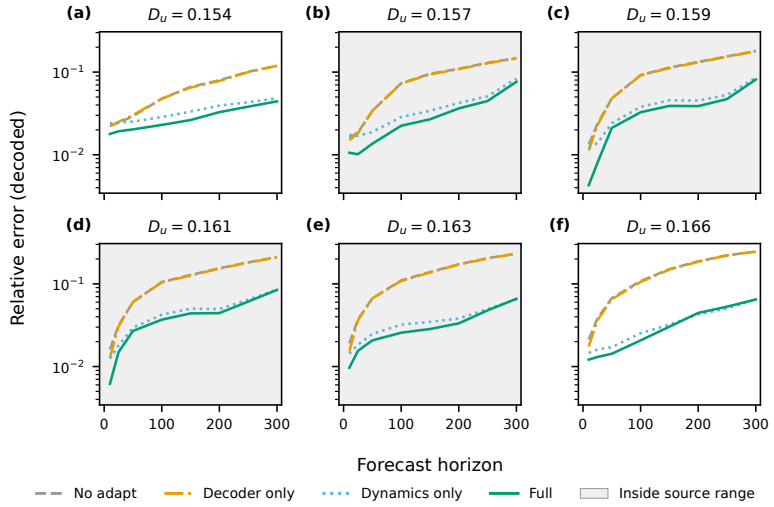


Figure 6: Decoded relative error versus forecast horizon on every tested regime of the Gray–Scott D_u sweep at $C = 150$, for seed 0. Gray panels mark regimes inside the source range.

B.3 Calibration Window Length

Figures 7–9 report the decoded error at $H = 300$ as the calibration window C varies, averaged over all tested regimes. At each window, seeds whose rollout diverges are excluded, and windows without a remaining seed are omitted. The No-adapt curve varies with C for two reasons. The forecast starts at the end of the window, so each C evaluates a different segment of the trajectory, and the set of seeds retained differs between windows.

Full is more accurate than No-adapt at every calibration window of all three sweeps. On Burgers (Figure 7), its error decreases monotonically from 0.201 at $C = 50$ to 0.093 at $C = 250$, consistent with longer windows constraining the context more tightly. The D_v sweep (Figure 8) reaches its lowest error, 0.113, at $C = 125$ and rises to 0.151 at $C = 200$. In the D_u sweep (Figure 9), the error varies between 0.064 and 0.091 without a clear trend. The main-text windows, marked by the dashed lines, coincide with the lowest Full error for Burgers and the D_v sweep.

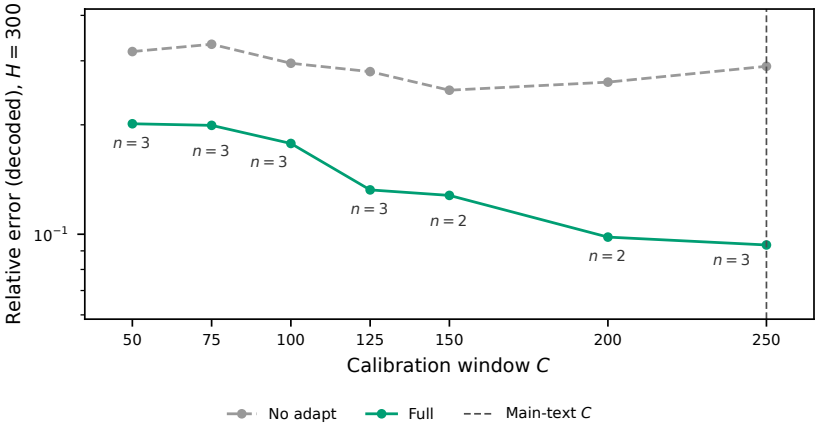


Figure 7: Decoded relative error at $H = 300$ versus calibration window C for Full and No-adapt on Burgers. Each point is the median over seeds within each test regime, averaged over all tested regimes. Labels give the number of seeds, and the dashed line marks the main-text window.

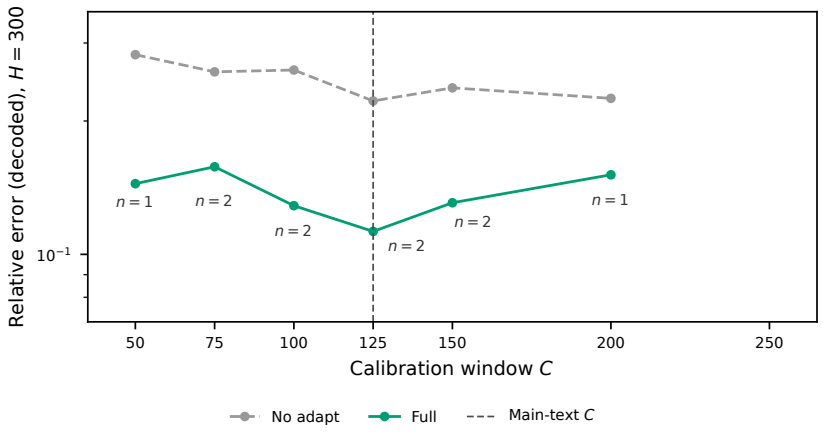


Figure 8: Decoded relative error at $H = 300$ versus calibration window C for Full and No-adapt on the Gray–Scott D_v sweep, aggregated as in Figure 7.

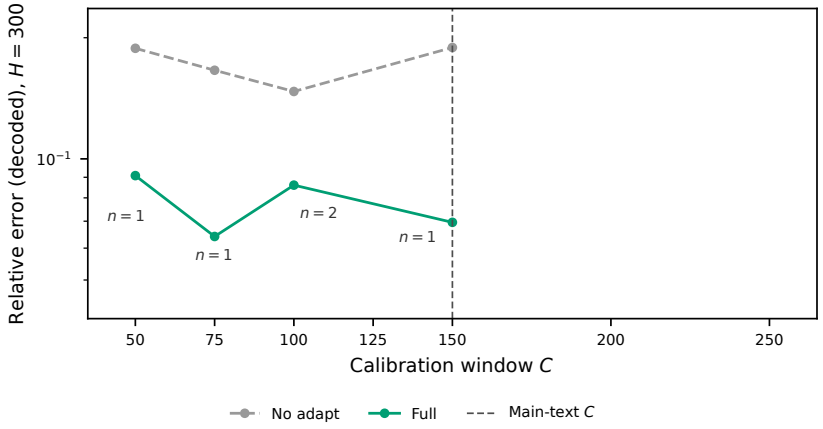


Figure 9: Decoded relative error at $H = 300$ versus calibration window C for Full and No-adapt on the Gray–Scott D_u sweep, aggregated as in Figure 7.

B.4 Qualitative Results

Figures 10–12 show the ground truth, the No Adapt and Full Adapt forecasts, and their signed residuals, prediction minus ground truth, with one color scale shared across the field panels and one across the residual panels of each figure. The fields come from a single seed per regime. For Burgers and Gray–Scott, the displayed seed is the seed whose Full error is closest to the tabulated median, and the Navier–Stokes fields come from a single run of the main experiment.

Burgers. Figure 10 shows the space-time evolution at $\nu = 0.080$ and $\nu = 0.090$, the extrapolation regimes not displayed in Figure 1. For the displayed seeds, the pooled forecast errs in opposite directions at the two viscosities. At $\nu = 0.080$ it keeps a steeper, higher-amplitude front than the ground truth, and its residual grows with the forecast step on both sides of the shock. At $\nu = 0.090$ it underestimates the amplitude on both sides of the front. In both regimes, Full Adapt reproduces the diffusion of the front and confines its residual to a narrow band at the shock location.

Gray–Scott. Figure 11 adds the signed residuals to the fields of Figure 2. The pooled model smooths the pattern and underestimates its amplitude, most visibly in the D_u sweep, where the parameter acts on the unobserved substrate. Full Adapt restores the pattern at $D_v = 0.064$ and $D_u = 0.166$, and its residual at $D_u = 0.166$ is close to zero. The adapted forecast also shows

artifacts, namely isolated high-amplitude spots at $D_v = 0.070$ and a spurious four-lobed structure at the center of the domain at $D_u = 0.154$, whose residual in this frame exceeds that of the pooled model. The decoded errors of Table 1 aggregate both species over all 300 forecast steps, so a single V frame at the final step need not reflect them.

Navier–Stokes. Figure 12 shows the vorticity at $H = 550$ for the four test Reynolds numbers. The pooled model preserves the arrangement of the coherent vortices but overestimates their amplitude, with residuals concentrated at the vortex cores. The Full Adapt residuals are small at all four Reynolds numbers, consistent with the decoded errors of Tables 1 and 11.

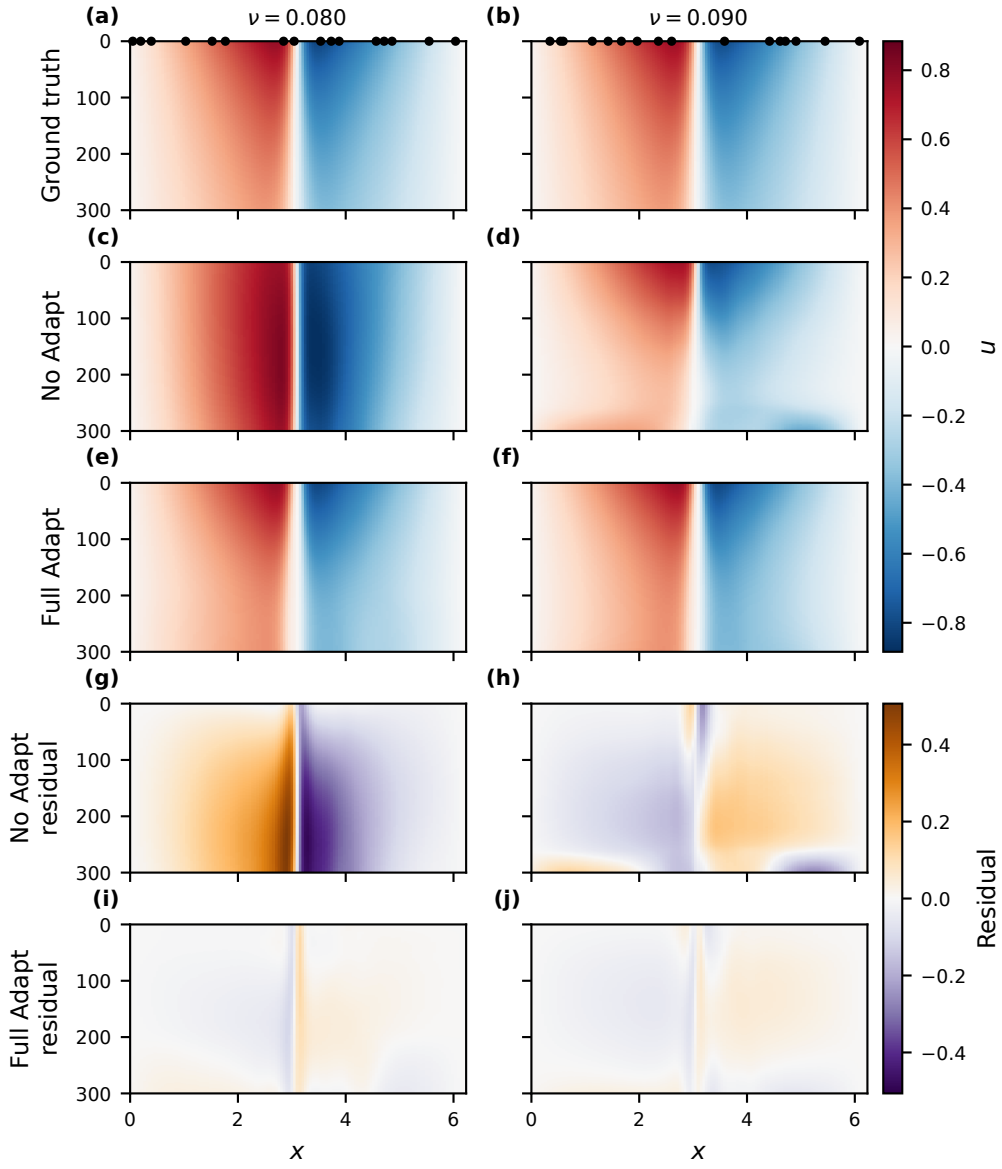


Figure 10: Burgers space-time evolution of u over the $H = 300$ step forecast at $\nu = 0.080$ and $\nu = 0.090$ with $C = 250$, showing the ground truth, the No Adapt and Full Adapt forecasts, and their signed residuals, with time increasing downward. Black dots mark the 16 sensors.

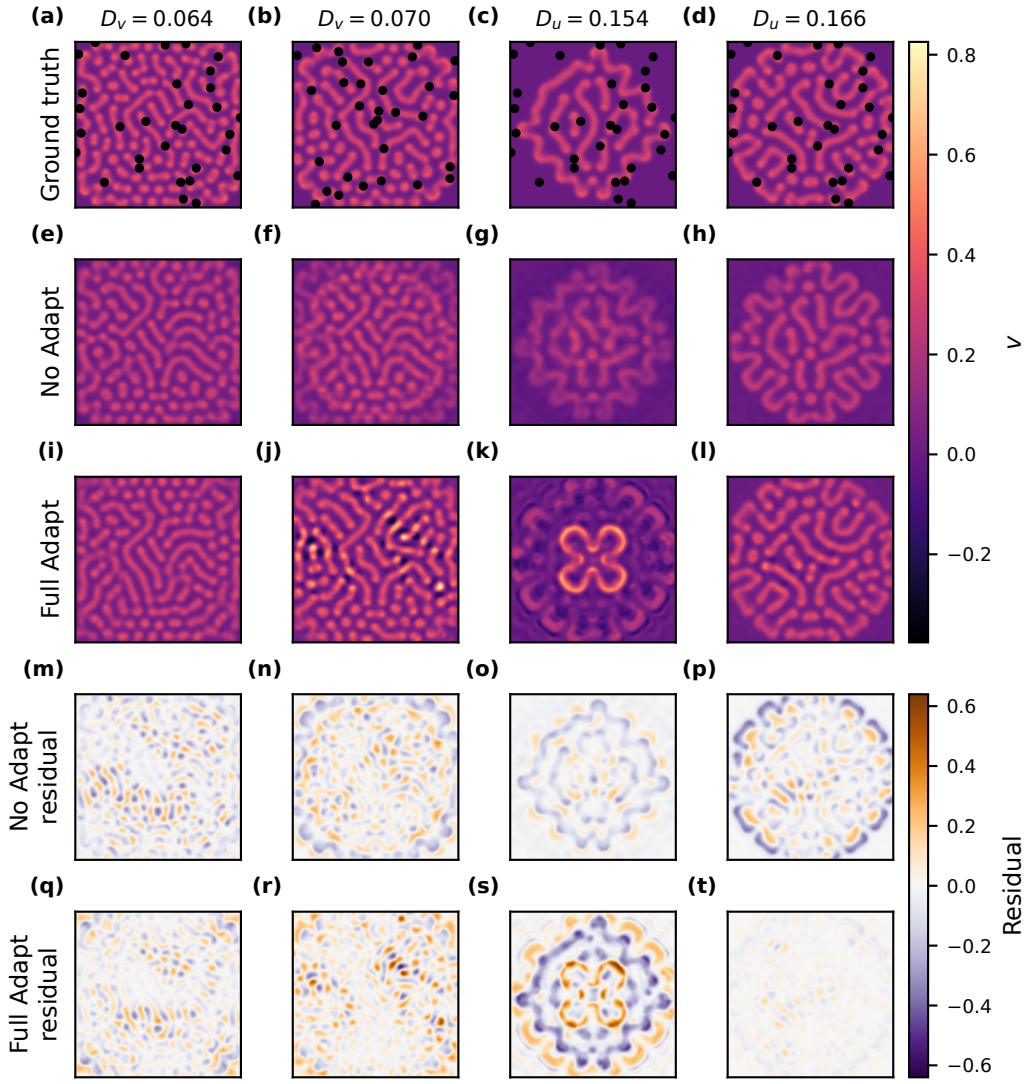


Figure 11: Gray–Scott V field at $H = 300$ for two regimes of each sweep, with $C = 125$ (D_v) and $C = 150$ (D_u), showing the ground truth, the No Adapt and Full Adapt forecasts, and their signed residuals. Black dots mark the 32 V sensors.

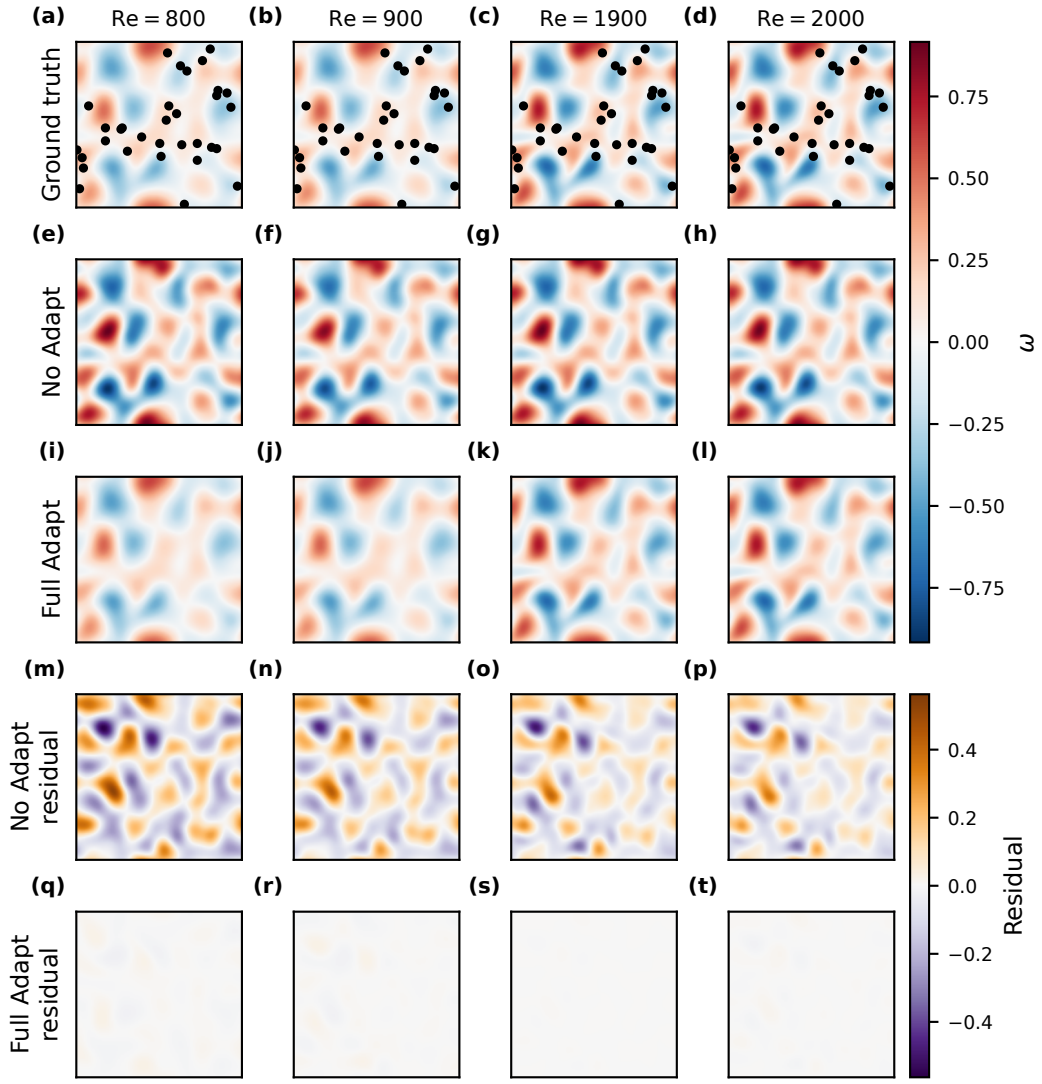


Figure 12: Navier–Stokes vorticity at $H = 550$ with $C = 200$ for the four test Reynolds numbers, showing the ground truth, the No Adapt and Full Adapt forecasts, and their signed residuals. Black dots mark the 32 sensors, which are shared across seeds and regimes.

C Sensitivity Analysis

We conduct a sensitivity study on the two-dimensional Navier–Stokes system to assess how each architectural and adaptation hyperparameter affects reconstruction quality under distribution shift. The system is simulated on a 128×128 spatial grid across various Reynolds numbers. In the results below, we focus on the below-training test regimes, which require the model to extrapolate toward lower Reynolds numbers than any observed during training. Each experiment varies a single parameter while holding all others at the default configuration used throughout Section 4: $s = 32$ sensors, lag window $L = 20$, latent dimension $d_z = 12$, calibration window $C = 200$, $\lambda_{\text{dyn}} = 10^{-3}$, and $\lambda_{\text{dec}} = 1$. Every configuration is evaluated over five independent seeds, and we report the decoded relative ℓ^2 error (mean $\pm$ std) as a function of forecast horizon. Figure 13 summarizes the results across five parameter axes and a four-arm adaptation comparison.

Latent dimension. Varying the latent dimension $d_z \in \{4, 6, 8, 10, 12\}$ shows that the smallest value tested ($d_z = 4$) incurs noticeably higher error at longer forecast horizons, consistent with

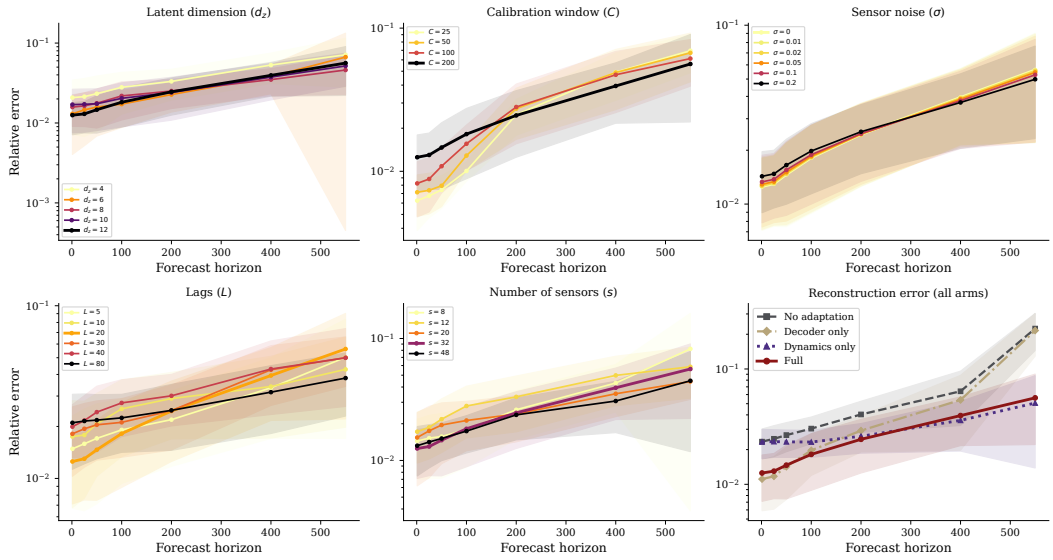


Figure 13: One-at-a-time sensitivity study on the Navier–Stokes system at below-training Reynolds numbers ($\mathrm{Re} \in \{800, 900\}$). Each of the first five panels varies one hyperparameter under full adaptation while all others are held at default; the bottom-right panel compares the four adaptation arms at the default configuration. Lines show the mean decoded relative ℓ^2 error over five seeds; shaded bands indicate $\pm$ one standard deviation. All panels use a logarithmic error axis.

an insufficient latent capacity to represent the dynamics across the rollout. The remaining values produce broadly similar performance, with errors converging as d_z increases. This suggests that the SINDy-based dynamics and the decoder correction are robust to moderate reductions in latent dimensionality, provided necessary capacity is met.

Calibration window. The calibration window $C \in \{25, 50, 100, 200\}$ controls the number of observed time steps available for inferring the dynamics and decoder contexts at test time. Smaller windows ($C = 25$, $C = 50$) yield higher and more variable error at longer horizons, reflecting the difficulty of estimating c_{dyn} from limited calibration data. Longer forecast horizons suffer more from insufficient calibration window, where a poorly estimated dynamics context leads to accumulating drift in the latent rollout.

Sensor noise. Adding Gaussian sensor noise at levels $\sigma \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2\}$ progressively degrades reconstruction accuracy on smaller forecast horizons, but only marginally. The adaptation mechanism continues to function at $\sigma = 0.2$, indicating that the sensor-space residual used for context inference is robust to measurement noise.

Lag window. The lag parameter $L \in \{5, 10, 20, 30, 40, 80\}$ determines the length of the temporal window processed by the GRU unit to produce the latent state. Performance improves as L increases from 5 to 20 and remains largely stable for $L \geq 20$, suggesting that the temporal unit's ability to exploit additional past observations saturates at sufficient temporal context, which could be dependent on the time scale of the dynamics.

Number of sensors. Varying the number of randomly placed sensors $s \in \{8, 12, 20, 32, 48\}$ shows that reconstruction error decreases with denser spatial sampling. This trend is expected: more sensors provide a richer spatial context, but with diminishing returns under the current single-physics setting.

Adaptation arm comparison. The bottom-right panel of Figure 13 compares the four adaptation arms at the default configuration. At short forecast horizons, decoder-only adaptation provides a meaningful reduction in error relative to the unadapted pooled model. At these short horizons the latent trajectory has not yet diverged substantially from the target, and the reconstruction correction

Table 12: Architectural components of TTA-SINDy-SHRED and the baselines.

	TTA-SINDy-SHRED	FNO	FNO-masked	DeepONet	Senseiver
Input	sensor window	full field	masked sensors	sensor window	sensor window
Architecture	GRU	spectral conv	spectral conv	branch MLP	cross-attention
Dynamics model	SINDy ODE $\Theta(\mathbf{z})\Xi(c)$	ConditionalMLP $g_\theta(\mathbf{z}, c)$			
Dynamics time	continuous (Euler)	discrete			
Dynamics training	joint (SINDy loss)	post-hoc (MSE on frozen latents)			
Interpretable dynamics	✓	—			
Decoder	$W\,\phi(\mathbf{z}) + \mathbf{b}$ (shared structure)				
Decoder family	$\Delta W(c), \Delta\mathbf{b}(c)$ (shared)				
Adaptation DOF	c_{dyn}, c_{dec} (two scalars)				

captured by c_{dec} can compensate for the initial regime mismatch in the observation map. Dynamics-only adaptation achieves comparable error at these short horizons. As the forecast horizon grows, however, the dynamics-only arm surpasses decoder-only: without dynamics adaptation, the latent rollout accumulates systematic error and the decoder correction alone cannot compensate for the resulting trajectory drift. Full adaptation, which combines both contexts, attains a balance at all target forecast horizons by applying the corrected latent dynamics before decoder adjustment. This crossover—decoder-only effective at short horizons, dynamics-only increasingly important at longer horizons—is consistent with the complementary roles of c_{dyn} and c_{dec}: the dynamics context corrects the temporal latent evolution over the forecast window, while the decoder context refines the spatial reconstruction at each time step.

D Baseline Comparison

To test whether the observed test-time adaptation gains generalize beyond the proposed the architectural choices—the GRU temporal unit, the SINDy-discovered sparse polynomial dynamics, and the joint training objective that co-optimizes both—we construct four baseline methods that replace both the temporal unit and the latent dynamics model while retaining the decoder structure and the two-scalar adaptation mechanism. Each baseline pairs a different representation-learning neural architecture, with a post-hoc neural dynamics model and a decoder family analogous to those in the TTA-SINDy-SHRED full method. The four baselines consist of a Fourier Neural Operator (FNO) operating on full-field snapshots (Li et al., 2020), a masked variant of the FNO that receives only sparse sensor observations (FNO-masked), a DeepONet-style branch network (Lu et al., 2021), and the Senseiver (Santos et al., 2023). While the downstream pipeline is held fixed across all five methods, the learning architectures span four distinct representation-learning paradigms—spectral convolution on complete and on sparse spatial inputs, branch-network function approximation, and set-valued cross-attention—collectively testing whether our adaptation mechanism is complementary to, rather than dependent on, the specific architecture and dynamics model of the proposed method. The goal is not to benchmark these architectures against one another, but to verify that the two-scalar adaptation mechanism remains effective across fundamentally different representation-learning choices. Table 12 summarizes the architectural components of each method.

D.1 Baseline Architectures

Fourier Neural Operator (FNO). The FNO was designed for mapping between full fields rather than sparse sensor readings (Li et al., 2020). A stack of spectral convolution layers processes a window of L consecutive snapshots in Fourier space, and global average pooling followed by a linear projection yields the latent state $\mathbf{z}_t \in \mathbb{R}^{d_z}$. Because this architecture requires the entire spatial field—and its standard formulation does not prescribe a mechanism for processing temporal sensor histories under extreme sparsity or adapting to unseen operating regimes—it serves as a reference point rather than a deployment-realistic alternative: full-field observations are typically unavailable in the sparse-sensing setting that motivates the SHRED framework.

Masked FNO (FNO-masked). The FNO-masked baseline applies the same spectral convolution architecture as the standard version, but adapted to operate on sparse sensor observations. At each time step, the input consists of sensor measurements and a binary mask indicating the observed

locations on the spatial grid. Because spectral convolutions are designed to exploit correlations across a fully populated grid, operating on masked input under extreme sparsity is a substantial deviation from the architecture's intended use. This variant nevertheless receives the same sparse information as the DeepONet, Senseiver, and TTA-SINDy-SHRED, providing a direct comparison of spectral convolution against the other representation-learning paradigms under the same input constraints.

DeepONet branch network. The DeepONet baseline flattens a window of L consecutive readings from s sensors into a single vector and maps it to $\mathbb{R}^{d_z}$ through a multi-layer perceptron serving as the branch network (Lu et al., 2021). Its branch architecture is well suited to fixed-location sensor measurements, as it operates directly on discrete evaluations of the input field. Spatial reconstruction is handled by the decoder described below.

Senseiver. The Senseiver baseline uses a Perceiver-style set-attention mechanism over sparse sensor inputs (Santos et al., 2023). Each sensor's time-lagged readings are embedded via a linear projection with learned positional encodings, and a set of learned latent tokens aggregates information from these sensor embeddings via cross-attention. The resulting tokens are refined through several self-attention layers, mean-pooled, and linearly projected to yield $\mathbf{z}_t \in \mathbb{R}^{d_z}$.

D.2 Decoder and Test-Time Adaptation

Each baseline uses a three-layer MLP decoder analogous to that of the TTA-SINDy-SHRED decoder:

$$\hat{\mathbf{x}}_t = W\,\phi(\mathbf{z}_t) + \mathbf{b}, \qquad \phi(\mathbf{z}) = \mathrm{ReLU}\big(W_2\ \mathrm{ReLU}(W_1\,\mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2\big), \tag{12}$$

where $\phi\colon \mathbb{R}^{d_z} \to \mathbb{R}^{p}$ is the composition of the decoder's hidden layers. The model is trained end-to-end by minimizing the mean squared reconstruction error on pooled source-regime data with pointwise standardization matching the preprocessing used in TTA-SINDy-SHRED. Unlike TTA-SINDy-SHRED, no dynamics regularization is applied during training; the latent representations are shaped solely by reconstruction quality. After convergence, the entire network is frozen.

At test time, the decoder context c_{dec} adapts the frozen output layer through the same decoder family used in TTA-SINDy-SHRED (Section 3). With $\Delta W(c) = \Delta W_0 + c\,\Delta W_1$ and $\Delta\mathbf{b}(c) = \Delta\mathbf{b}_0 + c\,\Delta\mathbf{b}_1$, the adapted output layer is

$$\hat{\mathbf{x}}_t = \big(W + \Delta W(c_{\mathrm{dec}})\big)\,\phi(\mathbf{z}_t) + \mathbf{b} + \Delta\mathbf{b}(c_{\mathrm{dec}}). \tag{13}$$

The family parameters $(\Delta W_0, \Delta W_1, \Delta\mathbf{b}_0, \Delta\mathbf{b}_1)$ are fitted by ridge regression on the source-regime decoder residuals (Equation (4)), and c^*_{dec} is determined in closed form at test time by scalar ridge regression on the sensor residual of the adapted rollout:

$$c^\star_{\mathrm{dec}} = \frac{\sum_{t=1}^{C} \big\langle S\big[\Delta W_1\phi(\hat{\mathbf{z}}_t) + \Delta\mathbf{b}_1\big],\, \mathbf{y}_t - S\big[(W + \Delta W_0)\phi(\hat{\mathbf{z}}_t) + \mathbf{b} + \Delta\mathbf{b}_0\big]\big\rangle}{\sum_{t=1}^{C} \big\|S\big[\Delta W_1\phi(\hat{\mathbf{z}}_t) + \Delta\mathbf{b}_1\big]\big\|_2^2 + \lambda_{\mathrm{dec}}\,\frac{1}{\sigma_c^2}}. \tag{14}$$

D.3 Latent Dynamics

In TTA-SINDy-SHRED, the SINDy loss is applied jointly during training so that the temporal unit learns a latent space in which the dynamics are well approximated by a sparse polynomial library, and the coefficient matrix is affine in the dynamics context c_{dyn}. For the neural baselines, the representation network is trained with reconstruction loss alone, so the latent space carries no such structural guarantee. Applying a SINDy library post-hoc to these unconstrained latent trajectories would be unreliable, as the dynamics may not be well approximated by sparse polynomial terms.

We therefore replace the SINDy ODE with a discrete-time ConditionalMLP $g_\theta\colon \mathbb{R}^{d_z+1} \to \mathbb{R}^{d_z}$:

$$\mathbf{z}_{t+1} = g_\theta(\mathbf{z}_t,\, c), \tag{15}$$

where the scalar context c is concatenated with the latent state before the first hidden layer, allowing the learned dynamics to vary with the operating regime. At $c = 0$ the model recovers the dynamics averaged over source regimes, analogous to the pooled coefficient matrix Ξ_0 in the SINDy formulation.

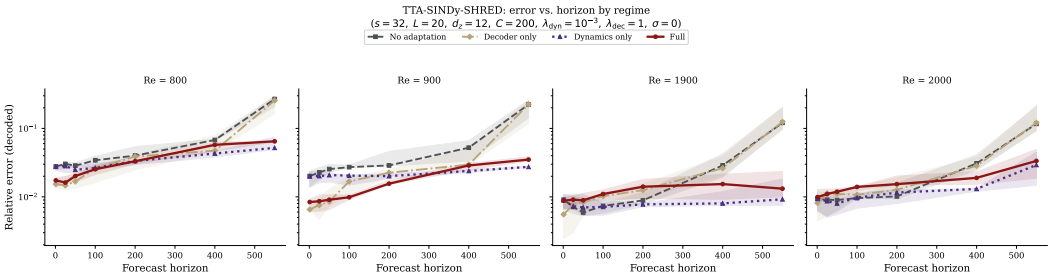


Figure 14: Per-regime ablation horizon curves for TTA-SINDy-SHRED. Each subplot shows one test Reynolds number; curves correspond to the four ablation arms (median over five seeds; shaded bands show the interquartile range). The crossover between decoder-only and dynamics-only depends on the extrapolation direction.

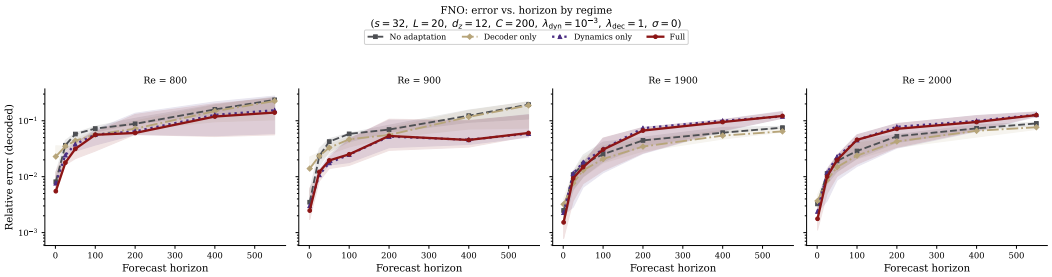


Figure 15: Per-regime ablation horizon curves for FNO (full-field input). Shaded bands show the interquartile range over five seeds.

The ConditionalMLP is trained post-hoc on the frozen representation network's latent trajectories by minimizing the one-step prediction loss

$$\mathcal{L}_{\text{dyn}} = \sum_i \sum_t \left\| g_\theta(\mathbf{z}_t^{(i)},\, c_i) - \mathbf{z}_{t+1}^{(i)} \right\|_2^2 \tag{16}$$

over all source regimes i with context coordinates c_i. After training, the network is frozen; only the scalar c is adapted at test time.

Given the anchor state $\mathbf{z}_1$ produced by the frozen network from the initial lagged sensor sequence in the calibration window, the ConditionalMLP generates the subsequent trajectory $\{\mathbf{z}_t(c)\}_{t=2}^{C}$ autoregressively via Equation (15). The dynamics context is then chosen to minimize the sensor-space calibration residual:

$$c_{\text{dyn}}^* = \arg\min_c \sum_{t=2}^{C} \left\| \mathbf{y}_t - S\big[(W + \Delta W_0)\,\phi(\mathbf{z}_t(c)) + \mathbf{b} + \Delta \mathbf{b}_0\big] \right\|_2^2 + \lambda_{\text{dyn}} \frac{c^2}{\sigma_c^2}, \tag{17}$$

where S is the sampling operator and the prior $\lambda_{\text{dyn}}\, c^2/\sigma_c^2$ regularizes the dynamics context toward the source-regime average. This objective is minimized by grid search followed by Levenberg–Marquardt refinement, following the same procedure as in TTA-SINDy-SHRED. The dynamics model thus constrains the latent rollout to follow a learned dynamical law parameterized by a single scalar, preventing the trajectory from drifting arbitrarily and providing a physics-informed prior on the latent evolution—although without the interpretability of the sparse polynomial structure.

D.4 Results and Discussion

FNO-masked exhibits the highest error among all methods across regimes and horizons (Figure 3), an expected outcome given that spectral convolution relies on dense spatial structure that a masked sparse input could not provide; it illustrates why sparse sensing requires specific architectures designed for sparse measurements. The other four methods, which operate more or less within their intended regime, are analyzed in further detail below.

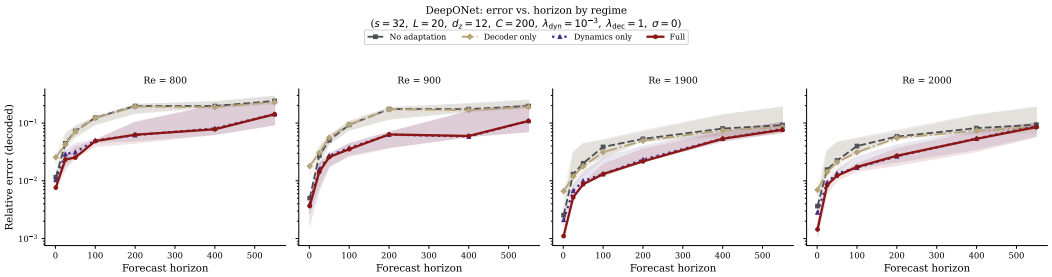


Figure 16: Per-regime ablation horizon curves for DeepONet. Shaded bands show the interquartile range over five seeds.

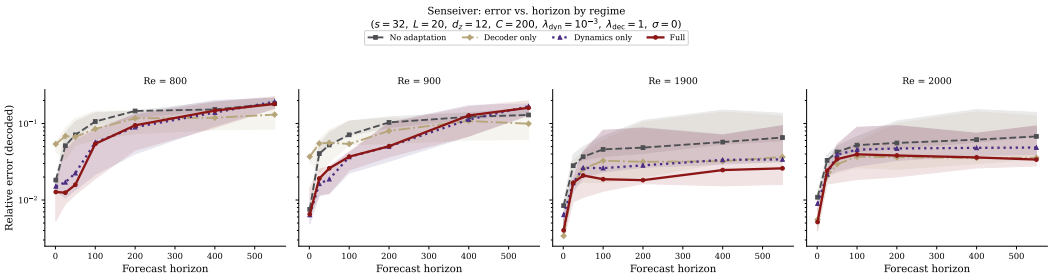


Figure 17: Per-regime ablation horizon curves for Senseiver. Shaded bands show the interquartile range over five seeds.

Between-method comparison. Figure 3 plots the full-adaptation error against forecast horizon for all five methods; Table 2 reports the numerical values. While FNO and DeepONet achieve the lowest single-step error ($H{=}1$), TTA-SINDy-SHRED attains the lowest error at every horizon evaluated, followed by DeepONet, Senseiver, and FNO. All methods perform comparably at short horizons ($H \leq 25$), where the local rollout has not yet departed far from the initial state; beyond $H = 50$, the curves separate markedly, with the baseline errors growing several times faster than that of TTA-SINDy-SHRED. By $H = 550$ the error of TTA-SINDy-SHRED remains below 0.04, while the best-performing baseline (DeepONet) exceeds 0.06 and the remaining methods reach 0.10–0.13. The interquartile range (shaded regions, 25th–75th percentile over seeds and regimes) of TTA-SINDy-SHRED remains narrow across all horizons, whereas the baselines exhibit substantially wider spread at $H \geq 200$. These results indicate that the jointly trained SINDy dynamics under the same test-time adaptation mechanism provide a stronger and more stable prior for long-range latent rollouts than the ConditionalMLP used by the baselines.

Ablation across methods. Figures 14–17 show the per-regime ablation horizon curves for TTA-SINDy-SHRED, FNO, DeepONet, and Senseiver, with one subplot per test Reynolds number ($\mathrm{Re} = 800, 900, 1900, 2000$). Full adaptation yields the lowest error for all methods and regimes, confirming that the two-scalar mechanism transfers across encoder architectures. Beyond this shared pattern, the relative contribution of each context differs between the proposed method and the baselines.

E Extrapolation Far from the Source Range

We extend the Navier–Stokes experiment of Section 4.3 to test regimes up to four source standard deviations from the centroid, $\mathrm{Re} \in \{300, 500, 700, 900\}$ below the source range and $\mathrm{Re} \in \{1900, 2100, 2300, 2500\}$ above it, with the source regimes unchanged at $\mathrm{Re} \in \{1000, 1100, \ldots, 1800\}$. The data, sensors, architecture, families, and inference settings follow Appendix A, with $C = 200$ and $H = 550$. The pooled models are trained separately from those of Section 4, for 200 epochs with thresholding at the final epoch, on the concatenated source trajectories split into contiguous training, validation, and test segments, with seeds 42, 52, and 62. The box is

widened to $c_{\max} = 4.1$ so that it contains every test context. Because the models differ, the errors at $\mathrm{Re} = 900$ and 1900 are not directly comparable with Table 11. Table 13 and Figure 18 report the decoded error in physical units at $H = 550$.

Table 13: Decoded relative error at $H = 550$ for the far-out-of-distribution Navier–Stokes regimes, averaged over the seeds whose rollout remains bounded. Dashes mark regimes where no pooled rollout remains bounded, and * marks averages over fewer than three seeds. The context coordinate $c = (\mathrm{Re} - 1400)/273.9$ and the viscosity ratio $\nu/\bar{\nu} = 1400/\mathrm{Re}$ give the distance from the source centroid.

Re	c	$\nu/\bar{\nu}$	No-adapt	Decoder only	Dynamics only	Full
300	−4.02	4.67	2.027*	1.881*	0.572*	0.675*
500	−3.29	2.80	—	—	0.257	0.263
700	−2.56	2.00	—	—	0.145	0.143
900	−1.83	1.56	0.086*	0.082*	0.096	0.090
1900	1.83	0.74	0.067	0.069	0.061	0.047
2100	2.56	0.67	0.080	0.082	0.056	0.038
2300	3.29	0.61	0.093	0.096	0.060	0.037
2500	4.02	0.56	0.105	0.110	0.067	0.040

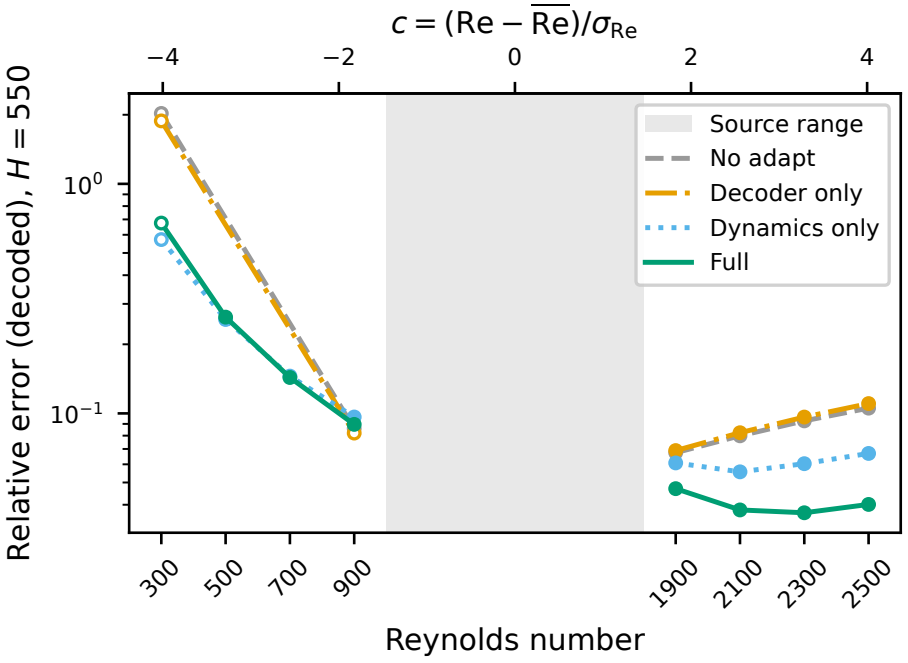


Figure 18: Decoded relative error at $H = 550$ versus Reynolds number for the four arms, averaged over the seeds whose rollout remains bounded. Hollow markers average fewer than three seeds, and an arm has no marker where none of its rollouts remains bounded. The gray band marks the source range, and the top axis gives the context coordinate c.

Above the source range. Full adaptation attains the lowest decoded error at all four Reynolds numbers and improves on No-adapt for every seed at every regime. The pooled error grows steadily with distance from the source range, from 0.067 at $\mathrm{Re} = 1900$ to 0.105 at $\mathrm{Re} = 2500$, while the Full

error stays between 0.037 and 0.047. The latent error of the adapted rollout falls from 0.34 to 0.13 over the same range, against 0.65 to 0.69 for the pooled rollout. The decoder context contributes throughout, with Full improving on Dynamics only by 0.014 to 0.027 in every regime.

Below the source range. Below the source range, the pooled model does not remain bounded over the 550-step horizon at $\mathrm{Re} \in \{500, 700\}$, and Decoder only, which decodes the same latent trajectory, behaves identically. Adapting the dynamics yields bounded forecasts at every test Reynolds number. The error of the adapted forecasts increases smoothly as the Reynolds number decreases, from 0.090 at $\mathrm{Re} = 900$ to 0.143 at $\mathrm{Re} = 700$, 0.263 at $\mathrm{Re} = 500$, and 0.675 at $\mathrm{Re} = 300$, following the growing change in dissipation discussed below. Dynamics only and Full perform comparably in this direction, so the dynamics context accounts for nearly all of the benefit.

Asymmetry of the two directions. The test regimes are symmetric about the source centroid in c but not in viscosity. Because $\nu = 12/\mathrm{Re}$ is convex in Re, the regime at $c = -4.02$ multiplies the dissipation of the source centroid by 4.67, whereas the regime at $c = 4.02$ reduces it by 44%. An affine family in the standardized Reynolds number therefore extrapolates over a much larger change in the governing operator below the source range than above it, which is consistent with the rapid growth of the error at low Re.

Inferred contexts. At these distances from the source range, the dynamics context is placed at the edge of the admissible box, $|c^{\star}_{\mathrm{dyn}}| = c_{\max}$, in nearly all cases, with the sign of the shift matching the direction of extrapolation. The largest available correction of the latent dynamics is thus applied, and the decoder context accounts for the remaining regime dependence. Its mean over seeds increases monotonically with the Reynolds number across all eight regimes, from -10.9 at $\mathrm{Re} = 300$ to 1.81 at $\mathrm{Re} = 2500$, so the decoder correction tracks the direction and magnitude of the parameter shift. As in the main experiments, c_{dyn} is best read as an adaptation coordinate rather than an estimate of the Reynolds number.