
Anatomy of Rollout Error in Chaotic Flow Surrogates

Emma Gaillard^{1*} Armand Kassai Koupai¹ Etienne Meunier² Patrick Gallinari^{1,3}

¹Sorbonne Université, CNRS, ISIR, F-75005 Paris, France

²Sorbonne Université, CNRS, LOCEAN, F-75005 Paris, France

³Criteo AI Lab, Paris, France

Abstract

Autoregressive neural surrogates for turbulent PDEs are typically evaluated by rollout RMSE against a reference trajectory. However, in chaotic flows, trajectories often decorrelate due to sensitivity to initial conditions, making pointwise prediction errors meaningless. We investigate how surrogate errors develop and what aspects of the learned dynamics can still be evaluated beyond this predictability horizon. We compare deterministic and generative surrogates on 2D Kolmogorov flow using a shared backbone, focusing on models whose rollouts remain bounded. Our analysis reveals three findings. First, errors emerge at fine spatial scales before large-scale trajectory decorrelation. Second, despite trajectory divergence, local one-step errors measured against a numerical solver on model-generated states remain approximately constant with rollout depth. Third, one-step prediction accuracy does not necessarily translate into accurate long-term state statistics. This highlights the need to complement trajectory-based evaluation with measures of local dynamical accuracy and long-term statistical fidelity.

1 Introduction

Neural surrogates are increasingly used to solve time-dependent PDEs, offering a cheaper alternative to numerical solvers [Li et al., 2021, Hao et al., 2024]. They are most often trained to predict the next time step from one or several input states [McCabe et al., 2024], on pairs drawn from reference trajectories. At inference the model is deployed autoregressively: it starts from an initial condition ω_0 , or a set of input states, and consumes its own output up to the target horizon [Koupai et al., 2025, Hao et al., 2024]. Teacher forcing keeps training efficient, but it misaligns training and deployment, where the model is no longer conditioned on clean states [Liu et al., 2025].

Two error sources follow from this setting. *Spectral bias*, where ℓ^2 training emphasizes energetic large-scale modes over dynamically important fine-scale structure [Rahaman et al., 2019, Oommen et al., 2025], and *exposure bias*, where conditioning on predicted states progressively drifts trajectories away from the physical manifold [Brandstetter et al., 2023]. On simple systems their effect stays limited and autoregressive surrogates perform well. Many systems in nature are chaotic, and there small errors are amplified exponentially by the underlying dynamics [Strogatz, 2000]: surrogates lose short-term precision and long-range physical coherence alike, eventually reaching numerical blow-up [Pedersen et al., 2025].

Several directions address these failure modes. Regularization penalizes the Jacobian of the learned transition so the model cannot amplify small perturbations without bound [Janvier et al., 2026], while frequency-aware losses prevent the training signal from being dominated by energetic large scales [Chakraborty et al., 2026]. Generative surrogates have proved particularly effective on chaotic systems, learning a distribution of plausible states rather than a point estimate [Kohl et al.,

*Correspondence to: emma.gauillard@isir.upmc.fr

2024, Albergo et al., 2025, Li et al., 2025], refining a deterministic solution [Lippe et al., 2023, Dai et al., 2026b], or projecting drifting states back onto the data manifold, independently of the underlying deterministic model [Pedersen et al., 2025, Liu et al., 2026b]. Physical priors can also be injected, through the training loss [Baldan et al., 2026] or by projection onto the constraint set at sampling [Utkarsh et al., 2025, Li et al., 2026, Blanke et al., 2026].

What these methods share is how they are judged: whether rollouts stay bounded, how a pointwise RMSE behaves against a single reference trajectory, and how well the spatial energy spectrum is recovered. They rarely ask whether generated trajectories still follow the dynamics at long horizon, or how far their distribution has moved from the physical one.

In this work, we investigate how to assess autoregressive neural surrogates once chaotic dynamics make trajectory-wise prediction unreliable. We distinguish three aspects of prediction quality: short-term trajectory fidelity, long-term local dynamical accuracy, and statistical fidelity. Using six deterministic and generative surrogates of 2D Kolmogorov flow as case studies, we identify three main findings:

- **Early decorrelation is scale-dependent.** Errors first affect fine-scale structures and high-frequency phase coherence, before extending to larger spatial scales.
- **Local dynamical accuracy does not deteriorate across rollout steps.** In the bounded regime studied here, one-step errors measured against a numerical solver on model-generated states remain approximately constant across rollouts.
- **Local accuracy does not guarantee long-term statistical fidelity.** The model with the best one-step accuracy does not generate the state distribution closest to the physical one.

2 Anatomy of rollout error

We analyze rollout error in three steps. We first look at short horizons, where the rollout can still be compared to a reference trajectory, and identify the decorrelation horizon t^* in Lyapunov times, beyond which that comparison stops being informative (§2.2). Past t^* , two aspects of the learned dynamics can still be evaluated without trajectory alignment, following the conditional/unconditional decomposition of Liu et al. [2026a]: whether the learned one-step kernel still matches the true dynamics on the states the model actually visits (local dynamical accuracy, §2.3), and whether those states remain distributed like physical ones (long-term statistical fidelity, §2.4).

2.1 Setup

We consider the 2D incompressible Navier–Stokes equations in vorticity form with periodic boundary conditions,

$$\partial_t \omega + \mathbf{u} \cdot \nabla \omega = \nu \Delta \omega + f, \quad \nabla \cdot \mathbf{u} = 0, \quad (1)$$

where ω is the vorticity, $\mathbf{u}$ the velocity, $\nu = 10^{-3}$, and f a fixed sinusoidal forcing (Kolmogorov flow). The state predicted by all models is the vorticity field ω . We train a one-step forecaster $f_\theta(\omega_t) = \hat{\omega}_{t+1}$, where t indexes rollout steps and one step corresponds to $\Delta t = 0.5$ physical time units. At inference the model is applied autoregressively over 200-step rollouts, on which all reported metrics are computed; we additionally run 20 000-step rollouts only to check that trajectories remain bounded (Figure 6).

We study six models, all sharing the same FNO backbone and training data, so that differences in behaviour come from the training objective rather than from the architecture. MSE is a plain ℓ^2 regression baseline. Two variants extend it: MSE+SPECTRAL adds a spectral term to the training loss (Appendix E.1), and MSE+MASS applies a mass-conservation projection at each step (Appendix E.2). The remaining three are generative. FLOW learns a velocity field via conditional flow matching [Lipman et al., 2023]; SI uses stochastic interpolants directly between physical states [Albergo et al., 2025]; and REFINER refines a base ℓ^2 prediction through K denoising steps [Lippe et al., 2023]. Architecture, training settings and algorithms are in Appendices A and D.

2.2 Identifying a decorrelation horizon at early stage

Construction of the chaotic divergence baseline To isolate the sensitivity of chaotic dynamics, we build a reference, the *chaotic divergence baseline*, by comparing two numerical-solver trajectories

initialized from slightly different states. Their separation measures how an initial perturbation is amplified over time by chaos. Starting from an initial condition ω_0 , we apply a uniform Gaussian perturbation with relative amplitude α :

$$\tilde{\omega}_0 = \omega_0 + \varepsilon \xi, \quad \xi \sim \mathcal{N}(0, I), \quad \varepsilon = \alpha \sigma_\omega, \quad (2)$$

where σ_ω is the standard deviation of the vorticity field over the attractor, making α a dimensionless relative amplitude. Each amplitude $\alpha \in \{10^{-3}, 5 \times 10^{-2}\}$ defines its own baseline

$$D_\alpha(t) = \|S^t(\tilde{\omega}_0) - S^t(\omega_0)\|, \quad (3)$$

where S^t denotes the solver advanced by t rollout steps. Since the magnitude of one-step prediction errors of our surrogates lie between $0.11 \sigma_\omega$ and $0.29 \sigma_\omega$, we add $\alpha \in \{0.11, 0.29\}$. We estimate a Lyapunov time $\tau_\lambda = 1.63$ time units (derivation in Appendix A.1).

Comparing chaos and surrogates To see how errors develop across spatial scales over the first 25 rollout steps, we group Fourier modes into low, mid and high frequencies and evaluate two quantities that RMSE cannot show: the amplitude ratio, which measures how accurately the energy of each mode is reproduced (Figure 1, top), and the phase coherence, which measures whether these structures remain aligned with the reference (Figure 1, bottom).

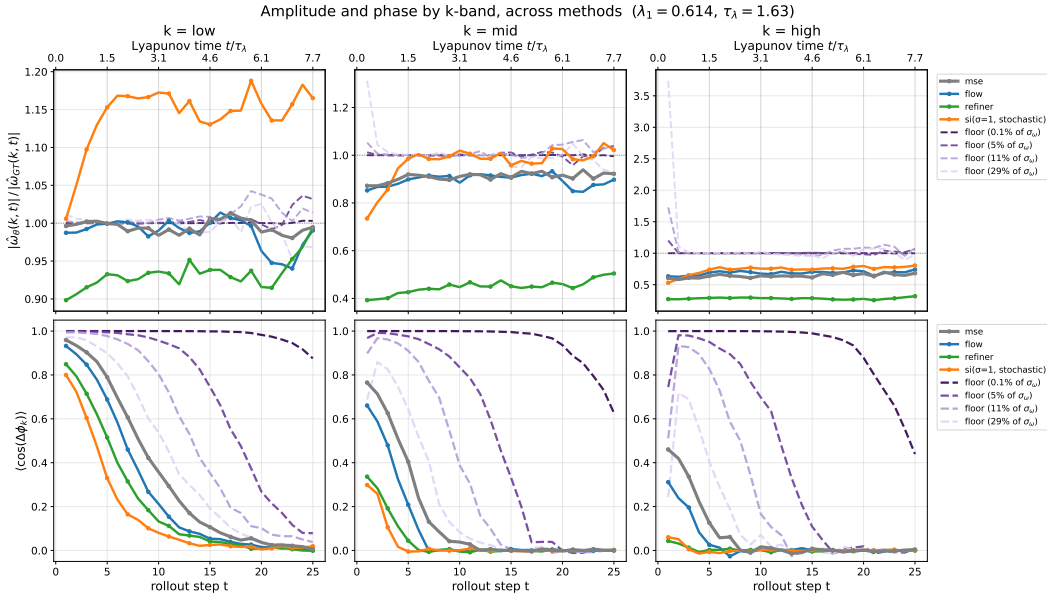


Figure 1: Fourier amplitude and phase accuracy over the first 25 rollout steps, averaged over the modes of each frequency band (columns, up to the 2/3 dealiasing cutoff) and over trajectories. The reference is $\omega_{GT}(t) = S^t(\omega_0)$. **Top:** ratio of predicted to reference Fourier amplitudes, $|\hat{\omega}_\theta(k, t)|/|\hat{\omega}_{GT}(k, t)|$; 1 indicates agreement in amplitude. **Bottom:** phase coherence $\langle \cos \Delta\phi_k(t) \rangle$, with $\Delta\phi_k = \arg \hat{\omega}_\theta(k, t) - \arg \hat{\omega}_{GT}(k, t)$; 1 indicates perfect alignment, 0 decorrelation. Solid: neural surrogates. Dashed: chaotic divergence baselines D_α , the same quantities computed between $S^t(\tilde{\omega}_0)$ and $S^t(\omega_0)$, one curve per α .

Across all models, high-frequency fourier mode lose phase coherence from the very first steps. The low-frequency modes phases decorrelate later. All methods also systematically under-predict high- k energy amplitudes, consistent with the well-known spectral bias of neural networks [Rahaman et al., 2019]. Thus, surrogate errors are already present at fine spatial scales before the trajectories lose the coarse scales. We define the decorrelation horizon t^* as the first rollout step at which the phase coherence $\langle \cos(\Delta\phi_k) \rangle$ has dropped to zero in all three frequency modes. We identify $t^* \approx 20$ rollout steps for all models, i.e. $6.1 \tau_\lambda$, earlier than for the chaotic divergence baselines.

What kind of perturbation is relevant? We chose an isotropic Gaussian perturbation for the chaotic divergence baselines. When matching the magnitude of the surrogates' first-step errors

(0.11 and $0.29 \sigma_\omega$), the phase coherence first decreases, then partially recovers, and decreases again (Figure 1). This can be explained by the spectrum of the perturbation: Gaussian noise carries much more high-frequency energy than the physical spectrum, so a large amount of small-scale energy is injected at the first step. The solver dissipates this energy and restores a physical spectrum, and then chaos decorrelates the trajectories. Even at matched error magnitude, the surrogates decorrelate faster than the solver. However, surrogate errors are not isotropic but concentrated at specific spatial scales (Figure 1). We therefore ask whether the solver amplifies perturbations faster along the directions of the surrogate errors (Appendix C.1).

2.3 Does the model still reproduce the local dynamics after decorrelation ?

After the decorrelation horizon, RMSE no longer measures whether the model follows the correct dynamics. We therefore ask whether the model still reproduces the local dynamics on the states it produces at long horizon. At different rollout steps, we initialize both the surrogate and the numerical solver from the same predicted state and compare their one-step outputs. This protocol measures whether exposure to model-generated states leads one-step dynamical error to increase over time.

Solver injection protocol We measure the pointwise one-step error $E_{\text{dyn}}(\omega) = \|f_\theta(\omega) - S(\omega)\|$, where S advances the state by the same $\Delta t = 0.5$ units the models were trained to predict. For generative framework $f_\theta(\omega)$ is a single sample, since samples drawn from a fixed context are nearly identical (Appendix C.3), this choice has little effect on E_{dyn} .

We evaluate E_{dyn} on states $\omega_t \sim \mu_t$, obtained by advancing independent initial conditions t steps with the solver (Figure 2, left), and on states $\hat{\omega}_t \sim \hat{\mu}_t$, the model’s rollout distribution at step t , obtained by applying f_θ autoregressively t times from the same initial conditions (Figure 2, right):

$$E_{\text{dyn}}^\mu(t) = \mathbb{E}_{u_t \sim \mu_t}[E_{\text{dyn}}(u_t)], \quad E_{\text{dyn}}^{\hat{\mu}}(t) = \mathbb{E}_{\hat{u}_t \sim \hat{\mu}_t}[E_{\text{dyn}}(\hat{u}_t)]. \quad (4)$$

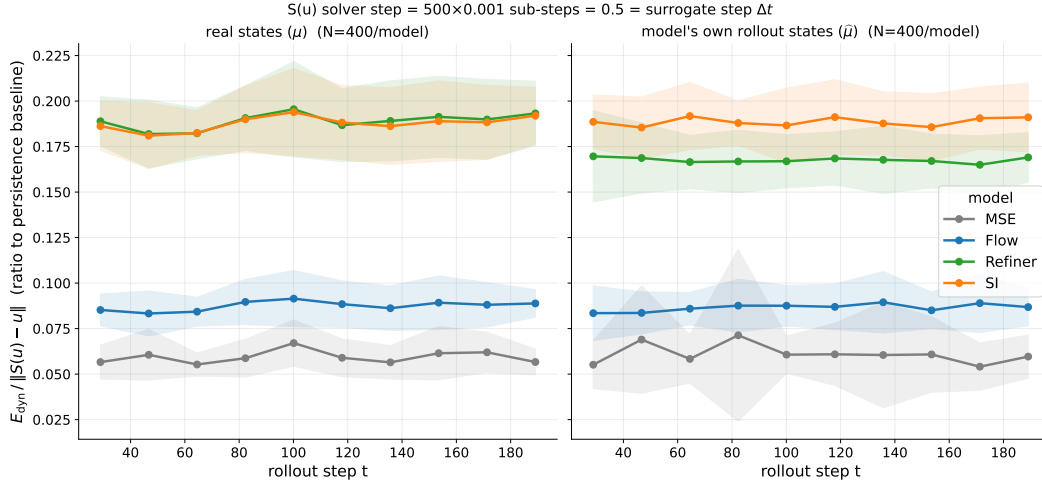


Figure 2: One-step dynamical error, normalized by the persistence baseline $\|S(\omega) - \omega\|$, for states drawn from the true measure μ_t (left) and the model’s own rollout distribution $\hat{\mu}_t$ (right, $t \in [20, 190]$). We compute this error for 400 samples per model and distribution, drawn uniformly at random from all (trajectory, ensemble member, rollout step) triplets in $t \in [20, 190]$. For generative models, ensemble members diverge after a few tens of steps and thus provide distinct samples of $\hat{\mu}_t$. Each plotted point averages the samples falling in one of 10 equally-spaced time bins.

Across all models, $E_{\text{dyn}}^{\hat{\mu}}(t)$ stays approximately flat with rollout depth (Figure 2) and is comparable in magnitude to E_{dyn}^μ : the number of autoregressive steps taken before a state is reached does not affect the one-step dynamical error at that state. In the bounded-rollout regime considered here, **the local dynamical error does not degrade with rollout depth, on the states the model actually visits.**

A possible explanation is that the models settle into a statistically stationary regime: their 20 000-step

rollouts remain bounded (Figure 6), and the one-step error would then be stable on the set of states they visit; we examine this in the next section.

2.4 Are generated states drawn from the true invariant measure?

Our 2D Kolmogorov system is dissipative: past an initial transient (discarded at dataset creation), trajectories converge to a compact attractor [Temam, 1997], which supports an invariant measure μ , a distribution of states preserved by the dynamics. Assuming ergodicity, μ governs the long-term statistics of the system and can be estimated from a finite number of solver trajectories and steps (formal definitions in Appendix A.1). A probability measure μ is invariant for the flow if

$$S_{\#}^k \mu = \mu, \quad \text{i.e.} \quad \mu(S^{-k}(A)) = \mu(A) \quad \text{for all Borel sets } A,$$

where S^k denotes the k -th iterate of the solver map S , $S_{\#}^k \mu$ denotes the pushforward of μ under S^k .

On the model side, we check whether $\hat{\mu}_t$ has become stationary, i.e. whether the distribution of predicted states doesn’t change over rollout time. The MMD between different time windows of the rollout remains flat for $t \geq 100$ across all models (Appendix C.2). We can therefore pool $\hat{\mu}_t$ over $t \geq 100$ into a single distribution $\hat{\mu}$, an empirical estimate of the model’s invariant measure. We compare it to the physical invariant measure μ with the Maximum Mean Discrepancy (MMD), using the kernel described in Appendix B, on $N = 1000$ samples. This measures whether the predicted states stay close to the physical distribution.

Table 1: Metrics over the full 200-step rollout, split by regime. [†]Degenerate for deterministic models, whose ensemble has zero spread. Details on metrics are given in Appendix B

Method	Short horizon (conditional)				Long horizon (unconditional)		
	RMSE _{t₁} ↓	CRPS ↓	RMSCE [†] ↓	SSR [†] → 1	TKED ↓	PSDD ↓	MMD ↓
MSE	0.1079	1.000	0.570	0.000	0.0069	4.718	0.203
MSE+MASS	0.1079	0.996	0.570	0.000	0.0072	4.424	0.209
MSE+SPECTRAL	0.1313	1.011	0.570	0.000	0.0062	5.044	0.467
FLOW	0.1460	0.506	0.265	0.576	0.0053	4.305	0.188
REFINER	0.2233	0.534	0.291	0.527	0.0055	5.315	1.098
SI	0.2759	0.516	0.254	0.592	0.0076	4.658	0.341

MSE has the lowest E_{dyn} of all models (Figure 2), so the best local dynamical accuracy. But it already shows vortex artifacts at short horizon (Figure 5). On long-horizon metrics, FLOW has the best TKED and PSDD (Table 1): its energy and its spectral shape stay closest to the true dynamics. Its MMD is also the lowest, but only barely – MSE is close behind however their 95% confidence intervals do not overlap. Generative models learn a conditional distribution $p_{\theta}(u_{t+1} | u_t)$, which may help them track distributions statistics over long rollouts better than direct regression.

3 Conclusion

In this work, we evaluate autoregressive neural surrogates of chaotic PDEs in two steps. At short horizon, we locate a decorrelation horizon [Boffetta, 2002] and compare the models against chaos floors. Past that horizon, we separate the local dynamics on the states the model visits from the distance between the generated and physical state distributions. On 2D Kolmogorov flow, the surrogate models remain bounded, settle into a statistically stationary regime, and their local dynamical error does not degrade with rollout depth. However, the model with the best one-step accuracy does not reproduce the physical state distribution most faithfully: trajectory error alone is insufficient to evaluate a chaotic surrogate.

Future work should investigate methods that explicitly target the invariant measure, through training regularization [Schiff et al., 2024, Jiang et al., 2024, Melo et al., 2026] or inference correctors [Pedersen et al., 2025, Liu et al., 2026b], and extend this analysis to models that blow up, freeze, or drift, the three failure modes described by Dai et al. [2026a].

Acknowledgments and Disclosure of Funding

We acknowledge the financial support provided by DEEPNUM (ANR-21-CE23-0017-02), PHLUSIM (ANR-23-CE23-0025-02), MELISSA (ANR-24-CE23-7140-01), PEPR Sharp (ANR-23-PEIA0008, ANR, FRANCE 2030), and cluster IA PostGenAI@Paris (ANR-23-IACL-0007).

This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grants AD010617285 and 2025-A0191016890 on the supercomputer Jean Zay’s V100/A100/H100 partitions.

References

- Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2025. URL <https://arxiv.org/abs/2303.08797>.
- Giacomo Baldan, Qiang Liu, Alberto Guardone, and Nils Thuerey. Physics vs distributions: Pareto optimal flow matching with physics constraints, 2026. URL <https://arxiv.org/abs/2506.08604>.
- Matthieu Blanke, Yongquan Qu, Sara Shamekh, and Pierre Gentine. Strictly Constrained Generative Modeling via Split Augmented Langevin Sampling, March 2026.
- G Boffetta. Predictability: a way to characterize complexity. *Physics Reports*, 356(6):367–474, January 2002. ISSN 0370-1573. doi: 10.1016/s0370-1573(01)00025-4. URL [http://dx.doi.org/10.1016/S0370-1573\(01\)00025-4](http://dx.doi.org/10.1016/S0370-1573(01)00025-4).
- Johannes Brandstetter, Daniel Worrall, and Max Welling. Message Passing Neural PDE Solvers, March 2023.
- Dibyajyoti Chakraborty, Arvind T. Mohan, and Romit Maulik. Binned spectral power loss for improved prediction of chaotic systems, 2026. URL <https://arxiv.org/abs/2502.00472>.
- Yilong Dai, Shaswata Mitra, Raj Patel, Yiming Sun, Shengyu Chen, Jiaqi Gong, Sudip Mittal, Shahram Rahimi, Xiaowei Jia, and Runlong Yu. Staying on the attractor: Supervising neural surrogates of 3d turbulence where they leave it, 2026a. URL <https://arxiv.org/abs/2609.32864>.
- Yilong Dai, Yiming Sun, Yiheng Chen, Shengyu Chen, Xiaowei Jia, and Runlong Yu. FlowRefiner: Flow Matching-Based Iterative Refinement for 3D Turbulent Flow Simulation, 2026b. URL <https://arxiv.org/abs/2604.17149>.
- Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. Dpot: Auto-regressive denoising operator transformer for large-scale pde pre-training. *41th International Conference on Machine Learning (ICML 2024)*, 2024.
- Maya Janvier, Julien Salomon, and Etienne Meunier. Jacobian regularization stabilizes long-term integration of neural differential equations, 2026. URL <https://arxiv.org/abs/2602.04608>.
- Ruoxi Jiang, Peter Y. Lu, Elena Orlova, and Rebecca Willett. Training neural operators to preserve invariant measures of chaotic attractors, 2024. URL <http://arxiv.org/abs/2306.01187>.
- Georg Kohl, Li-Wei Chen, and Nils Thuerey. Benchmarking Autoregressive Conditional Diffusion Models for Turbulent Flow Simulation, December 2024.
- Armand Kassaï Koupaï, Lise Le Boudec, Louis Serrano, and Patrick Gallinari. ENMA: Tokenwise autoregression for continuous neural PDE operators. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=3CYXSMFv55>.
- Xigui Li, Hongwei Zhang, Ruoxi Jiang, Deshu Chen, Chensen Lin, Limei Han, Yuan Qi, Xin Guo, and Yuan Cheng. Project and Generate: Divergence-Free Neural Operators for Incompressible Flows, 2026. URL <http://arxiv.org/abs/2603.24500>.

- Zijie Li, Anthony Zhou, and Amir Barati Farimani. Generative latent neural pde solver using flow matching. *arXiv preprint arXiv:2503.22600*, 2025. URL <http://arxiv.org/abs/2503.22600>.
- Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations, 2021. URL <https://arxiv.org/abs/2010.08895>.
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL <https://arxiv.org/abs/2210.02747>.
- Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving Accurate Long Rollouts with Neural PDE Solvers, 2023. URL <http://arxiv.org/abs/2308.05732>.
- Qi Liu, Laure Zanna, and Joan Bruna. Towards infinitely long neural simulations: Self-refining neural surrogate models for dynamical systems, 2026a. URL <https://arxiv.org/abs/2603.17750>.
- Qi Liu, Laure Zanna, and Joan Bruna. Towards Infinitely Long Neural Simulations: Self-Refining Neural Surrogate Models for Dynamical Systems, 2026b. URL <http://arxiv.org/abs/2603.17750>.
- Yuxuan Liu, Jingmin Sun, and Hayden Schaeffer. BCAT: A block causal transformer for pde foundation models for fluid dynamics. *arXiv preprint arXiv:2501.18972*, 2025.
- Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple physics pretraining for spatiotemporal surrogate models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=DKSI3bULiZ>.
- Gabriel Melo, Leonardo Santiago, and Peter Y. Lu. Learning to Emulate Chaos: Adversarial Optimal Transport Regularization, 2026. URL <http://arxiv.org/abs/2604.21097>.
- Vivek Oommen, Siavash Khodakarami, Aniruddha Bora, Zhicheng Wang, and George Em Karniadakis. Learning Turbulent Flows with Generative Models: Super-resolution, Forecasting, and Sparse Flow Reconstruction, September 2025.
- Chris Pedersen, Laure Zanna, and Joan Bruna. Thermalizer: Stable autoregressive neural emulation of spatiotemporal chaos, July 2025.
- Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks, 2019. URL <https://arxiv.org/abs/1806.08734>.
- Yair Schiff, Zhong Yi Wan, Jeffrey B. Parker, Stephan Hoyer, Volodymyr Kuleshov, Fei Sha, and Leonardo Zepeda-Núñez. DySLIM: Dynamics Stable Learning by Invariant Measure for Chaotic Systems, 2024. URL <http://arxiv.org/abs/2402.04467>.
- Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview Press, 1 edition, 2000.
- Roger Temam. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, volume 68 of *Applied Mathematical Sciences*. Springer, 1997.
- Utkarsh Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gomez-Bombarelli, and Christopher Vincent Rackauckas. Physics-Constrained Flow Matching: Sampling Generative Models with Hard Constraints, November 2025.

A Experimental setup

Table 2: Dataset: forced Kolmogorov flow (2D vorticity, pseudo-spectral solver `exponax`).

Domain	$[0, 2\pi]^2$, periodic
Resolution	256×256 (2/3 dealiasing)
Physical channel	vorticity ω (1 channel)
Viscosity ν	10^{-3}
Linear drag	-0.1
Forcing	sinusoidal, mode $k_{\text{inj}} = 4$, amplitude $\gamma = 1.0$
Reynolds number	$Re = \sqrt{\gamma}/\nu = 1000$
Solver step dt_{solver}	10^{-3}
Subsampling	1 snapshot every 500 solver steps
Effective step dt_{snap}	0.5 (time units)
Warm-up to stationary turbulence	7 000 solver steps
Simulated trajectories	350
Train / test trajectory split	280 / 70 (80/20, at generation time)
Train / val sub-split	252 / 28 (90/10 at training)
Train / test window	$T = 200$ steps each
Normalisation	$(\omega - \mu_{\text{train}})/\sigma_{\text{train}}$, statistics computed on the training set

Table 3: FNO backbone hyperparameters (`neuralop.models.FNO`).

Retained Fourier modes	64×64
Hidden channels	128
Temporal context T_{ctx}	1 frame
Input channels (conditional methods)	$C(T_{\text{ctx}} + 1) + 4 = 6$ (context, noised state + time embedding)
Input channels (mse)	$C \cdot T_{\text{ctx}} = 1$
Output channels	$C = 1$
Diffusion/flow time conditioning	sinusoidal embedding $\rightarrow$ MLP $\rightarrow$ 4 spatial channels

Table 4: Training and sampling details per method (FNO backbone).

Method	lr	epochs	batch	EMA	precision	NFE / sampling step
mse	10^{-4}	200	64	—	fp32	1 (not iterative)
flow	10^{-4}	200	64	—	fp32	50 (Euler, cosine schedule)
si	10^{-4}	200	64	—	fp32	20 (Euler/EM, linear grid)
refiner	10^{-4}	200	64	—	fp32	8 (PDERefiner, $k = 0 \dots 3$)

A.1 Chaotic dynamics: attractor, invariant measure, and ergodicity

Many chaotic systems, including 2D Kolmogorov flow, are dissipative, which implies the existence of a compact set often called an attractor $\mathcal{A}$, towards which all bounded sets of initial conditions converge in finite time. In addition, many of these systems also empirically exhibit **ergodicity**. The main consequence of this property is the existence of an **invariant measure** supported on the attractor, which remains unchanged by the dynamics and entirely captures the long-term statistical behavior of the system [Schiff et al., 2024].

Setting. We work on the state space $\mathcal{U}$ of vorticity fields, a separable metric space equipped with the distance d induced by the L^2 norm and with its Borel σ -algebra. $S : \mathcal{U} \rightarrow \mathcal{U}$ denotes one step of the reference solver, advancing the state by the snapshot interval Δt , and S^t its t -fold composition. For a measure μ on $\mathcal{U}$, $S_{\#}\mu$ is the pushforward of μ by S , defined by $S_{\#}\mu(A) = \mu(S^{-1}(A))$ for every Borel set A .

Attractor. A compact, invariant set $\mathcal{A} \subset \mathcal{U}$ is an attractor if it possesses a basin of attraction—an open neighbourhood $\mathcal{B}$ such that $\text{dist}(S^t(\omega_0), \mathcal{A}) \rightarrow 0$ as $t \rightarrow \infty$ for all $\omega_0 \in \mathcal{B}$ —and is strictly minimal among all closed sets satisfying this property.

Invariant measure. A probability measure μ is invariant for the dynamics S if it is preserved under the dynamics, meaning:

$$S_{\#}\mu = \mu, \quad \text{i.e.} \quad \mu(S^{-1}(A)) = \mu(A) \quad \text{for all Borel sets } A. \quad (5)$$

By definition, a measure μ is invariant relative to S if and only if $\mathbb{E}_{\mu}[g \circ S] = \mathbb{E}_{\mu}[g]$ for all integrable observables g . We call the expectation $\mathbb{E}_{\mu}[g] = \int g d\mu$ an *invariant statistic*.

Ergodicity. We further assume the dynamics to be ergodic: a typical trajectory visits every region of the attractor in proportion to its measure. By Birkhoff's theorem, for any observable $g \in L^1(\mu)$ and μ -almost every ω_0 ,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} g(S^t(\omega_0)) = \int_{\mathcal{U}} g d\mu, \quad (6)$$

which is what allows μ , and any statistic derived from it, to be estimated from a single sufficiently long trajectory.

Lyapunov exponent. Let $\delta(t) = S^t(\omega_0 + \delta_0) - S^t(\omega_0)$ be the separation at time t between two trajectories started a small δ_0 apart, with $\delta_0 = \varepsilon \xi$ as in Eq. (2) — the quantity the chaos floor measures. Then

$$\lambda_1 = \lim_{t \rightarrow \infty} \lim_{\|\delta_0\| \rightarrow 0} \frac{1}{t} \log \frac{\|\delta(t)\|}{\|\delta_0\|}, \quad (7)$$

with t expressed in physical time units and $\tau_{\lambda} = 1/\lambda_1$ setting the predictability horizon [Boffetta, 2002]. The exponential growth it describes is a transient linear regime: once the separation reaches the diameter of $\mathcal{A}$, two trajectories on the attractor cannot separate further.

Computation of the Lyapunov exponent in §2.2. In practice $t \rightarrow \infty$ and $\|\delta_0\| \rightarrow 0$ are not accessible: $D_{\alpha}(t)$ grows exponentially only for a finite window, bounded above by saturation (once $\|\delta(t)\|$ reaches the diameter of $\mathcal{A}$). We estimate λ_1 as the slope of $\log D_{\alpha}(t)$ against physical time t , fit by linear least squares over the points where $D_{\alpha}(t)$ lies between 5% and 60% of its own maximum. We use D_{10-3} , the smallest α , since it stays closest to the $\|\delta_0\| \rightarrow 0$ linear regime for longest. This gives $\lambda_1 = 0.614$ ($R^2 = 0.996$) and $\tau_{\lambda} = 1/\lambda_1 = 1.63$ time units.

B Metrics

All metrics are computed over the 200-step rollout on 15 held-out trajectories (ensemble size $M = 3$), on the normalised vorticity field. Short-horizon metrics are read at the step indicated; long-horizon metrics are averaged over $t \geq 50$, well past the decorrelation horizon. Deterministic models are treated as a degenerate M -member ensemble with identical members, which is why their calibration and spread entries are constant.

RMSE _{t_1} (one-step error). Root-mean-square error of the first predicted frame averaged over ensemble members and initial conditions,

$$\text{RMSE}_{t_1} = \left\langle \sqrt{\frac{1}{N} \sum_{\text{px}} (\hat{\omega}_1 - \omega_1)^2} \right\rangle. \quad (8)$$

Since the field is normalised to unit variance, this already reads as a fraction of σ_{ω} : it reaches 1 against the zero field and $\sqrt{2}$ for two decorrelated fields of matching variance.

CRPS. Energy-form CRPS, pointwise and averaged over space and the rollout window:

$$\text{CRPS} = \frac{1}{M} \sum_{m=1}^M |\hat{\omega}^{(m)} - \omega| - \frac{1}{M(M-1)} \sum_{m \neq m'} |\hat{\omega}^{(m)} - \hat{\omega}^{(m')}|. \quad (9)$$

The spread term penalises overconfidence; for M identical (deterministic) members it vanishes and CRPS reduces to the MAE.

RMSCE. Calibration of the ensemble’s central intervals. For nominal coverage levels $\{\alpha_i\}$ (19 evenly-spaced values in $(0, 1)$), let $\hat{\alpha}_i$ be the empirical fraction of points where the reference falls inside the ensemble’s central α_i -interval $[q_{(1-\alpha_i)/2}, q_{(1+\alpha_i)/2}]$. Then

$$\text{RMSCE} = \sqrt{\frac{1}{19} \sum_i (\hat{\alpha}_i - \alpha_i)^2}. \quad (10)$$

A degenerate ensemble covers every interval with probability 0 or 1 regardless of α_i , giving the same constant score for all deterministic models.

SSR (spread–skill ratio).

$$\text{SSR} = \frac{\sqrt{\text{Var}_m(\hat{\omega}^{(m)})}}{\langle \text{RMSE}(\hat{\omega}^{(m)}, \omega) \rangle_m}, \quad (11)$$

spread and per-member skill both averaged over space and time. $\text{SSR} \rightarrow 1$ is well-calibrated; below 1 overconfident, above 1 overdispersed. Deterministic models have zero spread, so $\text{SSR} = 0$ by construction.

TKED. Absolute kinetic-energy drift from the reference, averaged over the evaluation window:

$$\text{TKED} = \langle |E_{\text{pred}}(t) - E_{\text{ref}}(t)| \rangle_t, \quad E = \frac{1}{2} \|u\|_2^2, \quad (12)$$

u recovered from the vorticity ω by the Biot–Savart law. Flags systematic energy gain or loss, independent of the admissible band.

PSDD. Log-spectral RMS distance to the reference spectrum *at the same rollout step* (not time-averaged), over resolved shells $\mathcal{K}$:

$$\text{PSDD}(t) = \sqrt{\frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} (\log E_{\text{pred}}(k, t) - \log E_{\text{ref}}(k, t))^2}, \quad (13)$$

averaged over t . Unlike TKED, sensitive to how energy is distributed across scales, not just its total.

MMD. Distance between the model’s rollout distribution and the true invariant measure, NOT on the raw state (curse of dimensionality on a 256×256 field swamps the kernel) but on a compact feature map $\phi(u)$: the radial amplitude spectrum up to the dealiasing cutoff, plus enstrophy, z -scored by the reference pool’s own per-feature std. Kernel: RBF, $\kappa(u, v) = \exp(-\|\phi(u) - \phi(v)\|_2^2 / 2\sigma^2)$, bandwidth σ set by the median heuristic on the pooled features. Unbiased estimator (self-pairs $i = j$ excluded):

$$\widehat{\text{MMD}}^2(\mu, \hat{\mu}) = \frac{1}{m(m-1)} \sum_{i \neq j} \kappa(u^{(i)}, u^{(j)}) + \frac{1}{n(n-1)} \sum_{i \neq j} \kappa(v^{(i)}, v^{(j)}) - \frac{2}{mn} \sum_{i,j} \kappa(u^{(i)}, v^{(j)}), \quad (14)$$

$\{u^{(i)}\} \sim \mu$, $\{v^{(i)}\} \sim \hat{\mu}$. We also tried a biased mixture-of-IMQ kernel [Schiff et al., 2024] on the same features: it discriminates too, but its bias from including self-pairs depends on sample size which leads to unfair evaluation between deterministic and generative models –we kept the unbiased one. Reducing to spectral shape rather than pixels is what makes it sensitive to the high- k bias in Section 2.2.

C Further analysis

C.1 Chaos baselines with model-error perturbations

To test whether the solver amplifies perturbations faster along the directions of the surrogate errors, we replace the random perturbation with each surrogate’s own one-step error $f_\theta(\omega_0) - S(\omega_0)$, used as-is (no rescaling). We evolve the perturbed and unperturbed states with the solver, and compare against an isotropic Gaussian perturbation D_α of the same norm.

Solver trajectories perturbed by the model error decorrelate faster than those perturbed by an isotropic Gaussian of the same norm (Figure 3, dashed curves): the dynamics are less stable along the directions of the surrogate errors than along random directions. Part of this difference may also come from the

Gaussian perturbation spreading energy into high frequencies, which the solver dissipates before it can be amplified.

The surrogate’s actual rollout (solid) decorrelates faster still than its own model-error chaos floor, they keep introducing error at every subsequent step.

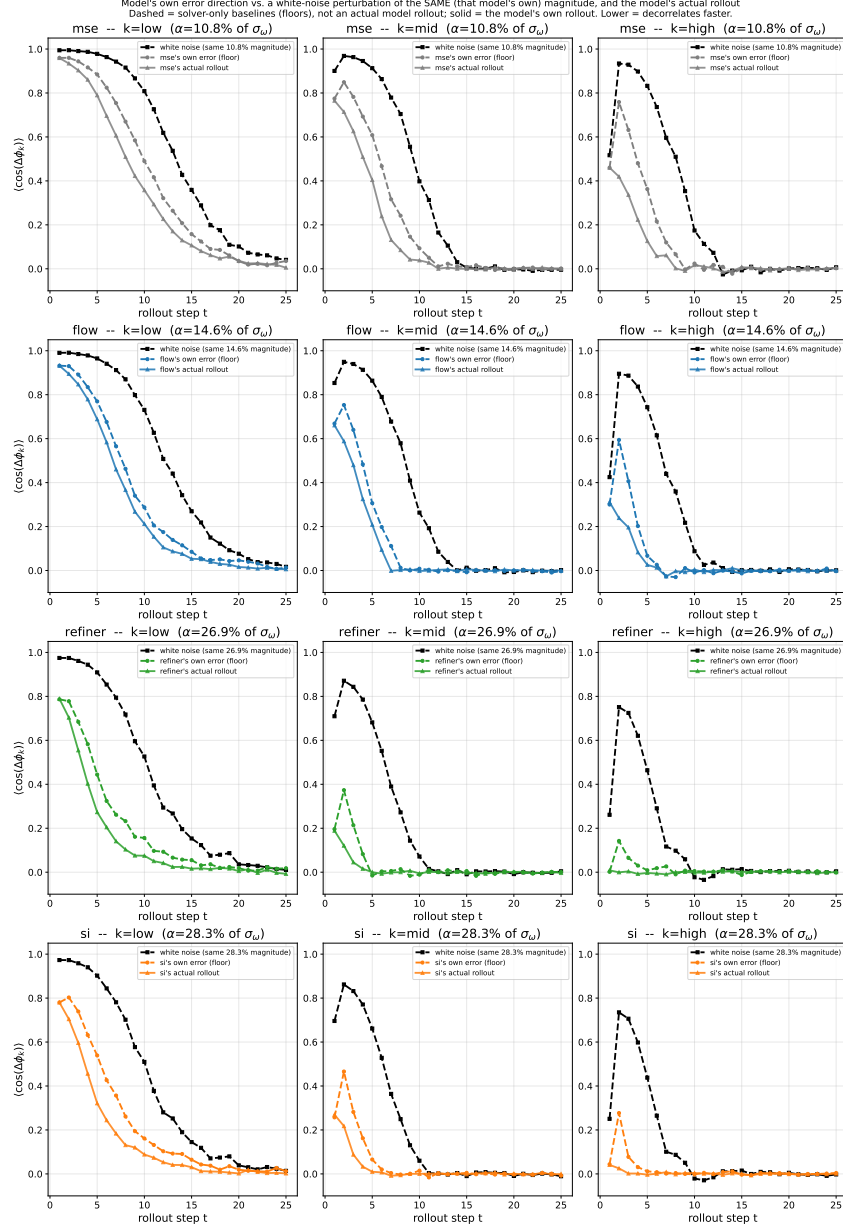


Figure 3: Dashed: solver-only baselines, perturbed either by the surrogate’s own one-step error (colored) or by an isotropic Gaussian of the same norm (black). Solid: the surrogate’s own actual rollout, for comparison. Lower curves decorrelate faster.

C.2 Stationarity of the model rollout distribution

The results of Section 2.3 raise the question of whether the distribution of rolled-out states settles into a statistically stationary regime. To check this, we compute the MMD between different time windows of $\hat{\mu}_t$, to see whether its properties change with rollout time. We split the interval $t \geq 100$

into 6 windows of about 17 steps. For each model, we compute the MMD^2 between the states of each window and the full pooled sample, with the same kernel as in the main text.

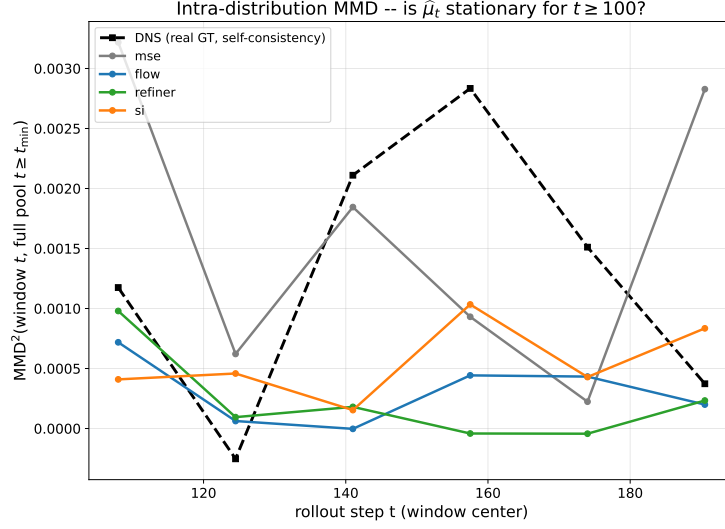


Figure 4: Intra-distribution MMD^2 between the rollout states in each time window (x-axis: window center) and the full pooled sample over $t \geq 100$, per model. Dashed: the same quantity for solver trajectories (DNS), giving the finite-sample fluctuation level.

For all models, the intra-distribution MMD fluctuates without a clear trend across windows, at a level comparable to the DNS reference’s own noise (Figure 4). These fluctuations ($\lesssim 0.003$) are two to three orders of magnitude below the model-to- μ MMD in Table 1 (0.2–1.1) – negligible next to what we are trying to measure. All windows are at a similar distance from the pooled distribution. We therefore pool $\hat{\mu}_t$ over $t \geq 100$ in Section 2.4: each model has reached its own stationary regime, and its distance to μ is reported in Table 1. MSE fluctuates more than the generative models: its ensemble members are identical copies (one distinct state per trajectory, vs. three for the generative models), so each window is noisier over only 6 windows. Only the absence of a trend matters here, not the fluctuation size.

C.3 Ensemble behaviour of the generative methods

Figure 2 does not show the ensemble behaviour of the generative methods. For a fixed context, sampled from either μ or $\hat{\mu}$, their members are nearly identical: the learned kernel is close to deterministic, so E_{dyn} is essentially the same for any draw. The members nevertheless end up completely different after a few tens of steps, since the chaotic flow amplifies whatever small spread they start with.

D Training and inference algorithms

Algorithm 1 Conditional Flow Matching — Training

Require: Context ω_t , target ω_{t+1} , model f_θ

- 1: Sample $\tau \sim \mathcal{U}(0, 1)$, $\varepsilon \sim \mathcal{N}(0, I)$
 - 2: $x_\tau \leftarrow (1 - \tau)\varepsilon + \tau\omega_{t+1}$
 - 3: $v^* \leftarrow \omega_{t+1} - \varepsilon$
 - 4: $\mathcal{L} \leftarrow \|f_\theta([\omega_t, x_\tau], \tau) - v^*\|^2$
-

Algorithm 2 Conditional Flow Matching — Inference

Require: Context ω_t , model f_θ , steps N

- 1: $x \leftarrow \varepsilon \sim \mathcal{N}(0, I)$
 - 2: **for** $i = 0$ **to** $N - 1$ **do**
 - 3: $\tau \leftarrow i/N$
 - 4: $x \leftarrow x + \frac{1}{N} f_\theta([\omega_t, x], \tau)$ ▷ Euler step
 - 5: **end for**
 - 6: **return** x
-

Algorithm 3 PDE-Refiner — Training

Require: Context ω_{in} , target ω_{t+1} , model f_θ , steps K , $\sigma_{\min}$

- 1: Sample $k \sim \mathcal{U}\{0, \dots, K - 1\}$, $\varepsilon \sim \mathcal{N}(0, I)$
 - 2: $\sigma_k \leftarrow \sigma_{\min}^{k/K}$
 - 3: **if** $k = 0$ **then**
 - 4: $x_k \leftarrow \mathbf{0}$, $y^* \leftarrow \omega_{t+1}$
 - 5: **else**
 - 6: $x_k \leftarrow \omega_{t+1} + \sigma_k \varepsilon$, $y^* \leftarrow \varepsilon$
 - 7: **end if**
 - 8: $\mathcal{L} \leftarrow \|f_\theta([\omega_{\text{in}}, x_k], k) - y^*\|^2$
-

Algorithm 4 PDE-Refiner — Inference

Require: Context ω_{in} , model f_θ , schedule $\{\sigma_k\}_{k=1}^K$

- 1: $\hat{\omega} \leftarrow f_\theta([\omega_{\text{in}}, \mathbf{0}], 0)$
 - 2: **for** $k = 1$ **to** K **do**
 - 3: $\varepsilon \sim \mathcal{N}(0, I)$
 - 4: $z \leftarrow \hat{\omega} + \sigma_k \varepsilon$
 - 5: $\hat{\omega} \leftarrow z - \sigma_k f_\theta([\omega_{\text{in}}, z], k)$
 - 6: **end for**
 - 7: **return** $\hat{\omega}$
-

E Physical and spectral losses

E.1 Spectral loss

Pixel-wise MSE weights every Fourier mode equally, but the high-wavenumber modes carry little energy, so errors at high frequencies barely affect the loss and the small scales end up under-resolved. We therefore add a term matching Fourier amplitudes:

$$\mathcal{L} = \frac{1}{N_x N_y} \|\tilde{\omega} - \omega\|_2^2 + \lambda \frac{1}{|\mathcal{K}|} \sum_{\mathbf{k} \in \mathcal{K}} (|\hat{\omega}_{\mathbf{k}}| - |\hat{\omega}_{\mathbf{k}}|)^2, \quad (15)$$

with $\lambda = 0.1$ and $\hat{\cdot}$ the orthonormal real 2D DFT on the half-spectrum $\mathcal{K}$ only including mid-high frequencies.

E.2 Physics-constrained model

Physical constraint projection. Given a constraint $c(u) = 0$ defining $\mathcal{C} = \{u : c(u) = 0\}$, enforcing physical consistency on an unconstrained prediction $\tilde{u}$ is the projection problem

$$u^* = \arg \min_{u \in \mathcal{C}} \frac{1}{2} \|u - \tilde{u}\|^2. \quad (16)$$

When $c(u) = Au - b$ is affine, (16) admits the closed form

$$u^* = \tilde{u} - A^\top (AA^\top)^{-1} (A\tilde{u} - b). \quad (17)$$

When c is nonlinear, linearizing about $u^{(k)}$ with Jacobian $J_k = \nabla c(u^{(k)})^\top$ and applying (17) locally yields the Gauss–Newton update

$$u^{(k+1)} = u^{(k)} - J_k^\top (J_k J_k^\top)^{-1} c(u^{(k)}). \quad (18)$$

The constraint used in MSE+MASS is linear, so only (17) is needed.

Mass conservation. Under periodic boundary conditions, the global vorticity is conserved. Integrating (1) over the spatial domain Ω gives

$$\frac{d}{dt} \int_{\Omega} \omega(x, t) dx = \int_{\Omega} (\nu \Delta \omega(x, t) + f(x) - \mathbf{u}(x, t) \cdot \nabla \omega(x, t)) dx. \quad (19)$$

By periodicity and the divergence theorem, $\int_{\Omega} \Delta \omega dx = 0$. The advective term vanishes as well,

$$\int_{\Omega} \mathbf{u} \cdot \nabla \omega dx = \int_{\Omega} \nabla \cdot (\mathbf{u} \omega) dx - \int_{\Omega} (\nabla \cdot \mathbf{u}) \omega dx = 0, \quad (20)$$

since the boundary integral vanishes under periodicity and $\nabla \cdot \mathbf{u} = 0$ by incompressibility. The forcing has zero spatial mean, $\int_{\Omega} f(x) dx = 0$. Together these give $\frac{d}{dt} \int_{\Omega} \omega(x, t) dx = 0$, so the global vorticity is conserved and the corresponding constraint is

$$c_{\text{mass}}(\omega) = \frac{1}{|\Omega|} \int_{\Omega} \omega(x, t) dx - \bar{\omega}_0 = 0. \quad (21)$$

It is linear in ω , so (17) applies directly with $u \equiv \omega \in \mathbb{R}^N$ the discretized vorticity field (N grid points), $A = \frac{1}{N} \mathbf{1}^\top \in \mathbb{R}^{1 \times N}$ the row-averaging operator, so that $Au = \frac{1}{N} \sum_{i=1}^N \omega_i$, and $b = \bar{\omega}_0$ the target mean vorticity. Since $AA^\top = 1/N$, substituting into (17) gives

$$\omega^* = \omega - \left(\frac{1}{|\Omega|} \int_{\Omega} \omega dx - \bar{\omega}_0 \right), \quad (22)$$

which subtracts the mean deviation at every point. The projection is exact, requires no iteration, and costs $O(N)$.

F Qualitative visualizations

This section collects rollout snapshots for each method, to give a sense of what the models actually produce beyond the aggregate metrics.

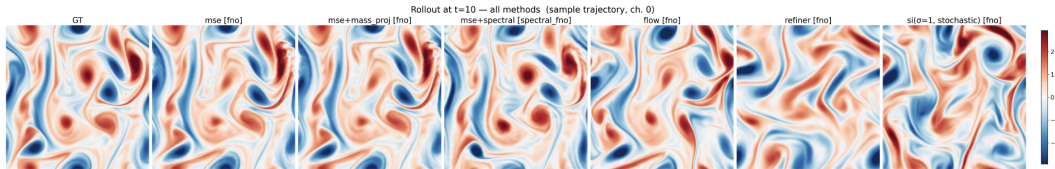


Figure 5: Rollout at step $t = 10$ for each method. The MSE variants develop vortex artifacts that the generative models do not, despite their lower one-step dynamical error.

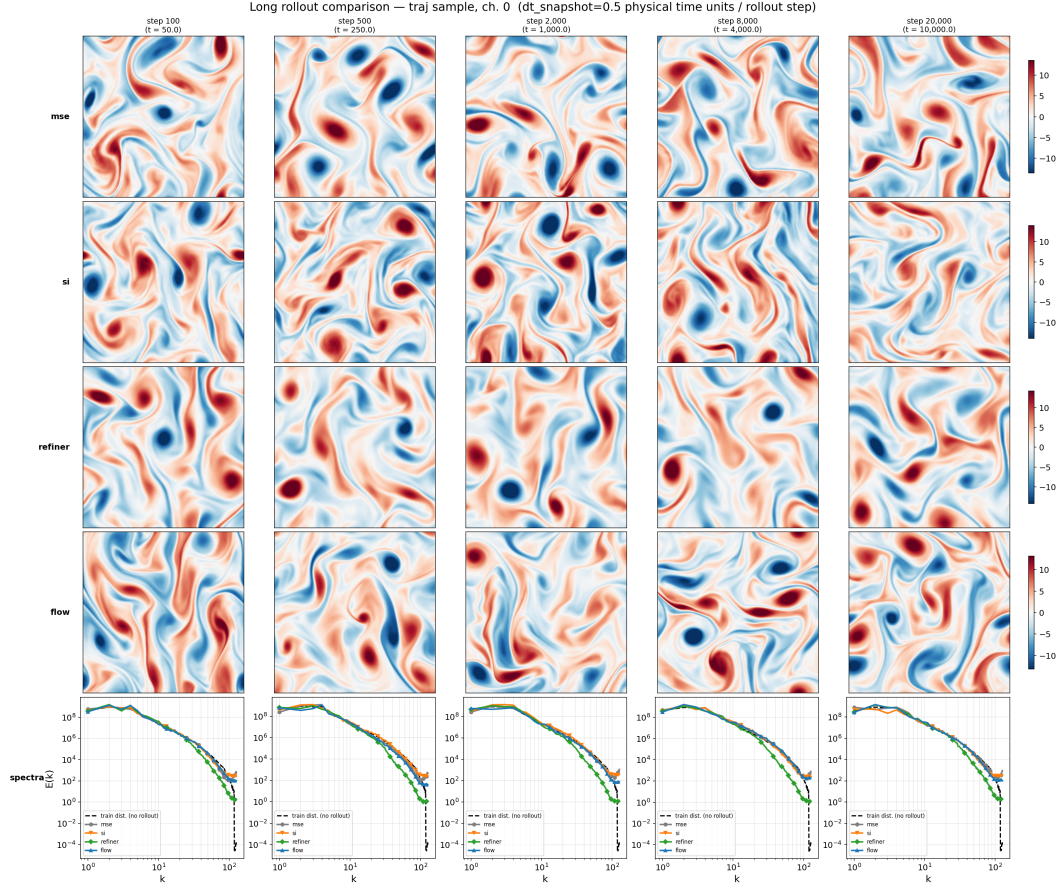


Figure 6: Rollouts of each model over 20 000 steps (10 000 time units) with the FNO backbone. All trajectories stay bounded over the full horizon.