Measuring Memorization and Generalization in Forecasting Models via Structured Perturbations of Chaotic Systems

Anonymous Authors[1]

Abstract

We introduce a benchmarking method for evaluating generalization and memorization in time series forecasting models for chaotic dynamical systems. By generating two complementary types of test sets—by perturbating training trajectories to minimally/maximally diverge over a fixed time horizon—we quantify each model's sensitivity to distribution shift. Our results reveal consistent trade-offs between training accuracy and OOD generalization across neural architectures, offering a lightweight diagnostic tool for model evaluation in the small-data regime.

1. Introduction

Dynamical systems describe how a state evolves over time according to fixed rules, underpinning phenomena from planetary orbits (Contopoulos, 2013) and climate dynamics (Reichstein et al., 2019) to economic cycles (Tu, 2012) and neural activity (Izhikevich, 2007). Accurate forecasts aid transportation, energy, and public-health planning.

Traditional approaches derive governing equations from physics and fit unknown parameters (Rapp et al., 1999; Abarbanel et al., 2009). Their built-in inductive bias yields interpretability and often reliable extrapolation but fails when key physics are missing—e.g., climate models that omit small eddies under-predict extreme weather (Ma et al., 2015; Czaja et al., 2019). Data-driven methods—from early local-linear predictors (Farmer & Sidorowich, 1987) to reservoir computing (Jaeger & Haas, 2004; Gauthier et al., 2021), recurrent nets (Bailer-Jones et al., 1998; Wang, 2017), neural ODEs (Chen et al., 2018), dynamic-mode decomposition (Schmid, 2010), and sparse-library regression (Brunton et al., 2016)—learn directly from data but often generalise poorly beyond the training regime.

Real deployments encounter unseen regimes, so *out-of-distribution* (OOD) generalisation is a critical benchmark. In computer vision this means new object classes (Liu et al., 2023); in RL, harder tasks (Cobbe et al., 2020). For dynamical systems, researchers perturb initial conditions, vary parameters, or corrupt observations, yet choices are usually ad-hoc, leading to incomparable robustness claims.

We propose a **small, reproducible stress test of OOD generalization in dynamical systems modelling**. For any reference trajectory we (1) linearise the dynamics to obtain Jacobian eigen-directions, (2) nudge the initial state by the same radius along a *stable* eigenvector (*near-OOD*) and an *unstable* eigenvector (*far-OOD*), and (3) compare forecasting error on the paired trajectories. The identical step size isolates sensitivity to the first unstable mode.

Contributions. (1) A Jacobian-based generator for paired near/far-OOD tests; (2) an open-source, Colab-ready toolkit; (3) empirical evaluation of 30 different chaotic dynamical systems predicted by KNN, LSTM, NBEATS, TIDE, and Transformer models.

2. Related Work

In dynamical-systems modeling and reconstruction, out-of-distribution/domain (OOD) evaluation is important for assessing how well a learned model performs under conditions that differ substantially from its training data. Because dynamical

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

systems are so easily simulated, researchers can introduce systematic "knobs" for OOD testing such as varying the system's equations, its physical parameters, or its initial conditions. They can also simulate realistic observational shifts such as added noise or coarser time resolution. Although many papers explore multiple types of distribution shifts (e.g. (Song et al., 2025; Mouli et al., 2024; McCabe et al., 2024)), the specific choices tend to be arbitrary and inconsistent, making it difficult to compare models and benchmark progress across the field.

Initial Conditions Evaluation on unseen initial conditions or trajectories is perhaps the most standard OOD test. Many studies measure generalization by reserving later time-steps of a trajectory for testing, forcing models to extrapolate beyond the temporal window seen during training (Zhang & Saxena, 2024; Volkmann et al., 2024; Park et al., 2024). Others deliberately sample training and test initial states randomly (Levine & Stuart, 2022; Vlachas et al., 2018; Kaptanoglu et al., 2023) or from different distributions (Bhamidipaty et al., 2023; Mouli et al., 2024; Song et al., 2025; Bihlo, 2024; Yang & Osher, 2024). For instance, DynaDojo provides a unified benchmarking platform for canonical systems like Lorenz, N-body, and Kuramoto, comparing model performance on in-distribution versus OOD initial conditions drawn from a separate region of state-space (Bhamidipaty et al., 2023). MetaPhysiCA further examines two axes of shift—novel initial states alone, and joint shifts in both initial conditions and ODE parameters—and demonstrates substantial error reductions over non-causal baselines under both settings (Mouli et al., 2024).

Dynamical Regime Not all initial-condition shifts are created equal. Small perturbations can dramatically alter long-term trajectory geometry in chaotic systems, while multimodal attractors admit qualitatively distinct dynamical regimes. McCabe et al. categorize Navier–Stokes initial states by the behaviors they induce (i.e. Random Periodic versus Turbulent) (McCabe et al., 2024). And Goring et al. argue that true OOD generalization requires crossing attractor basins. In their framework, a multistable system is partitioned by attractor, with some basins held out entirely for testing so that models must generalize to regimes never seen during training (Göring et al., 2024).

Dynamics/Parameters Beyond input changes, shifts in physical parameters probe a model's robustness to changes in system dynamics. The LEADS framework formalizes each "environment" as a distinct distribution over parameter vectors (e.g. Lotka–Volterra coefficients), drawing training and test values from non-overlapping ranges to evaluate generalization (Yin et al., 2021). Subramanian et al. show that pre-training a transformer-style neural operator on PDEs with diffusion coefficients in [0.1, 1.0] then fine-tuning on [1.1, 2.0] yields orders-of-magnitude gains in sample efficiency compared to learning from scratch (Subramanian et al., 2023). Li & Yang introduce a dual-branch neural operator in which an in-distribution branch handles $\lambda \in [1, 5]$ (Poisson eigenvalues) or $k \in [0.5, 2.0]$ (Helmholtz wavenumbers) and an OOD branch transfers pseudo-solutions to recover accuracy when $\lambda > 5$ or $k > 2$ without the need for OOD labels (Li & Yang, 2025).

Systems Transfer across entirely new dynamical systems and tasks represents an even more ambitious form of OOD evaluation. FMint, a foundation model pre-trained on a diverse corpus of ODE trajectories, can be fine-tuned with as few as 25 to 1000 samples to model unseen systems—from driven-damped pendulums to FitzHugh–Nagumo—outperforming state-of-the-art simulators and numerical solvers (Song et al., 2025). Even zero-shot forecasting is possible. A general-purpose time-series foundation model, despite never seeing dynamical-system data during pre-training, accurately predicts 135 chaotic systems for up to one Lyapunov time while preserving attractor geometry (Zhang & Gilpin, 2024). LLMs, too, have shown improved emergent reasoning for solving tasks from visual puzzles to chess moves when pre-trained on more complex cellular automata (Zhang et al., 2025).

Observation Finally, realistic application demands robustness to observational shifts such as sensor noise and sampling rates. Prior work has evaluated methods for sensitivity to noise (Stepaniants, 2023; Volkmann et al., 2024) and examined performance under coarser or irregular time-step strides (Song et al., 2025) to simulate non-ideal measurement scenarios.

Together, these varied OOD tests form a rich but fragmented landscape. It is clear that we need clarity to drive reproducible progress in dynamical systems forecasting.

3. Methods

Problem Setting

Let $\{x_{\text{ref}}^{(i)}(t)\}_{i=1}^{N}$ denote a training set of trajectories from a dynamical system, used to train a predictive model. To evaluate the model's ability to generalize beyond the training distribution, we construct in-distribution (ID) and out-of-distribution (OOD) test trajectories by perturbing each reference trajectory $x_{\text{ref}}(t)$ in two controlled ways. For each trajectory, we generate an ID variant $x_{\text{id}}(t)$ and an OOD variant $x_{\text{ood}}(t)$ by modifying the initial condition:

$$x_{\text{id}}(0) = x_{\text{ref}}(0) + \delta x_{\text{id}}(0), \quad x_{\text{ood}}(0) = x_{\text{ref}}(0) + \delta x_{\text{ood}}(0).$$

Both perturbations are designed to have the same integrated deviation ϵ over an initial window $[0, C]$ (with $C < T$), but differ maximally in their deviation over the full interval $[0, T]$. This setup enables us to isolate the model's performance under controlled deviations and quantify the generalization gap between ID and OOD scenarios.

Generating ID and OOD Test Trajectories via Constrained Perturbations

To construct perturbations for nonlinear systems, we linearize the dynamics around a reference trajectory and solve a generalised eigenvalue problem to identify the directions that minimize or maximize the deviation over a long time window $[0, T]$, under a fixed integrated deviation constraint over $[0, C]$.

Setup. Let $x_{\text{ref}}(t)$ be a reference trajectory governed by a nonlinear dynamical system $\dot{x} = f(x)$. The deviation $\delta x(t) = x(t; x_0 + \delta) - x_{\text{ref}}(t)$ evolves approximately as:

$$\dot{\delta x} \approx A(t)\delta x, \quad A(t) = Df(x_{\text{ref}}(t)).$$

We define the state transition matrix $\Phi(t, 0)$ as the solution to $\dot{\Phi} = A(t)\Phi$ with $\Phi(0, 0) = I$. Then:

$$\delta x(t) \approx \Phi(t, 0)\delta x(0).$$

Generalised Gramian Construction. Define two Gramians:

$$W_C = \int_0^C \Phi(t, 0)^\top \Phi(t, 0)\, dt,$$
$$W_T = \int_0^T \Phi(t, 0)^\top \Phi(t, 0)\, dt.$$

Constrained Optimisation. We solve the generalised eigenvalue problem:

$$W_T v = \lambda W_C v$$

and take the top eigenvector v_1 (maximally divergent) and bottom eigenvector v_n (minimally divergent). We then scale each to satisfy the constraint:

$$\delta^\top W_C \delta = C\epsilon^2.$$

Thus, the perturbations become:

$$\delta_{\text{ood}} = \epsilon \cdot \frac{v_1}{\sqrt{v_1^\top W_C v_1}}, \quad \delta_{\text{id}} = \epsilon \cdot \frac{v_n}{\sqrt{v_n^\top W_C v_n}}.$$

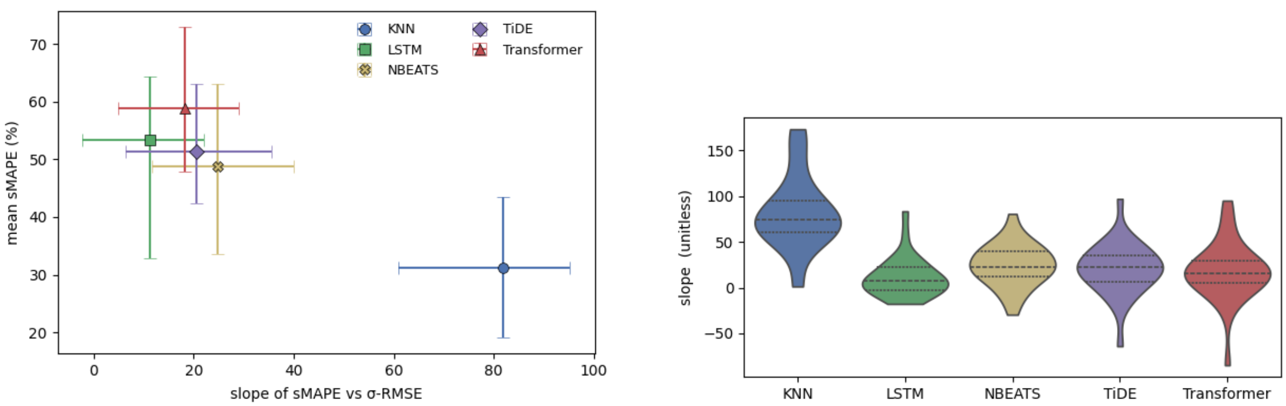
The full nonlinear system is then simulated forward from these perturbed initial conditions.

Algorithm 1 Construct ID/OOD Perturbations via Generalised Eigenproblem

Input: Nonlinear system $\dot{x} = f(x)$, reference trajectory $x_{\text{ref}}(t)$, horizon T, constraint window $C < T$, error level ϵ
Output: Perturbations $x_{\text{id}}(t), x_{\text{ood}}(t)$
1. Compute Jacobian trajectory $A(t) = Df(x_{\text{ref}}(t))$ for $t \in [0, T]$
2. Integrate tangent-linear system $\dot{\Phi} = A(t)\Phi$ to obtain $\Phi(t, 0)$
3. Compute Gramians $W_C = \int_0^C \Phi^\top \Phi \, dt$, $W_T = \int_0^T \Phi^\top \Phi \, dt$
4. Solve $W_T v = \lambda W_C v$; sort eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$
5. Scale eigenvectors v_1, v_n to satisfy $\delta^\top W_C \delta = C\epsilon^2$
6. Set $x_{\text{ood}}(0) = x_{\text{ref}}(0) + \delta_{\text{ood}}$, $x_{\text{id}}(0) = x_{\text{ref}}(0) + \delta_{\text{id}}$
7. Simulate nonlinear system from both initial conditions
8. **return** $x_{\text{id}}(t), x_{\text{ood}}(t)$

4. Results

We begin with global summaries (Fig. 1), then zoom into a representative system (`ThomasLabyrinth`) across all five models (Fig. 2). Full per–system heat-maps are deferred to Appendix A.



(a) Mean σ-RMSE *vs.* mean sMAPE with IQR error bars.

(b) Distribution of slopes ($\partial\,\text{sMAPE}/\partial\sigma$).

Figure 1. **Global error behaviour across 30 chaotic systems.** *(a)* Models positioned by their mean error and error–shift slope. *(b)* Spread of slopes per model.

Generalization Behavior Across Models. Figure 1 reveals a trade-off between forecasting accuracy and robustness to distribution shift. The KNN baseline achieves the **lowest mean sMAPE** across all systems, but also exhibits by far the **largest positive slope**—indicating that its performance degrades steeply as the test trajectories diverge from training data. This aligns with the fact that our test sets are perturbations of the training set: KNN performs well by memorization when test points lie close to training samples, but generalizes poorly beyond that neighborhood.

Among the neural models, the **relative ordering of mean sMAPE** matches prior findings in Zhang & Gilpin (2024), with NBEATS and TIDE outperforming TRANSFORMER and LSTM. However, our analysis goes further by also capturing *how performance changes with increasing distribution shift*. We observe an **inverse relationship between average sMAPE and slope**: models with *better average performance tend to have larger slopes*, suggesting they rely more heavily on memorization and degrade more rapidly when test inputs differ from training. Conversely, models with *flatter slopes* (e.g., LSTM) exhibit more stable generalization, even if their average error is slightly higher.

These findings demonstrate that our method provides an efficient, diagnostic proxy for comparing memorization and generalization behaviors across models on small chaotic datasets—information not available from sMAPE alone.

References

Abarbanel, H. D. I., Creveling, D. R., Farsian, R., and Kostuk, M. Dynamical State and Parameter Estimation. *SIAM Journal on Applied Dynamical Systems*, 8(4):1341–1381, 2009. doi: 10.1137/090749761. URL https://doi.org/10.1137/090749761. _eprint: https://doi.org/10.1137/090749761.

Bailer-Jones, C. A. L., MacKay, D. J. C., and Withers, P. J. A recurrent neural network for modelling dynamical systems. *Network: Computation in Neural Systems*, 9(4):531–547, January 1998. ISSN 0954-898X, 1361-6536. doi: 10.1088/0954-898X_9_4_008. URL https://www.tandfonline.com/doi/full/10.1088/0954-898X_9_4_008.

Bhamidipaty, L. M., Bruzzese, T., Tran, C., Ratl Mrad, R., and Kanwal, M. S. DynaDojo: An Extensible Platform for Benchmarking Scaling in Dynamical System Identification. *Advances in Neural Information Processing Systems*, 36: 15519–15530, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/32093649cbbcff773d9a991d8c30a7fe-Abstract-Datasets_and_Benchmarks.html.

Bihlo, A. Improving physics-informed neural networks with meta-learned optimization. *J. Mach. Learn. Res.*, 25(1): 14:755–14:780, January 2024. ISSN 1532-4435.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data: Sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, April 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1517384113. URL http://arxiv.org/abs/1509.03580. arXiv:1509.03580 [math].

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Cobbe, K., Hesse, C., Hilton, J., and Schulman, J. Leveraging Procedural Generation to Benchmark Reinforcement Learning, July 2020. URL http://arxiv.org/abs/1912.01588. arXiv:1912.01588 [cs].

Contopoulos, G. *Order and Chaos in Dynamical Astronomy*. Springer Science & Business Media, March 2013. ISBN 978-3-662-04917-4. Google-Books-ID: XdvzCAAAQBAJ.

Czaja, A., Frankignoul, C., Minobe, S., and Vannière, B. Simulating the Midlatitude Atmospheric Circulation: What Might We Gain From High-Resolution Modeling of Air-Sea Interactions? *Current Climate Change Reports*, 5(4): 390–406, December 2019. ISSN 2198-6061. doi: 10.1007/s40641-019-00148-5. URL https://doi.org/10.1007/s40641-019-00148-5.

Farmer, J. D. and Sidorowich, J. J. Predicting chaotic time series. *Physical Review Letters*, 59(8):845–848, August 1987. doi: 10.1103/PhysRevLett.59.845. URL https://link.aps.org/doi/10.1103/PhysRevLett.59.845. Publisher: American Physical Society.

Gauthier, D. J., Bollt, E., Griffith, A., and Barbosa, W. A. S. Next generation reservoir computing. *Nature Communications*, 12(1):5564, September 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25801-2. URL https://www.nature.com/articles/s41467-021-25801-2. Publisher: Nature Publishing Group.

Göring, N., Hess, F., Brenner, M., Monfared, Z., and Durstewitz, D. Out-of-Domain Generalization in Dynamical Systems Reconstruction, June 2024. URL http://arxiv.org/abs/2402.18377. arXiv:2402.18377 [cs].

Izhikevich, E. M. *Dynamical Systems in Neuroscience*. MIT Press, 2007. ISBN 978-0-262-09043-8.

Jaeger, H. and Haas, H. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, 304(5667):78–80, April 2004. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1091277. URL https://www.science.org/doi/10.1126/science.1091277.

Kaptanoglu, A. A., Zhang, L., Nicolaou, Z. G., Fasel, U., and Brunton, S. L. Benchmarking sparse system identification with low-dimensional chaos. *Nonlinear Dynamics*, 111(14):13143–13164, July 2023. ISSN 1573-269X. doi: 10.1007/s11071-023-08525-4. URL https://doi.org/10.1007/s11071-023-08525-4.

Levine, M. and Stuart, A. A framework for machine learning of model error in dynamical systems. *Communications of the American Mathematical Society*, 2(07):283–344, 2022. ISSN 2692-3688. doi: 10.1090/cams/10. URL https://www.ams.org/cams/2022-02-07/S2692-3688-2022-00010-9/.

Li, J. and Yang, M. Dual-branch neural operator for enhanced out-of-distribution generalization. *Engineering Analysis with Boundary Elements*, 171:106082, February 2025. ISSN 0955-7997. doi: 10.1016/j.enganabound.2024.106082. URL https://www.sciencedirect.com/science/article/pii/S0955799724005551.

Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards Out-Of-Distribution Generalization: A Survey, July 2023. URL http://arxiv.org/abs/2108.13624. arXiv:2108.13624 [cs].

Ma, X., Chang, P., Saravanan, R., Montuoro, R., Hsieh, J.-S., Wu, D., Lin, X., Wu, L., and Jing, Z. Distant Influence of Kuroshio Eddies on North Pacific Weather Patterns? *Scientific Reports*, 5(1):17785, December 2015. ISSN 2045-2322. doi: 10.1038/srep17785. URL https://www.nature.com/articles/srep17785. Publisher: Nature Publishing Group.

McCabe, M., Blancard, B. R.-S., Parker, L. H., Ohana, R., Cranmer, M., Bietti, A., Eickenberg, M., Golkar, S., Krawezik, G., Lanusse, F., Pettee, M., Tesileanu, T., Cho, K., and Ho, S. Multiple Physics Pretraining for Spatiotemporal Surrogate Models. November 2024. URL https://openreview.net/forum?id=DKSI3bULiZ.

Mouli, S. C., Alam, M. A., and Ribeiro, B. MetaPhysiCa: Improving OOD Robustness in Physics-informed Machine Learning. In *Proceedings of the 12th International Conference on Learning Representations*, ICLR '24, 2024.

Park, J., Yang, N. T., and Chandramoorthy, N. When are dynamical systems learned from time series data statistically accurate? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS '24, November 2024. URL https://openreview.net/forum?id=4t3ox9hj3z&referrer=%5Bthe%20profile%20of%20Nisha%20Chandramoorthy%5D(%2Fprofile%3Fid%3D~Nisha_Chandramoorthy1).

Rapp, P. E., Schmah, T. I., and Mees, A. I. Models of knowing and the investigation of dynamical systems. *Physica D: Nonlinear Phenomena*, 132(1):133–149, July 1999. ISSN 0167-2789. doi: 10.1016/S0167-2789(99)00035-4. URL https://www.sciencedirect.com/science/article/pii/S0167278999000354.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, February 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0912-1. URL https://www.nature.com/articles/s41586-019-0912-1. Publisher: Nature Publishing Group.

Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, August 2010. ISSN 1469-7645, 0022-1120. doi: 10.1017/S0022112010001217. URL https://www.cambridge.org/core/journals/journal-of-fluid-mechanics/article/abs/dynamic-mode-decomposition-of-numerical-and-experimental-data/AA4C763B525515AD4521A6CC5E10DBD4.

Song, Z., Yuan, J., and Yang, H. FMint: Bridging Human Designed and Data Pretrained Models for Differential Equation Foundation Model for Dynamical Simulation. *Advanced Theory and Simulations*, pp. 2500062, April 2025. ISSN 2513-0390, 2513-0390. doi: 10.1002/adts.202500062. URL https://advanced.onlinelibrary.wiley.com/doi/10.1002/adts.202500062.

Stepaniants, G. Learning partial differential equations in reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.*, 24(1): 86:3887–86:3958, January 2023. ISSN 1532-4435.

Subramanian, S., Harrington, P., Keutzer, K., Bhimji, W., Morozov, D., Mahoney, M. W., and Gholami, A. Towards foundation models for scientific machine learning: characterizing scaling and transfer behavior. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 71242–71262, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Tu, P. N. V. *Dynamical Systems: An Introduction with Applications in Economics and Biology*. Springer Science & Business Media, December 2012. ISBN 978-3-642-78793-5. Google-Books-ID: 7Ff9CAAAQBAJ.

Vlachas, P. R., Byeon, W., Wan, Z. Y., Sapsis, T. P., and Koumoutsakos, P. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2213):20170844, May 2018. doi: 10.1098/rspa.2017.0844. URL https://royalsocietypublishing.org/doi/10.1098/rspa.2017.0844. Publisher: Royal Society.

Volkmann, E., Brändle, A., Durstewitz, D., and Koppe, G. A scalable generative model for dynamical system reconstruction from neuroimaging data. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS '24, November 2024. URL https://openreview.net/forum?id=exATQD4HSv.

Wang, Y. A new concept using LSTM Neural Networks for dynamic system identification. In *2017 American Control Conference (ACC)*, pp. 5324–5329, May 2017. doi: 10.23919/ACC.2017.7963782. URL https://ieeexplore.ieee.org/abstract/document/7963782. ISSN: 2378-5861.

Yang, L. and Osher, S. J. PDE generalization of in-context operator networks: A study on 1D scalar nonlinear conservation laws. *Journal of Computational Physics*, 519:113379, December 2024. ISSN 0021-9991. doi: 10.1016/j.jcp.2024.113379. URL https://www.sciencedirect.com/science/article/pii/S0021999124006272.

Yin, Y., Ayed, I., de Bézenac, E., Baskiotis, N., and Gallinari, P. LEADS: Learning Dynamical Systems that Generalize Across Environments. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7561–7573. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/3df1d4b96d8976ff5986393e8767f5b2-Abstract.html.

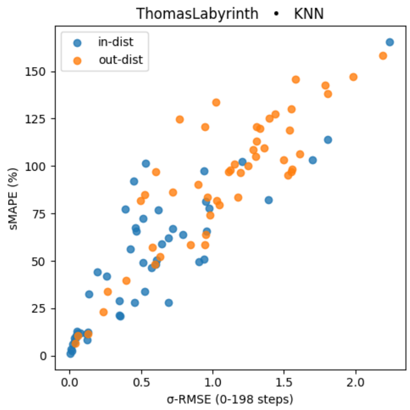
Zhang, S., Patel, A., Rizvi, S. A., Liu, N., He, S., Karbasi, A., Zappala, E., and Dijk, D. v. Intelligence at the Edge of Chaos, March 2025. URL http://arxiv.org/abs/2410.02536. arXiv:2410.02536 [cs].

Zhang, Y. and Gilpin, W. Zero-shot forecasting of chaotic systems. October 2024. URL https://openreview.net/forum?id=TqYjhJrp9m.

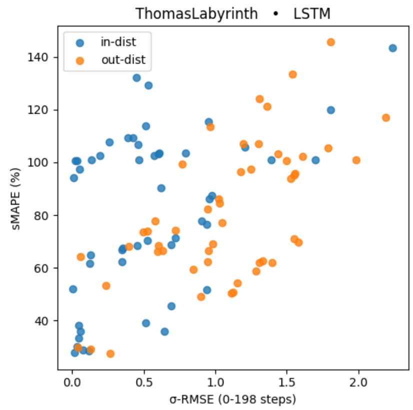
Zhang, Y. and Saxena, S. Inference of Neural Dynamics Using Switching Recurrent Neural Networks. November 2024. URL https://openreview.net/forum?id=zb8jLAh2VN.

A. Per-System Results Heatmaps

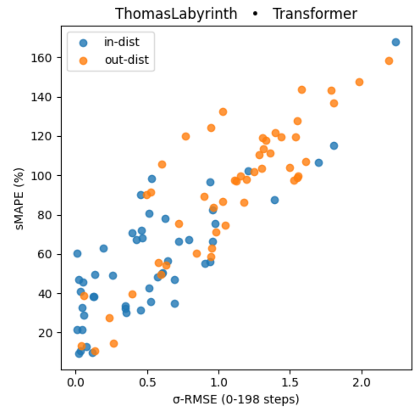
This appendix provides representative slopes observed when models are trained on a particular system and prediction errors are plotted against test vs. training trajectory distances. The complete system-by-model breakdown for both mean sMAPE and slope values. The heatmaps supplement the aggregate findings presented in Section 4.



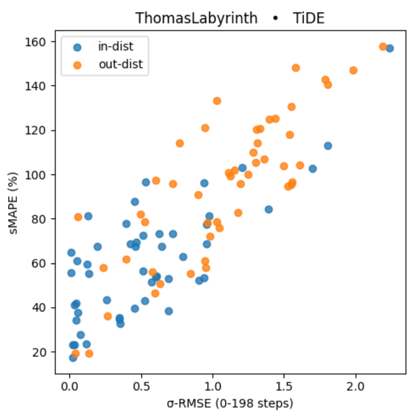
(a) KNN



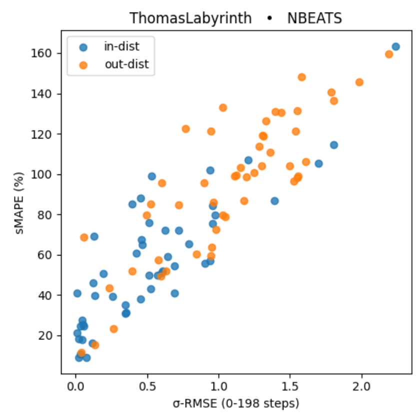
(b) LSTM



(c) Transformer



(d) TiDE



(e) N-BEATS

Figure 2. **Per-trajectory scatter on the `TL` system.** Each point is a single trajectory: standardized-RMSE(test || train) on the x-axis and sMAPE(pred || test) on the y-axis. In-distribution points are shown in blue, out-of-distribution in orange. The pattern is representative of what one sees across systems where the larger the deviation between training and test trajectories, the larger the prediction sMAPE is. However, the slope differs for each model, which relates to each model's degree of generalization.

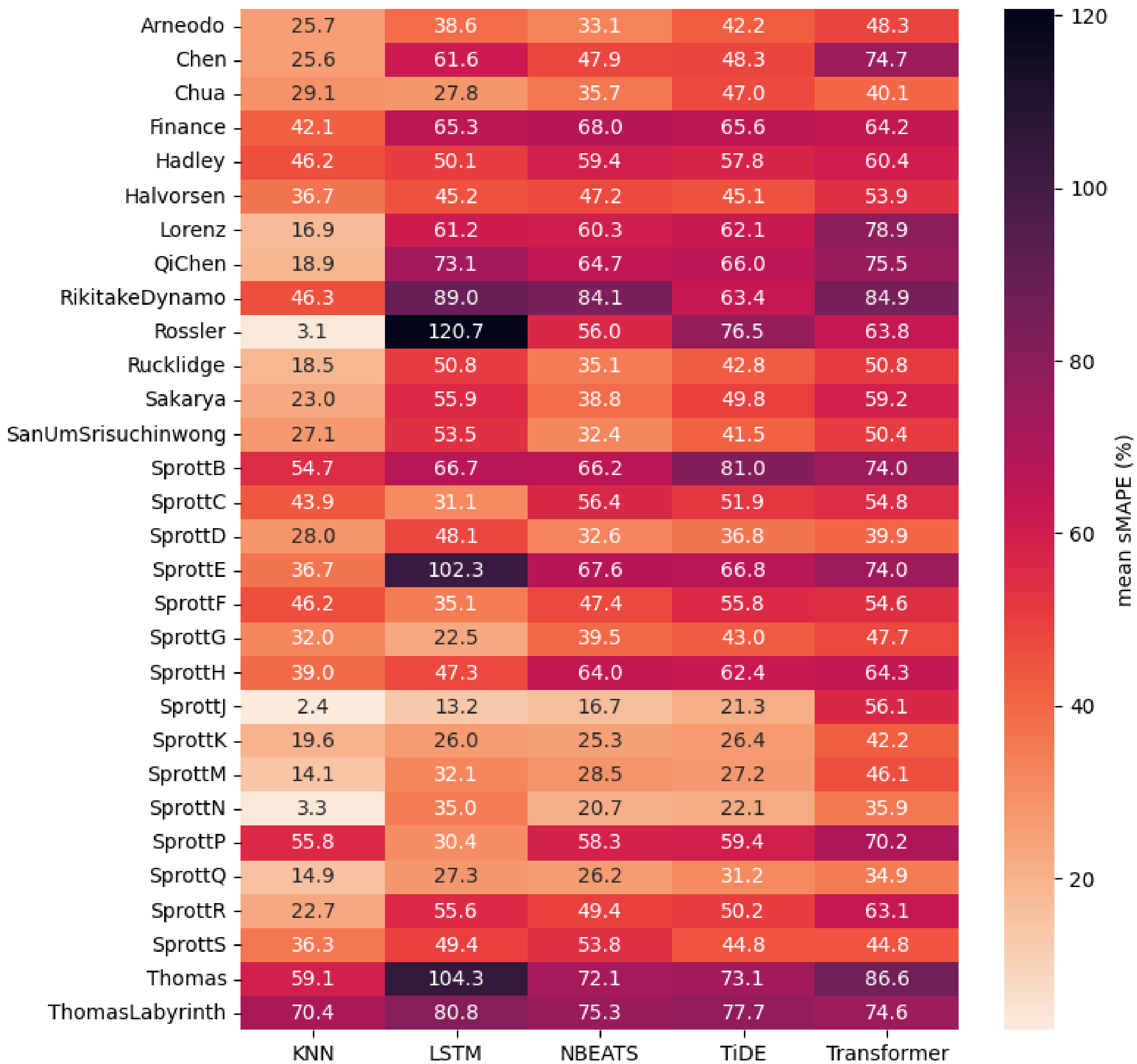


Figure 3. **Per-system mean sMAPE.** Each cell reports the mean symmetric mean absolute percentage error for a given chaotic system (row) and model (column). Darker shades indicate better performance (lower error). This figure allows fine-grained comparison across all 30 systems.

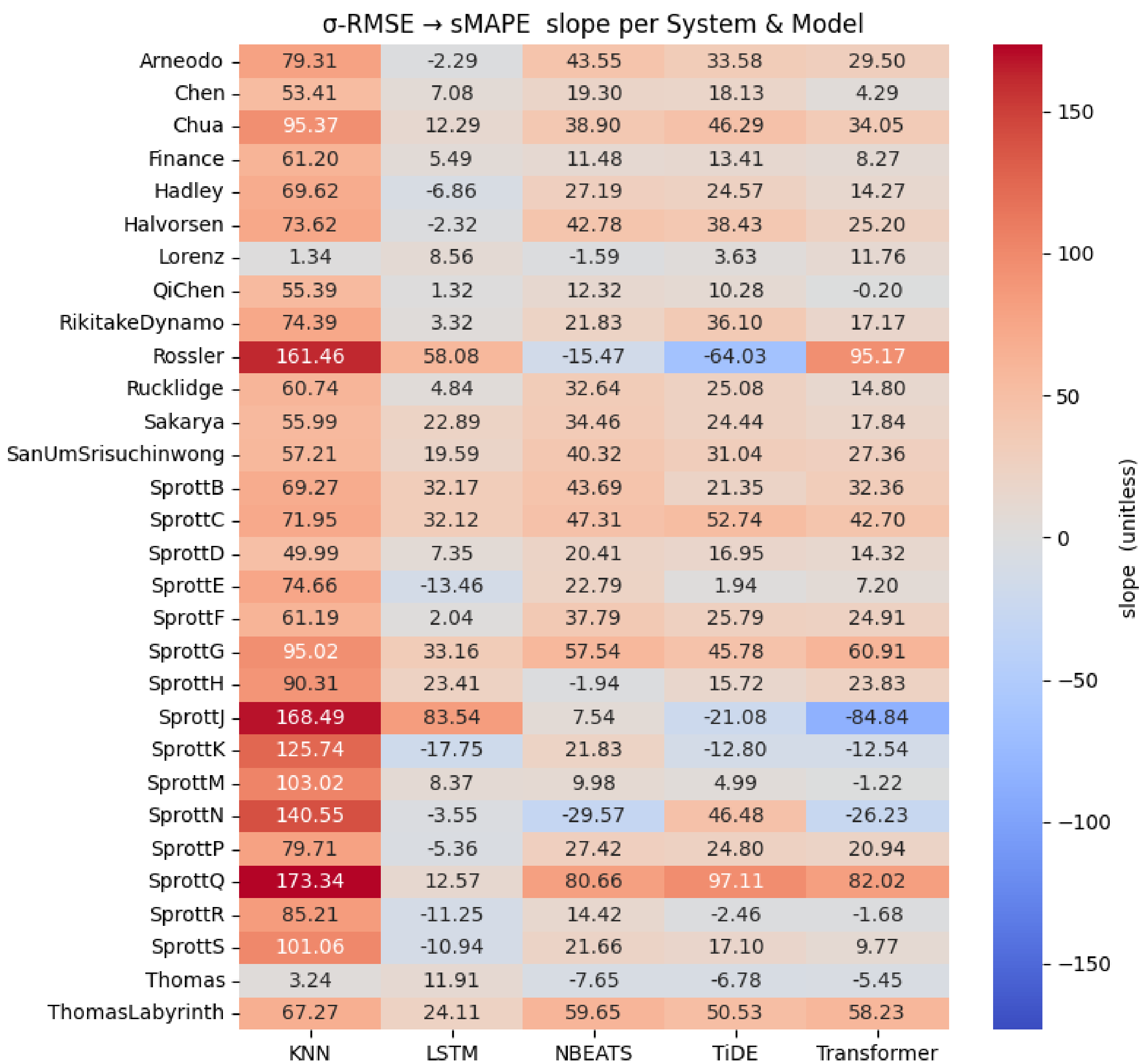


Figure 4. **Per-system slopes: ∂ sMAPE / ∂ shift.** Red indicates systems where model performance degrades sharply as the input distribution shifts away from the training data. Blue indicates stability or error reduction with shift. These values correspond to the slopes used in Figures 1 and 2.

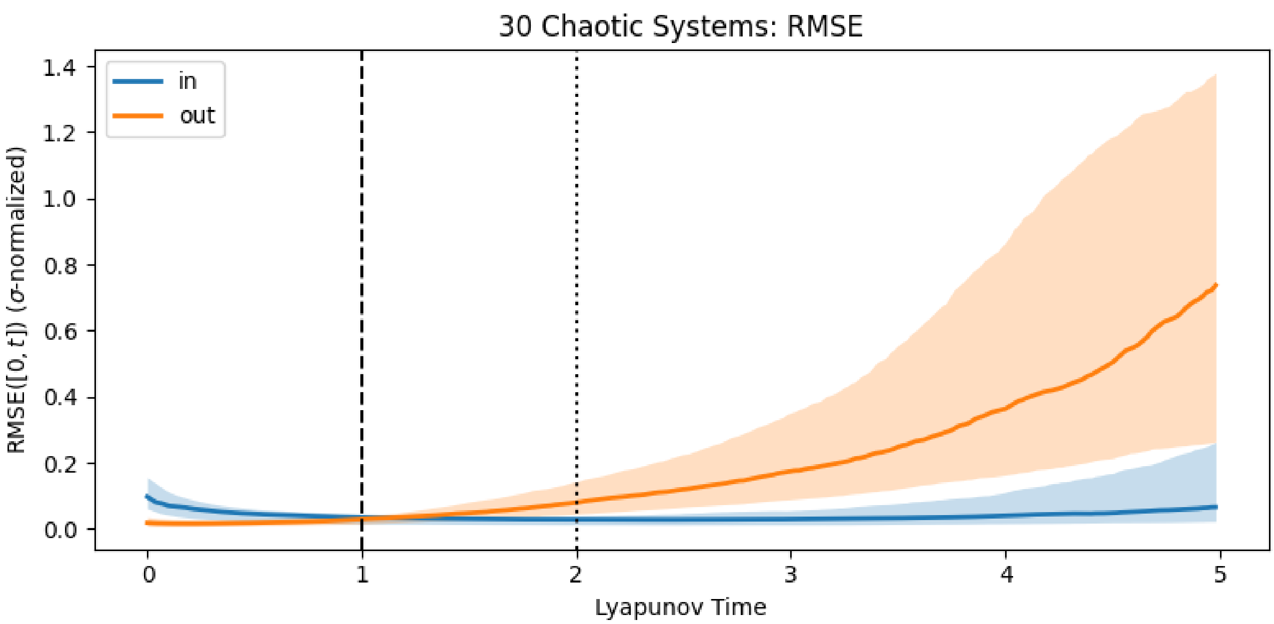


Figure 5. We also include evidence that the perturbation method works effectively across a variety of 30 chaotic dynamical systems by plotting the integrated standardized-RMSE over Lyapunov time between the in- and out-of-distribution test sets that were perturbed to lead to maximal and minimal divergence, respectively, relative to the reference training trajectory over window of $T = 2$ Lyapunov times. The in and out initial condition perturbations were constrained to cause equal RMSE divergence over a shorter window $C = 1$ Lyapunov time.