Improved deep learning of chaotic dynamical systems with multistep penalty losses

Dibyajyoti Chakraborty[*]
School of Information Sciences and Technology
Pennsylvania State University
University Park, PA-16802, USA.

Seung Whan Chung
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
Livermore, CA-94550, USA

Ashesh Chattopadhyay
Department of Applied Mathematics
University of California Santa Cruz
Santa Cruz, CA-95064, USA

Romit Maulik
School of Information Sciences and Technology
Pennsylvania State University
University Park, PA-16802, USA.

Abstract

Predicting the long-term behavior of chaotic systems remains a formidable challenge due to their extreme sensitivity to initial conditions and the inherent limitations of traditional data-driven modeling approaches. This paper introduces a novel framework that addresses these challenges by leveraging the recently proposed multi-step penalty (MP) optimization technique. Our approach extends the applicability of MP optimization to a wide range of deep learning architectures, including Fourier Neural Operators and UNETs. By introducing penalized local discontinuities in the forecast trajectory, we effectively handle the non-convexity of loss landscapes commonly encountered in training neural networks for chaotic systems. We demonstrate the effectiveness of our method through its application to two challenging use-cases: the prediction of flow velocity evolution in two-dimensional turbulence and ocean dynamics using reanalysis data. Our results highlight the potential of this approach for accurate and stable long-term prediction of chaotic dynamics, paving the way for new advancements in data-driven modeling of complex natural phenomena.

1 Introduction

Chaotic systems are ubiquitous in nature, encompassing fields as diverse as meteorology, fluid dynamics, and chemical reactions. They exhibit complex multi-scale dynamics, without any scale separation, making their forecasts extremely challenging. They are also characterized by their extreme sensitivity to initial conditions, meaning a small perturbation in the initial condition leads to completely diverging trajectories over time. A deterministic long-term prediction for chaotic systems is irrelevant due to their very nature. Therefore, several current works on data driven long term prediction of chaotic systems focus on preserving the invariant statistics of the system[Linot et al., 2023, Schiff et al., 2024, Li et al., 2021, Guan et al., 2024]. However, minimizing the deviations from ground truth in one autoregressive time step of prediction using a mean-squared error, typically used to optimize ML models, is not effective for long term dynamics. Recent developments have tried to tackle this limitation by several techniques like using multiple timesteps for accumulating errors before gradient computation [Keisler, 2022], including structures from governing differential equations[Linot et al., 2023] and implementing physical laws in optimization [Raissi et al., 2019].

In this work, we focus on the challenges data-driven ML models trained using multiple timesteps (rollouts) face for predicting chaotic systems. Gradient-based optimization used in neural networks, aiming to minimize the difference between predictions and ground truth, proves particularly difficult for such systems. The extreme sensitivity to perturbations leads to exploding gradients during

[*] Corresponding author: d.chakraborty@psu.edu

1st Conference on Applied AI and Scientific Machine Learning (CASML 2024).

optimization for any long term objective with underlying chaotic dynamics [Lea et al., 2000]. Additionally, even a theoretically convex objective function becomes highly non-convex in numerical implementation when involving long chaotic trajectories, often trapping the optimization process in sub-optimal local minima [Chung and Freund, 2022]. For more details on non-convexity and loss landscape we refer readers to Chakraborty et al. [2024]. This challenge shares similarities with the well-known exploding/vanishing gradient problem in deep learning [Hanin, 2018, Philipp et al., 2017]. Similar to how chaotic dynamics evolve, the repeated application of deep neuron layers can cause extreme gradients while automatic differentiation, obstructing the training process. Although there are some previous works by the machine learning community(Refer section 2 in Chakraborty et al. [2024]) trying to tackle this issue, it is still an open problem and an area of active research.

A solution to the exploding gradients problem, proposed by Blonigan et al. [2014] is the Lease Square Shadowing (LSS) method. They use the shadowing lemma, which states that there exists a trajectory that always stays close to the reference trajectory with slightly perturbed initial conditions. The brute-force computation of shadowing lemma requires a cubic cost with respect to the number of parameters rendering it unusable for deep learning. Chung and Freund [2022] introduced the multi-step penalty(MP) optimization which uses segmented time intervals and introduces penalized local discontinuities to optimize the objective along with minimizing the discontinuities. As an alternative to LSS method, Chakraborty et al. [2024] showed that the MP optimization can reduce the gradient computation cost from cubic to linear with respect to the number of parameters. They implemented it on several chaotic systems like Lorenz, 2D turbulence and weather to achieve long term stability. However, the dynamical core of their work was based specifically on the Neural Ordinary Differential Equations [Chen et al., 2018]. In this paper, we propose a modified extension of the MP optimization algorithm to other deep learning algorithms that can predict dynamics autoregressively, thereby assessing the general applicability of the optimization technique. We implement it on two popular deep learning architectures for dynamical systems, namely the Fourier Neural Operator(Li et al. [2020]) and UNET (Ronneberger et al. [2015]). We focus on two chaotic dynamical systems - High Reynolds number (Re $\sim 10^5$) 2D turbulence with Kolmogorov forcing and the northwest Atlantic Ocean western boundary current (the Gulf Stream) dataset introduced in Chattopadhyay et al. [2024].

2 Methodology

Let us consider an operator S that advances the state q of a dynamical system in time. It can be considered as the theoretical solution of any underlying governing differential equation that controls the dynamics of a system. So, the state evolution can be given as,

$$q_t = S(q_{t-1}) = S(S(S(...S(q_0)))) = S^t(q_0) \tag{1}$$

where q_t is the state at time t. S can be approximated by a neural network architecture $F_\theta(q)$, depending on learnable parameters θ. The parameters are obtained by minimizing the mismatch from ground truth data (with discrete index i) given by a 1 step loss function,

$$L_1 = \mathbb{E}_i \left[\| F_\theta(q_i) - S(q_i) \| \right] \tag{2}$$

The popular multi-rollout loss function, L_M used to train several state-of-the-art models is defined as

$$L_M = \mathbb{E}_i \left[\sum_{t=1}^{t=n} \left\| \lambda(t)(F_\theta^t(q_i) - S^t(q_i)) \right\| \right] \tag{3}$$

where n is the number of rollouts that the training sees and $\lambda(t)$ is a hyper-parameter that gives lower weights to mismatch in trajectories that are farther in time. Furthermore, the 'Pushforward Trick' introduced in Brandstetter et al. [2022] can be used to reduce the computational cost and induce stability by breaking the computational graph between intermediate rollouts. However, these techniques are themselves insufficient to capture the invariant metric of the underlying dynamical system for prediction of chaotic systems (Schiff et al. [2024]). The MP method introduces learnable intermediate discontinuities in the long trajectory and adds a penalty term to the rollout loss defined in Equation 3 penalizing the magnitude of the discontinuities. Therefore, the problem of extreme gradients can be overcome and the trajectory learned is stable without explicitly specifying invariant properties in the loss function as in Schiff et al. [2024] which are either unknown or computationally expensive to calculate.

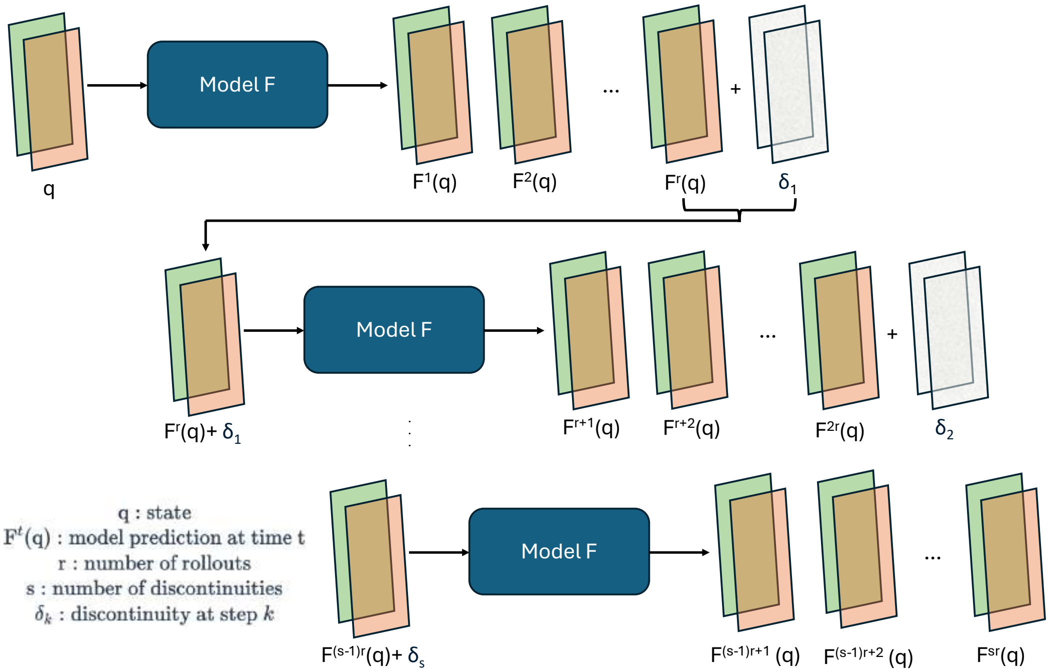

Figure 1: A schematic for MP optimization. The model F can be any autoregressive machine learning model. The intermediate discontinuity δ is introduced after every r rollouts.

As shown in Figure 1, we augment the ground truth loss(L_{GT}) with a penalty loss(L_P) as,

$$L_{MP} = \underbrace{\mathbb{E}_i\left[\sum_{t=1}^{t=sr}\left\|\lambda(t)(F_\theta^t(q_i) - S^t(q_i))\right\|\right]}_{L_{GT}} + \underbrace{\mu\sum_{k=1}^{s}\|\delta_k\|}_{L_P} \tag{4}$$

where r is the number of rollouts before introducing a discontinuity, s is the number of splits (discontinuities) and δs are the introduced discontinuities (learnable parameters). This is a modification from the previous implementations of MP [Chung and Freund, 2022, Chakraborty et al., 2024] where the intermediate states were learnable. We found the proposed approach to be more scalable for larger systems and stable during training. However, after every r rollouts we detached the computational graph before introducing the discontinuities δ, so that gradients are not propagated through the entire trajectory. The penalty strength μ is a hyperparameter that is gradually increased to achieve continuity in time. It is typically started with a very low value(10^{-5} in our experiments) and then gradually increased. We also start with a single rollout (r=1) and a single discontinuity (s=1) in the trajectory which are gradually increased to learn longer trajectoies. Further details on tuning the hyperparameters of MP method are provided in Chakraborty et al. [2024]. The loss is backpropagated through the computational graph to compute the gradients with respect to the set of learnable parameters θs and δs using automatic differentiation. The intermediate discontinuities are introduced only in training and not used in inference. Techniques like the Pushforward Trick [Brandstetter et al., 2022] and weighting trajectories that are closer in time can also be used with the MP optimization. Any other improvement like the Maximum Mean Discrepancy (MMD) loss introduced in Schiff et al. [2024] can be easily extended to MP method for further improvement. However, we note that the these additional invariant statistics based losses add significant computational overhead (requiring long-term integration for each gradient computation). The MP approach seeks to improve on standard autoregressive model training without this overhead.

3 Results

3.1 Kolmogorov Flow

This section evaluates the performance of our proposed framework on two-dimensional homogeneous isotropic turbulence driven by Kolmogorov forcing, governed by the incompressible Navier-Stokes equations. These experiments aim to assess MP optimization's capabilities for improving performance of the Fourier Neural Operator. Forced two-dimensional turbulence, a classic example of chaotic dynamics, has become a standard benchmark for ML methods used in dynamical system prediction

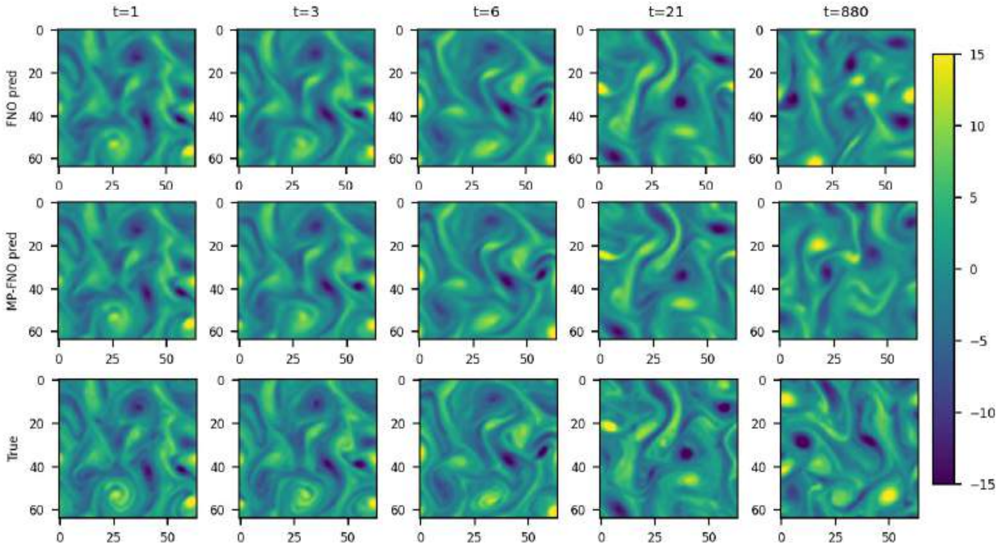

Figure 2: Vorticity of 2D Kolmogorov flow from predicted velocity fields. 't' here is the rollout step of the model.

(Stachenfeld et al. [2021], Brandstetter et al. [2022], Schiff et al. [2024]). The Reynolds number $Re = 10^5$ chosen for this study. The initial condition is a randomly generated divergence-free velocity field Kochkov et al. [2021]. For more details on dataset construction, refer to the work by Shankar et al. [2023]. The trajectories are temporally sub-sampled empirically after flow reaches the chaotic regime to guarantee sufficient separation between snapshots. It can be observed in Figure 2 that the the predictions match the ground truth closely in the starting timesteps and then diverge as a property of chaotic system. However, both FNO and MP-FNO shows no sign of instability even after a high number of autoregressive rollout. The MP optimization clearly improves upon the vanilla FNO for an invariant metric - the energy spectrum, and the correlation with DNS as shown in Figure 3. The latter also demonstrates how the MP optimization improves accuracy with greater integration duration.

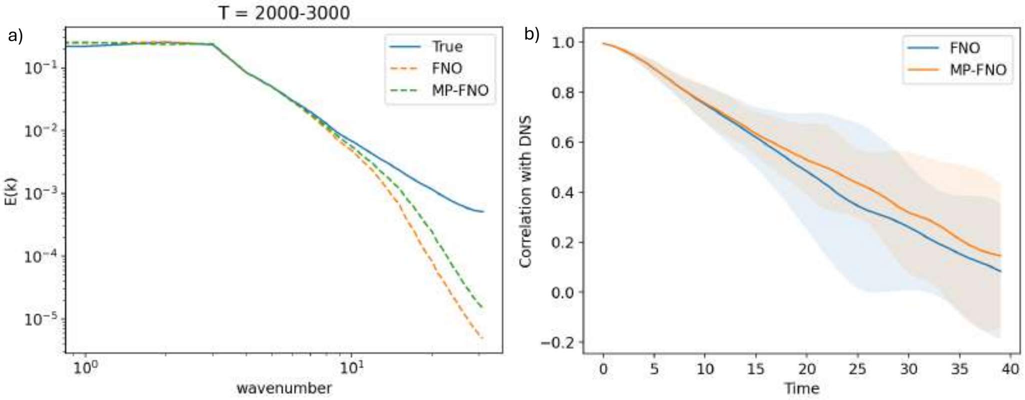

Figure 3: A comparison between FNO and MP-FNO for (a) angle-averaged total kinetic energy spectrum and (b) correlation with DNS. In (a) we check the performance for an invariant statistic, and for (b) we assess how the MP FNO technique improves accuracy with forecast duration compared to vanilla FNO.

3.2 Ocean Reanalysis Data

In this experiment we implement the MP algorithm to predict the sea-surface height (SSH), longitudinal (SSU), and meridional (SSV) velocities of the northwest Atlantic Ocean's western boundary extending from 92°W into the Atlantic 75°W in the Gulf of Mexico (GoM). For this, we have used the GLORYS version 4 [Garcia and Brown, 2021] reanalysis dataset, which is an eddy-permitting dataset at $\frac{1}{12}^{\circ}$ (8 Km). The training data (available daily) is temporally sub-sampled by a factor of 3 to keep sufficient distinction between the snapshots. We implement the MP algorithm with a UNET

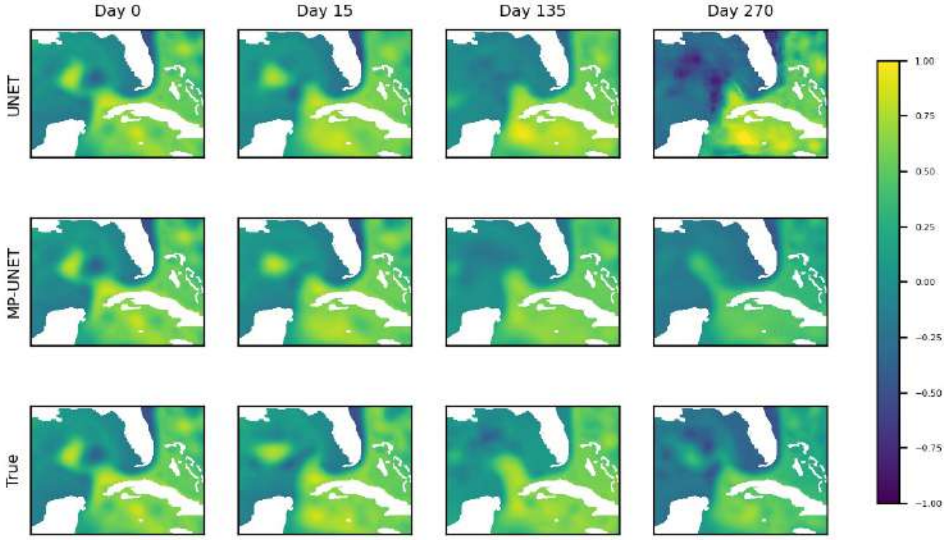

Figure 4: Prediction performance of UNET and MP-UNET for the GoM LCE shedding event: Eddy Sverdrup

[Ronneberger et al., 2015] architecture and compare the predictions for a test (unseen) time period of a major GoM Loop Current Eddies (LCE) shedding event: Eddy Sverdrup (Jul 2019-Jan 2020).

Figure 4 shows that both UNET and MP-UNET captures the dynamics of the data accurately for the time-period of the eddy event. We also found out that the vanilla UNET shows signs of instability after longer periods of time whereas MP optimization makes it stable for over 270 days while testing. However, to delve deeper into the results we compare root-mean-square error (RMSE) from ground truth for the model predictions. MP-UNET performs the best in long term as evident from Figure 5.

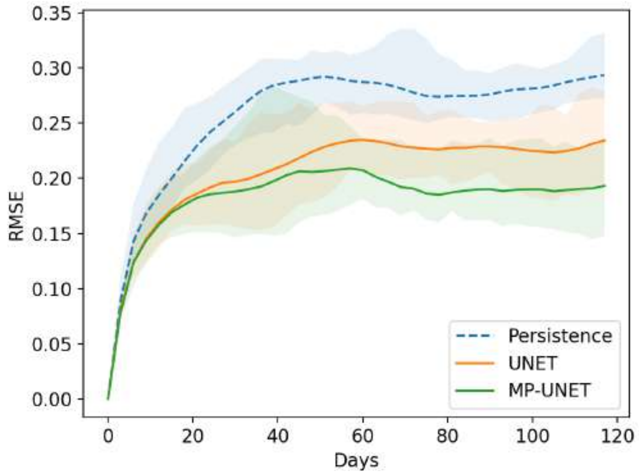

Figure 5: RMSE comparison between UNET, MP-UNET and Persistence. Persistence is an elementary model used to compare the performance of other models. It assumes that the weather is static and the initial condition itself is the forecast.

This also demonstrates the potential for the MP technique to improve neural forecasting applications in real-world use cases, for example in the earth sciences [Kashinath et al., 2021].

4 Conclusion

This paper focuses on the challenges posed by the long-term prediction of chaotic systems. Our proposed method provides a modified extension of the multi-step penalty(MP) optimization framework to a broader class of deep learning models such as Fourier Neural Operators and UNETs. We demonstrate its advantage by forecasting challenging chaotic systems such as high Reynolds number 2D turbulence and the Gulf Stream ocean reanalysis dataset. The MP optimization based architectures show more stability in the long term and is more accurate in short term compared to their vanilla counterparts without any significant overhead in computational cost. This work contributes to the field of data-driven modeling of chaotic systems and opens new avenues to explore and gain insight into complex natural phenomena.

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357. This research was funded in part and used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. RM and DC acknowledge support from U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research grant DOE-FOA-2493: “Data-intensive scientific machine learning”, and computational support from Penn-State Institute for Computational and Data Sciences.

References

Patrick J Blonigan, Steven A Gomez, and Qiqi Wang. Least squares shadowing for sensitivity analysis of turbulent fluid flows. In *52nd Aerospace Sciences Meeting*, page 1426, 2014.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural pde solvers. *arXiv preprint arXiv:2202.03376*, 2022.

Dibyajyoti Chakraborty, Seung Whan Chung, and Romit Maulik. Divide and conquer: Learning chaotic dynamical systems with multistep penalty neural ordinary differential equations. *arXiv preprint arXiv:2407.00568*, 2024.

Ashesh Chattopadhyay, Michael Gray, Tianning Wu, Anna B Lowe, and Ruoying He. Oceannet: A principled neural operator-based digital twin for regional oceans. *Scientific Reports*, 14(1):21181, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Seung Whan Chung and Jonathan B Freund. An optimization method for chaotic turbulent flow. *Journal of Computational Physics*, 457:111077, 2022.

Maria Garcia and Robert Brown. *Introduction to Neuromorphic Computing*. TechBooks Publishing, New York, 2021. ISBN 978-1234567890.

Haiwen Guan, Troy Arcomano, Ashesh Chattopadhyay, and Romit Maulik. Lucie: A lightweight uncoupled climate emulator with long-term stability and physical consistency for o (1000)-member ensembles. *arXiv preprint arXiv:2405.16297*, 2024.

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? *Advances in neural information processing systems*, 31, 2018.

Karthik Kashinath, M Mustafa, Adrian Albert, JL Wu, C Jiang, Soheil Esmaeilzadeh, Kamyar Azizzadenesheli, R Wang, Ashesh Chattopadhyay, A Singh, et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194):20200093, 2021.

Ryan Keisler. Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*, 2022.

Dmitrii Kochkov, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118 (21), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2101784118. URL `https://www.pnas.org/content/118/21/e2101784118`.

Daniel J Lea, Myles R Allen, and Thomas WN Haine. Sensitivity analysis of the climate of a chaotic system. *Tellus A: Dynamic Meteorology and Oceanography*, 52(5):523–532, 2000.

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.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Markov neural operators for learning chaotic systems. *arXiv preprint arXiv:2106.06898*, pages 2–3, 2021.

Alec J Linot, Joshua W Burby, Qi Tang, Prasanna Balaprakash, Michael D Graham, and Romit Maulik. Stabilized neural ordinary differential equations for long-time forecasting of dynamical systems. *Journal of Computational Physics*, 474:111838, 2023.

George Philipp, Dawn Song, and Jaime G Carbonell. The exploding gradient problem demystified-definition, prevalence, impact, origin, tradeoffs, and solutions. *arXiv preprint arXiv:1712.05577*, 2017.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

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. *arXiv preprint arXiv:2402.04467*, 2024.

Varun Shankar, Vedant Puri, Ramesh Balakrishnan, Romit Maulik, and Venkatasubramanian Viswanathan. Differentiable physics-enabled closure modeling for burgers' turbulence. *Machine Learning: Science and Technology*, 4(1):015017, 2023.

Kimberly Stachenfeld, Drummond B Fielding, Dmitrii Kochkov, Miles Cranmer, Tobias Pfaff, Jonathan Godwin, Can Cui, Shirley Ho, Peter Battaglia, and Alvaro Sanchez-Gonzalez. Learned coarse models for efficient turbulence simulation. *arXiv preprint arXiv:2112.15275*, 2021.