EXPLORING DESIGN CHOICES FOR AUTOREGRESSIVE DEEP LEARNING CLIMATE MODELS

Florian Gallusser, Simon Hentschel, Anna Krause, Andreas Hotho
Data Science Chair, Center for Artificial Intelligence and Data Science (CAIDAS),
University of Würzburg
{gallusser,hentschel,anna.krause, hotho}@informatik.uni-wuerzburg.de

ABSTRACT

Deep Learning (DL) models have achieved state-of-the-art performance in medium-range weather prediction (MWP) but often fail to maintain physically consistent rollouts beyond 14 days. In contrast, a few atmospheric models demonstrate stability over decades, though the key design choices enabling this remain unclear. This study quantitatively compares the long-term stability of three prominent DL-MWP architectures — FourCastNet, SFNO, and ClimaX — trained on ERA5 reanalysis data at $5.625°$ resolution. We systematically assess the impact of autoregressive training steps, model capacity, and choice of prognostic variables, identifying configurations that enable stable 10-year rollouts while preserving the statistical properties of the reference dataset. Notably, rollouts with SFNO exhibit the greatest robustness to hyperparameter choices, yet all models can experience instability depending on the random seed and the set of prognostic variables.[1]

1 INTRODUCTION

Over the past few years autoregressive Deep Learning (*DL*) models have emerged that are en par with physics-based state-of-the-art medium range weather prediction systems while only requiring a fraction of the computational costs for inference (Lam et al., 2023; Bi et al., 2023; Price et al., 2025). Given an initial state of the atmosphere, those models iteratively predict future states for a set of prognostic variables, which is commonly referred to as autoregressive rollout. However, the forecasts of popular weather models tend to degrade beyond 14 days and result in prediction of system states with physically unplausible distributions (Karlbauer et al., 2024a; McCabe et al., 2023; Chattopadhyay & Hassanzadeh, 2023). This instability poses a fundamental challenge when considering DL models for earth system simulations, where interactions between atmosphere, oceans, land, and ice are simulated over climate timescales, i.e., decades to hundreds of years. While simulations with physics-based earth system models are a foundational scientific tool to understand and project past, present, and future climate change Eyring et al. (2016), reducing their immense computational costs while maintaining physical accuracy is a key motivation for exploring DL-based alternatives.

Several DL approaches have demonstrated the ability to generate stable autoregressive trajectories for atmospheric dynamics over at least one year, leveraging either ERA5 reanalysis data (Hersbach et al., 2020) or outputs from atmospheric general circulation models (*GCMs*). Most of these methods employ specialized grid representations or architectures to mitigate distortions introduced by the regular latitude-longitude grid (Weyn et al., 2020; Karlbauer et al., 2024a; Bonev et al., 2023). Also, the complexity of the emulated GCM (Scher & Messori, 2019) or the selection of variables from the reanalysis dataset seem to play an important role for stability, since proposed models are often trained on a very limited set of prognostic variables (Guan et al., 2024; Weyn et al., 2020; Cresswell-Clay et al., 2024). However, different forward stepping schemes have been employed, such as predicting only the next timestep (Bonev et al., 2023; Karlbauer et al., 2024b) or predicting two timesteps at the same time (Weyn et al., 2020; Karlbauer et al., 2024a; Cresswell-Clay et al., 2024), which were both combined with multiple autoregressive training steps. The inclusion of top-of-the-atmosphere incident solar radiation as an external forcing variable has also been claimed to be a crucial factor for long-term stability (Guan et al., 2024). While first works that provided boundary

[1] Code will be made publicly available on publication.

conditions from the ocean just used annually repeating sea surface temperature (Watt-Meyer et al., 2023; Cachay et al., 2024), recent efforts have also successfully coupled the atmospheric models to basic data-driven ocean models (Wang et al., 2024; Cresswell-Clay et al., 2024). Despite these advancements, the fundamental design principles underpinning long-term stability remain relatively unclear, as existing studies primarily focus on demonstrating and evaluating stable models rather than systematically analyzing the factors that ensure stability.

A preliminary comparison of the long-term behavior of common weather model architectures was conducted by Karlbauer et al. (2024b), who demonstrated that even models lacking explicit spherical geometry and trained to predict a single future state per forward pass can achieve stable rollouts when trained on ERA5 data at a coarse resolution of $5.625°$. While their results qualitatively indicate stability over 50-year rollouts, a rigorous quantitative assessment is still missing.

Building on this foundation, our study quantitatively evaluates the long-term stability of three prominent DL architectures: FourCastNet, SFNO, and ClimaX. We systematically investigate fundamental design choices, including the number of autoregressive steps during training, model capacity, and the choice of prognostic variables. Using ERA5 reanalysis data at $5.625°$ resolution, we identify model configurations that allow stable 10-year rollouts while preserving the statistical properties of the reference dataset. Moreover, we highlight the geometry-aware SFNO architecture as the least sensitive to hyperparameter variations, yet we observe that all models can experience severe instability for some training random seeds.

2 Experimental Setup

Dataset For our study, we use the Weatherbench1 dataset (Rasp et al., 2020) at a horizontal resolution of $5.625°$ (64x32 grid points) and at a temporal resolution of 6 hours. Weatherbench1 was created by bilinear interpolation to a regular latitude-longitude grid from the ERA5 reanalysis dataset (Hersbach et al., 2020) and is available at an hourly temporal resolution covering the years 1979 to 2018. We train our models on data from 1979 to 2007, use the year 2008 for validation during training and keep 10 years from 2009 to 2018 for testing. During our experiments, we use two distinct sets of prognostic variables (see Appendix Table 2). The first set contains 33 prognostic variables based on Nguyen et al. (2023) minus relative humidity and the level $50\,\mathrm{hPa}$; the second set contains the eight prognostic variables from Karlbauer et al. (2024b). As a forcing variable, we use the top-of-the-atmosphere incident solar radiation (*TISR*) computed from the total solar irradiance from the recommended solar forcing datataset (Matthes et al., 2017) for CMIP6[2]. We employ z-score normalization to normalize each variable at each level by its global mean and standard deviation computed over the entire training set.

Models We experiment with three model architectures f_θ: (i) **ClimaX** (Nguyen et al., 2023) is based on a vision-transformer (Dosovitskiy et al., 2020) and employs a special variable tokenization and aggregation scheme, which was designed to use multiple physical quantities as input data. (ii) **FourCastNet** (*FCN*) (Pathak et al., 2022) is based on the Adaptive Fourier Neural Operator (Guibas et al., 2021) which is an efficient token mixer in frequency space that allows FCNs to scale efficiently to large spatial resolutions. (iii) **Spherical Neural Operators** (*SFNO*) (Bonev et al., 2023) generalizes Fourier Neural Operators (Li et al., 2020) to the sphere by replacing the discrete Fourier transform by the spherical harmonics transform.

Training We compute the loss for a single training example with initial condition at timestep t, given a tensor of K_p prognostic variables $X_t \in \mathbb{R}^{K_p \times W \times H}$, a tensor of K_c of constant variables $C \in \mathbb{R}^{K_c \times W \times H}$ and K_f initial forcing variables $F_t \in \mathbb{R}^{K_f \times W \times H}$ as

$$\mathcal{L}_t = \sum_{m=0}^{M-1} \mathcal{L}_t^m = \sum_{m=0}^{M-1} \sum_{k=1}^{K_p} \sum_{h=1}^{H} a_h \sum_{w=1}^{W} (X_{t+m\cdot\Delta t} + f_\theta(X_{t+m\cdot\Delta t}, F_{t+m\cdot\Delta t}, C) - X_{t+(m+1)\cdot\Delta t})^2$$

where M is the number of autoregressive steps, a_h is the area weight accounting for the area distortion introduced in the regular latitude-longitude grid of size (W, H) and Δt is equal to 6 hours.

[2]The code for computing TISR was adapted from the GraphCast repository `https://github.com/google-deepmind/graphcast`

We sample initial conditions at `06:00`, `12:00`, `18:00` and `00:00` UTC for all days in the training period which results in $42\,368$ training samples. Note that gradients are only computed after summing the loss $\mathcal{L}_t^m$ for all M-steps.

Experiment We perform a grid search over the following parameters: $f_\theta \in \{$ClimaX, FCN, SFNO$\}$, the number of prognostic variables $K_p \in \{8, 33\}$, the number of steps $M = \{1, 2, 4\}$ during training, the number of layers $L = \{4, 6, 8\}$ and the size of the hidden dimension $D = \{128, 256, 512\}$. We train each configuration with 10 different random seeds. Other hyperparameters were chosen based on Karlbauer et al. (2024b) and are documented in the Appendix in Table 1.

Evaluation After training, we initialize each model on the 1st of January 2009 `00:00` UTC, the first timestep of the test dataset, and compute an autoregressive rollout for 10 years. We compare the average over time of the predicted prognostic variables with the average over time of the target for the 10-year time period as a measure for long-term stability. As comparison metrics, we use the area-weighted mean squared error (RMSE). To obtain a single score for each model, we compare normalized predictions and targets, and average the per variable RMSE across the commonly evaluated variables `tas`, `uas`, `vas`, `ta850` and `zg500`. We refrain to rollout all configurations from multiple initial conditions since we observed in preliminary experiments that this does not effect the long-term behavior of the rollout. We consider climatology, i.e. the temporal average over the 29 years of the training data set, as the lower bound baseline method. The RMSE of the climatology indicates how much climate changed from the training period compared to the evaluation period, and since our model is not forced by essential climate drivers such as greenhouse gas emissions and the state of the ocean, we can not expect to outperform climatology with our models.

3 Results & Discussion

We present the area-weighted normalized RMSE scores averaged across the evaluated variables for all trained models with a fixed number of layers $L = 4$ in Fig. 1. However, the results discussed here generalize to configurations with $L = 6$ and $L = 8$ layers (see Appendix Fig. 2 and Fig. 3).

Our results reveal that all model architectures can yield 10-year rollouts with comparably low RMSE scores close to the RMSE of climatology (see Fig. 1). However, across all configurations, the total number of SFNO models with RMSE score close to climatology is substantially higher than for FCN and ClimaX. We verified that those rollouts do not only match well the temporal mean of the reference period, but also the temporal standard deviation (see Appendix Fig. 4). Hence, those models preserve important distributional characteristics of the climate they were trained for.

Generally, we observe that multiple autoregressive steps during training reduces the RMSE of the 10-year rollouts. This effect is very dominant for the SFNO trained on 33 prognostic variables: multi-step training is much more stable than one-step training. Also, FCN and ClimaX configurations with RMSE scores near climatology are all trained with a multi-step objective. However, multi-step training alone is not sufficient to obtain stable rollouts and choosing the right capacity of the model is crucial as well. ClimaX and FCN only achieve RMSE scores close to climatology, when trained with the small $D = 128$ or medium $D = 256$ hidden size across all number of layers. Contrarily, for SFNO models with $L = 4$ layers, we see a clear trend that increasing the hidden dimension decreases the RMSE score. We hypothesize that implicitly satisfied geometrical constraints in the SFNO architecture allow to increase the model's capacity while FCN and ClimaX overfit to noisy signals close to the poles when capacity is increased.

Furthermore, we see that even for configurations that achieve RMSE scores close to climatology, some random seeds lead to severe instability of the 10-year rollout, e.g., SFNO trained on 8 prognostic variables, with the 4-step training objective and with hidden dimension 256. We verified that this can not be traced back to a decrease in 1-step prediction accuracy. Additionally, we highlight that model configurations that are stable for one set of prognostic variables become unstable for the other set of prognostic variables. Interestingly, rollouts with ClimaX tend to become more stable with the larger set of prognostic variables, while contrarily the RMSE of rollouts with SFNO and FCN explodes with the larger set of prognostic variables for some configurations.

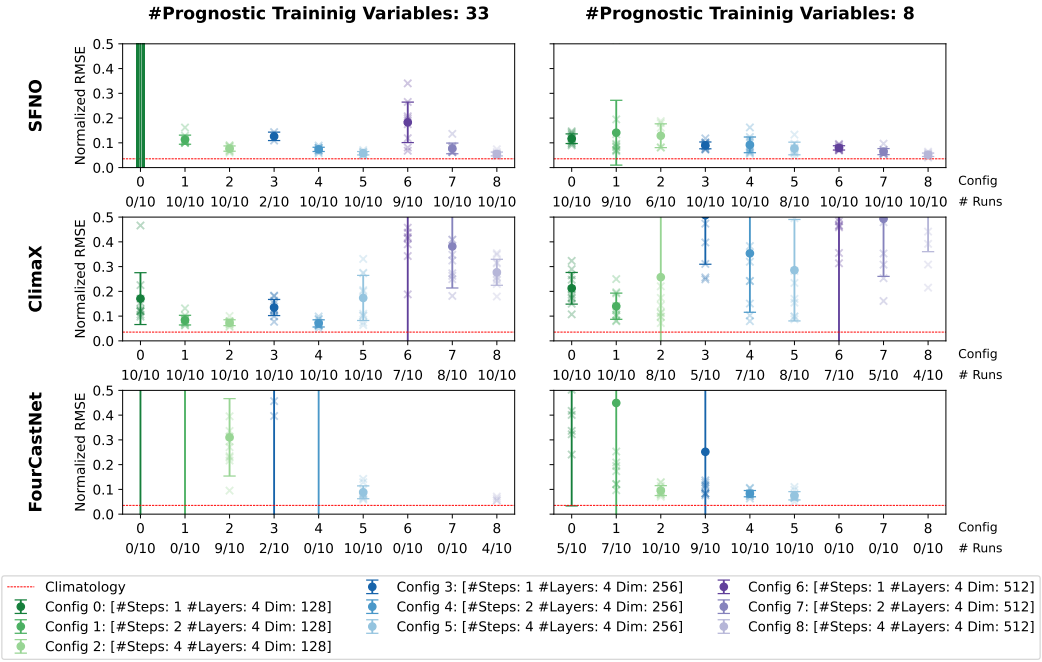

Figure 1: Area-weighted normalized RMSE scores averaged over the variables `tas`, `uas`, `vas`, `ta850` and `zg500` for model configurations $0-8$ with layer size $L = 4$. The mean (dots) and standard deviation (error bars) were computed across 10 seeds (crosses) with finite RMSE. For better readability, the y-axis is cut-off at 0.5 and the number of displayed runs out of 10 is shown on the x-axis.

Discussion While our results validate and quantitatively extend the findings of Karlbauer et al. (2024b) that also non-geometry-aware architectures can achieve long stable rollouts with minimal drift, we showed that stability of rollouts with SFNO are much less sensitive to the choice of hyperparameters. Our results further confirm that multi-step training, which has been shown to mitigate autoregressive error growth in other experimental settings such as weather prediction (Lam et al., 2023), hybrid numerical-DL climate models (Kochkov et al., 2024), and PDE learning (Brandstetter et al., 2021), also enhances long-term stability in climate projections. Also Watt-Meyer et al. (2023) report performance variations of up to a factor of four with different training seeds for SFNO, but we observe that rollouts with all tested architectures completely degenerate for some random seeds. We point out that our results were obtained on a relatively coarse resolution of $5.625°$ and it is unclear if they generalize to higher spatial resolutions. While we focus on stability as a necessary criterion for autoregressive DL climate models, we acknowledge that evaluating the overall quality of a climate model is much more complex.

4 Conclusion & Future Work

With this study, we provide insights into basic hyperparameter choices for training DL models on reanalysis data which help to build autoregressive atmospheric models suitable for climate timescales. While models based on the geometry-aware SFNO architecture yield stable rollouts across many tested hyperparameter configurations, we find that non geometry-aware architectures can yield rollouts of equal stability with the right hyperparameter settings. Furthermore, we show that the multi-step objective helps to stabilize long rollouts across all architectures. The sensitivity to the training random seed and the used set of prognostic variables illustrates that further empirical and theoretical analysis is needed to create a more profound understanding of criteria that influence stability of long rollouts. Future work should continue to study the design choices which might impact the behavior of long autoregressive rollouts, e.g., optimization, regularization techniques and the number of input/output timesteps. While long-term stability of autoregressive atmospheric DL models

is a prerequisite for achieving fully data-driven climate projections, integrating the external forcings that lead to changing climate and the coupling to other earth system components are necessary steps towards this goal.

ACKNOWLEDGEMENT

This research was conducted in the BigData@Geo 2.0 project which is co-financed by the European Regional Development Fund (ERDF). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project ID b214cb. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683.

REFERENCES

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023. doi: 10.1038/s41586-023-06185-3.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 2806–2823. PMLR, July 2023.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message Passing Neural PDE Solvers. In *International Conference on Learning Representations*, October 2021.

Salva Rühling Cachay, Brian Henn, Oliver Watt-Meyer, Christopher S. Bretherton, and Rose Yu. Probablistic Emulation of a Global Climate Model with Spherical DYffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024.

Ashesh Chattopadhyay and Pedram Hassanzadeh. Long-term instabilities of deep learning-based digital twins of the climate system: The cause and a solution. (arXiv:2304.07029), April 2023. doi: 10.48550/arXiv.2304.07029.

Nathaniel Cresswell-Clay, Bowen Liu, Dale Durran, Andy Liu, Zachary I. Espinosa, Raul Moreno, and Matthias Karlbauer. A Deep Learning Earth System Model for Stable and Efficient Simulation of the Current Climate. (arXiv:2409.16247), September 2024. doi: 10.48550/arXiv.2409.16247.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, October 2020.

Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, May 2016. doi: 10.5194/gmd-9-1937-2016.

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:2405.16297), May 2024. doi: 10.48550/arXiv.2405.16297.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient Token Mixing for Transformers via Adaptive Fourier Neural Operators. In *International Conference on Learning Representations*, October 2021.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana

Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, May 2020. doi: 10.1002/qj.3803.

Matthias Karlbauer, Nathaniel Cresswell-Clay, Dale R. Durran, Raul A. Moreno, Thorsten Kurth, Boris Bonev, Noah Brenowitz, and Martin V. Butz. Advancing Parsimonious Deep Learning Weather Prediction Using the HEALPix Mesh. *Journal of Advances in Modeling Earth Systems*, 16(8):e2023MS004021, July 2024a. doi: 10.1029/2023MS004021.

Matthias Karlbauer, Danielle C. Maddix, Abdul Fatir Ansari, Boran Han, Gaurav Gupta, Yuyang Wang, Andrew Stuart, and Michael W. Mahoney. Comparing and Contrasting Deep Learning Weather Prediction Backbones on Navier-Stokes and Atmospheric Dynamics. (arXiv:2407.14129), October 2024b. doi: 10.48550/arXiv.2407.14129.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, August 2024. doi: 10.1038/s41586-024-07744-y.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, December 2023. doi: 10.1126/science.adi2336.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. In *International Conference on Learning Representations*, October 2020.

Katja Matthes, Bernd Funke, Monika E. Andersson, Luke Barnard, Jürg Beer, Paul Charbonneau, Mark A. Clilverd, Thierry Dudok de Wit, Margit Haberreiter, Aaron Hendry, Charles H. Jackman, Matthieu Kretzschmar, Tim Kruschke, Markus Kunze, Ulrike Langematz, Daniel R. Marsh, Amanda C. Maycock, Stergios Misios, Craig J. Rodger, Adam A. Scaife, Annika Seppälä, Ming Shangguan, Miriam Sinnhuber, Kleareti Tourpali, Ilya Usoskin, Max van de Kamp, Pekka T. Verronen, and Stefan Versick. Solar forcing for CMIP6 (v3.2). *Geoscientific Model Development*, 10 (6):2247–2302, June 2017. doi: 10.5194/gmd-10-2247-2017.

Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards Stability of Autoregressive Neural Operators. *Transactions on Machine Learning Research*, June 2023.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 25904–25938. PMLR, July 2023.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. (arXiv:2202.11214), February 2022. doi: 10.48550/arXiv.2202.11214.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044): 84–90, January 2025. doi: 10.1038/s41586-024-08252-9.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020. doi: 10.1029/2020MS002203.

Sebastian Scher and Gabriele Messori. Weather and climate forecasting with neural networks: Using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, 12(7):2797–2809, July 2019. doi: 10.5194/gmd-12-2797-2019.

Chenggong Wang, Michael S. Pritchard, Noah Brenowitz, Yair Cohen, Boris Bonev, Thorsten Kurth, Dale Durran, and Jaideep Pathak. Coupled Ocean-Atmosphere Dynamics in a Machine Learning Earth System Model. (arXiv:2406.08632), June 2024. doi: 10.48550/arXiv.2406.08632.

Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K. Clark, Brian Henn, James Duncan, Noah D. Brenowitz, Karthik Kashinath, Michael S. Pritchard, Boris Bonev, Matthew E. Peters, and Christopher S. Bretherton. ACE: A fast, skillful learned global atmospheric model for climate prediction. (arXiv:2310.02074), December 2023. doi: 10.48550/arXiv.2310.02074.

Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. Improving Data-Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109, June 2020. doi: 10.1029/2020MS002109.

A Parameter Settings

We train each model configuration for 20 epochs at float32 precision and employ early stopping when the validation loss has not improved during the last 5 epochs. We use the Adam optimizer with an initial learning rate of 4×10^{-3} for FCN and ClimaX, and a learning rate of 1×10^{-3} for SFNO models. A smaller LR for SFNO was necessary as the multi-step training loss exploded at the start of training with the LR of 4×10^{-3} The LR rate was chosen based on the training protocol of Karlbauer et al. (2024b) accounting for training with the 4 times larger effective batch size of 64. The learning rate decays every epoch with a cosine annealing schedule towards 0. We clip the norm of gradients to a value of 0.001 similar to Karlbauer et al. (2024b). The following random seeds were used for training: 597, 1152, 1826, 3909, 6153, 5513, 5707, 9813, 9941, 9982. Model specific hyperparameters are listed in Table 1.

Table 1: Model Specific Hyperparameters

ClimaX	FourCastNet	SFNO
Patch size: (2,2)	Patch size: (1,1)	Scale Factor: 1
# Heads: 8	# Blocks: 4	Big Skip: False
Decoder Depth: 2	Use Pos. Embed.: False	Use Pos. Embed.: False
MLP Ratio: 4	MLP Ratio: 4	Use MLP: True
Drop Rate: 0.0	Drop Rate: 0.0	Grid: "Equiangular"
Drop Path: 0.0	Drop Path: 0.0	factorization: Null
	Hard Threshold Fraction: 1.0	Hard Threshold Fraction: 1.0
	Sparsity Threshold: 0.01	

Table 2: List of prognostic, constant, and forcing variables used as input variables.

Category	**Vars**	**Level**	**Description**
Prognostic (33 Variables)	tas	2 m	Near-surface temperature
	uas	10 m	Near-surface eastward wind comp.
	vas	10 m	Near-surface northward wind comp.
	ta	925, 850, 700, 600, 500, 250 hPa	Temperature
	zg	925, 850, 700, 600, 500, 250 hPa	Geopotential height
	hus	925, 850, 700, 600, 500, 250 hPa	Specific humidity
	ua	925, 850, 700, 600, 500, 250 hPa	Eastward wind comp.
	va	925, 850, 700, 600, 500, 250 hPa	Northward wind comp.
Prognostic (8 Variables)	tas	2 m	Near-surface temperature
	uas	10 m	Near-surface eastward wind comp.
	vas	10 m	Near-surface northward wind comp.
	ta	850 hPa	Temperature
	zg	1000, 700, 500, 300 hPa	Geopotential height
Constant	lsm	-	Binary land-sea representation
	orog	-	Surface elevation data
	lat	-	Latitude information as 2D grid
	lon	-	Longitude information as 2D grid
Forcing	tisr	-	Top-of-the-atmosphere incident solar radiation

B Additional Figures

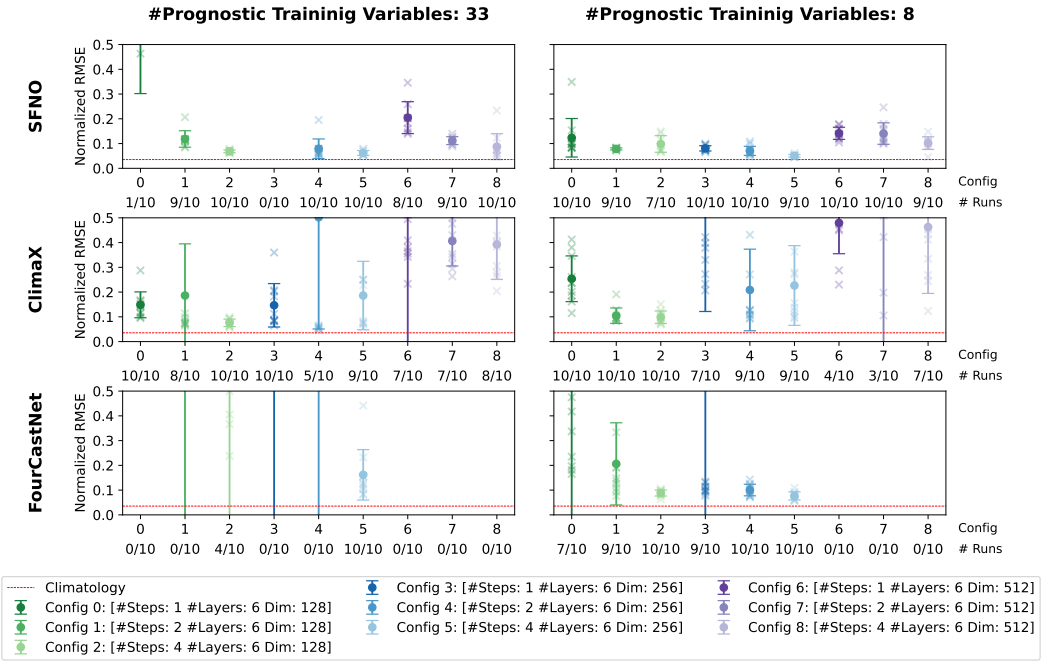

Figure 2: As Fig. 1 with $L = 6$: Area-weighted normalized RMSE scores averaged over the variables `tas`, `uas`, `vas`, `ta850` and `zg500` for model configurations $0 - 8$ with layer size $L = 6$. The mean (dots) and standard deviation (error bars) were computed across 10 seeds (crosses) with finite RMSE. For better readability, the y-axis is cut-off at 0.5 and the number of displayed runs out of 10 is shown on the x-axis.

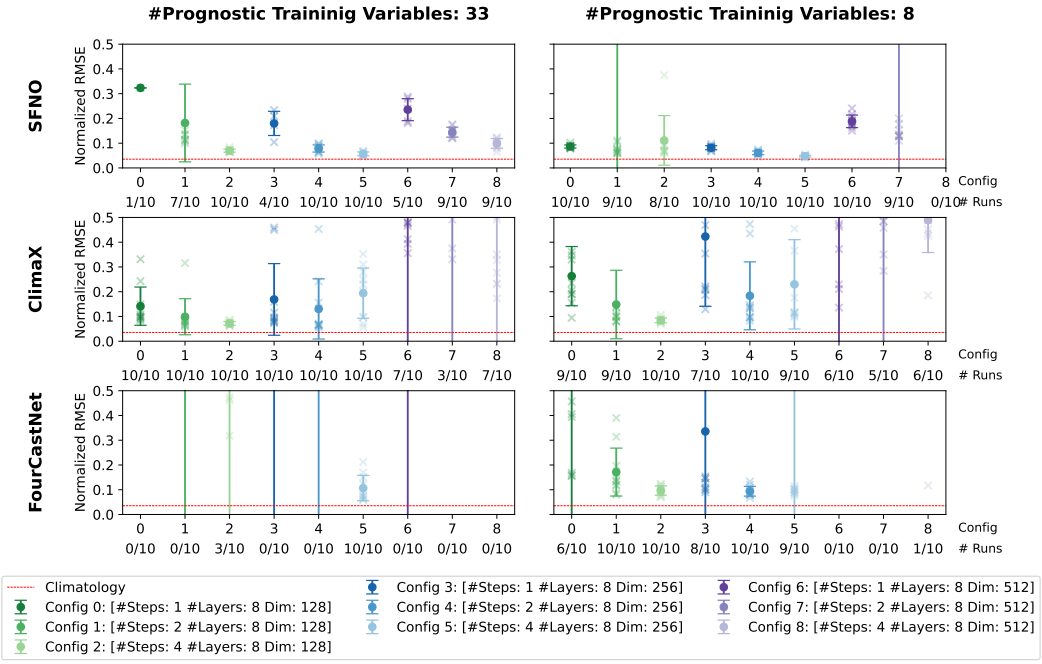

Figure 3: As Fig. 1 with $L = 8$: Area-weighted normalized RMSE scores averaged over the variables `tas`, `uas`, `vas`, `ta850` and `zg500` for model configurations $0 - 8$ with layer size $L = 8$. The mean (dots) and standard deviation (error bars) were computed across 10 seeds (crosses) with finite RMSE. For better readability, the y-axis is cut-off at 0.5 and the number of displayed runs out of 10 is shown on the x-axis.

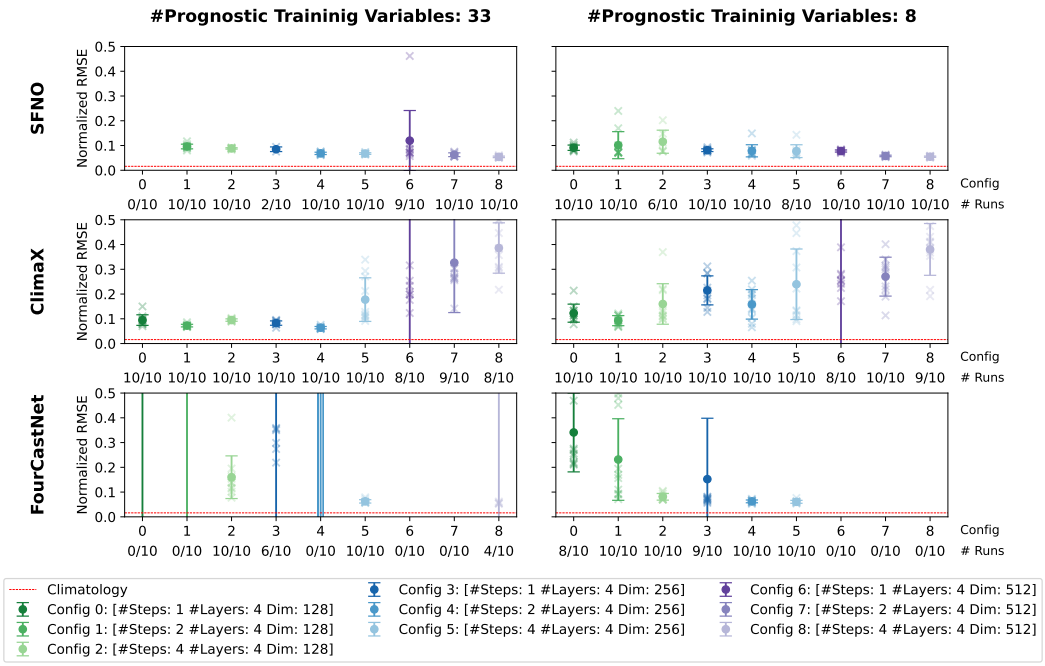

Figure 4: As Fig. 1, but temporal standard deviations are compared instead of temporal means: Area-weighted normalized RMSE scores for the temporal standard deviation averaged over the variables `tas`, `uas`, `vas`, `ta850` and `zg500` for model configurations $0-8$ with layer size $L=4$. The mean (dots) and standard deviation (error bars) were computed across 10 seeds (crosses) with finite RMSE. For better readability, the y-axis is cut-off at 0.5 and the number of displayed runs out of 10 is shown on the x-axis.

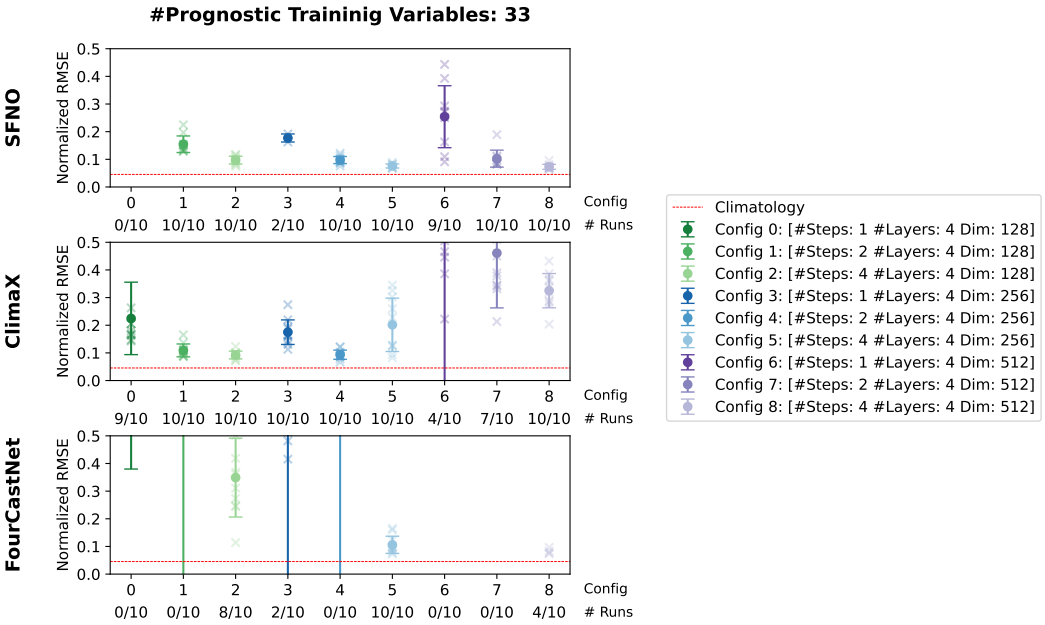

Figure 5: As Fig. 1 averaged over all 33 variables. Area-weighted normalized RMSE scores averaged over all 33 prognostic variables for model configurations $0-8$ with layer size $L=4$. The mean (dots) and standard deviation (error bars) were computed across 10 seeds (crosses) with finite RMSE. For better readability, the y-axis is cut-off at 0.5 and the number of displayed runs out of 10 is shown on the x-axis.

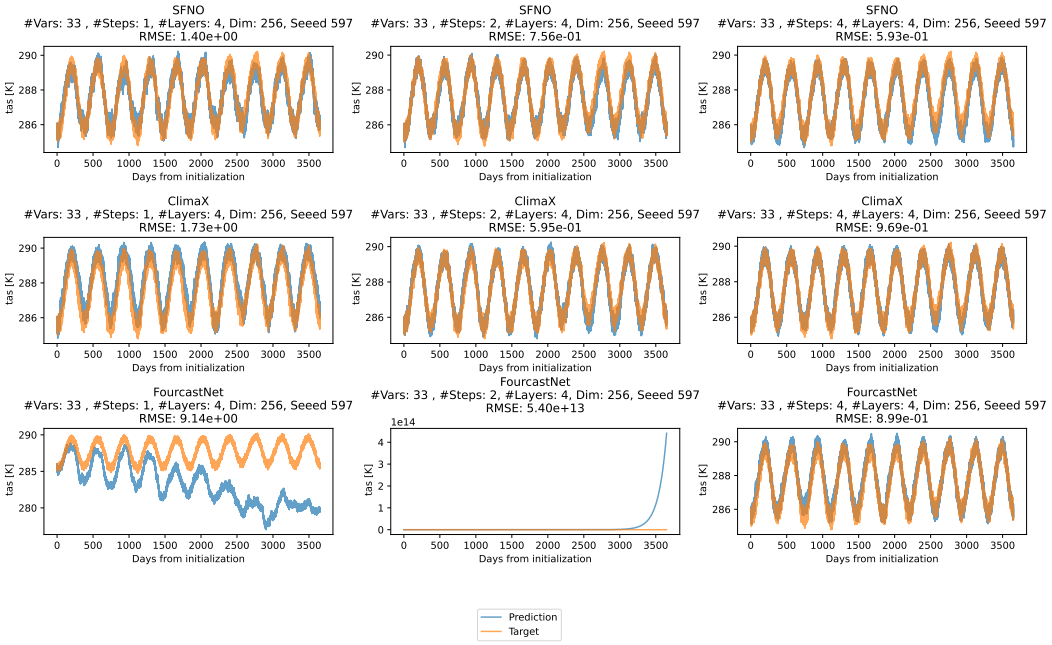

Figure 6: Near-surface temperature timeseries for selected models and ERA5 data that show the area-weighted global means for the 10-year rollout period. The RMSE score for the shown variable (not normalized) is indicated above each plot.

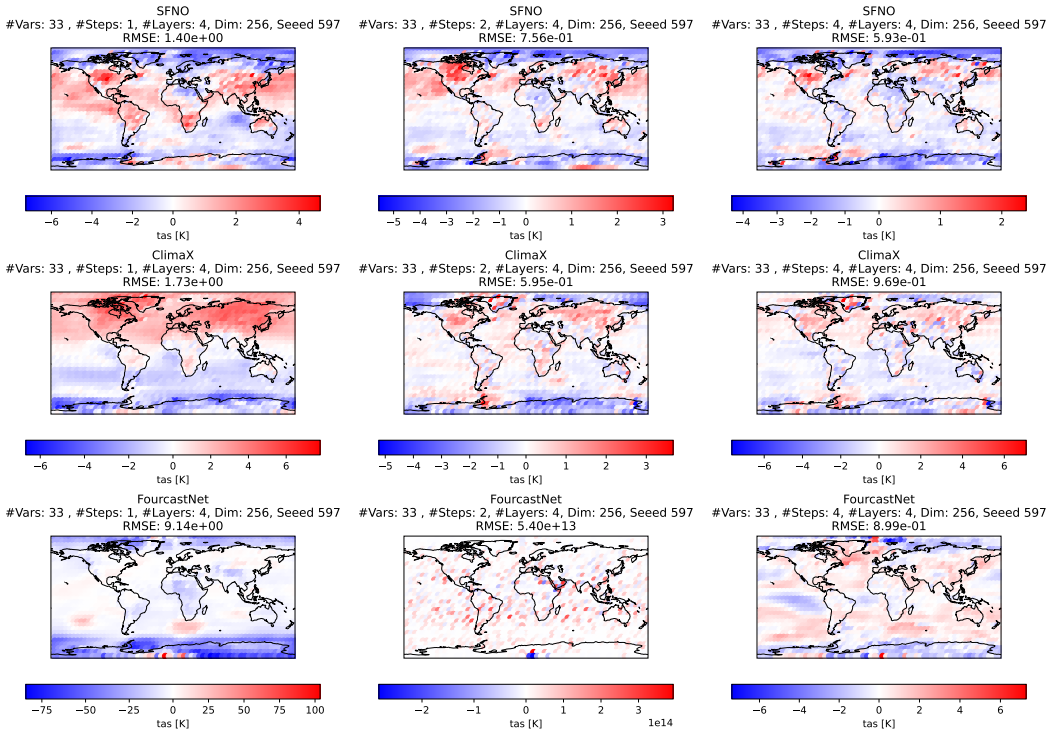

Figure 7: Near-surface temperature maps for selected models that show the difference of the temporal means of the 10-year rollouts minus the the temporal mean of ERA5 data. The RMSE score for the shown variable (not normalized) is indicated above each plot.

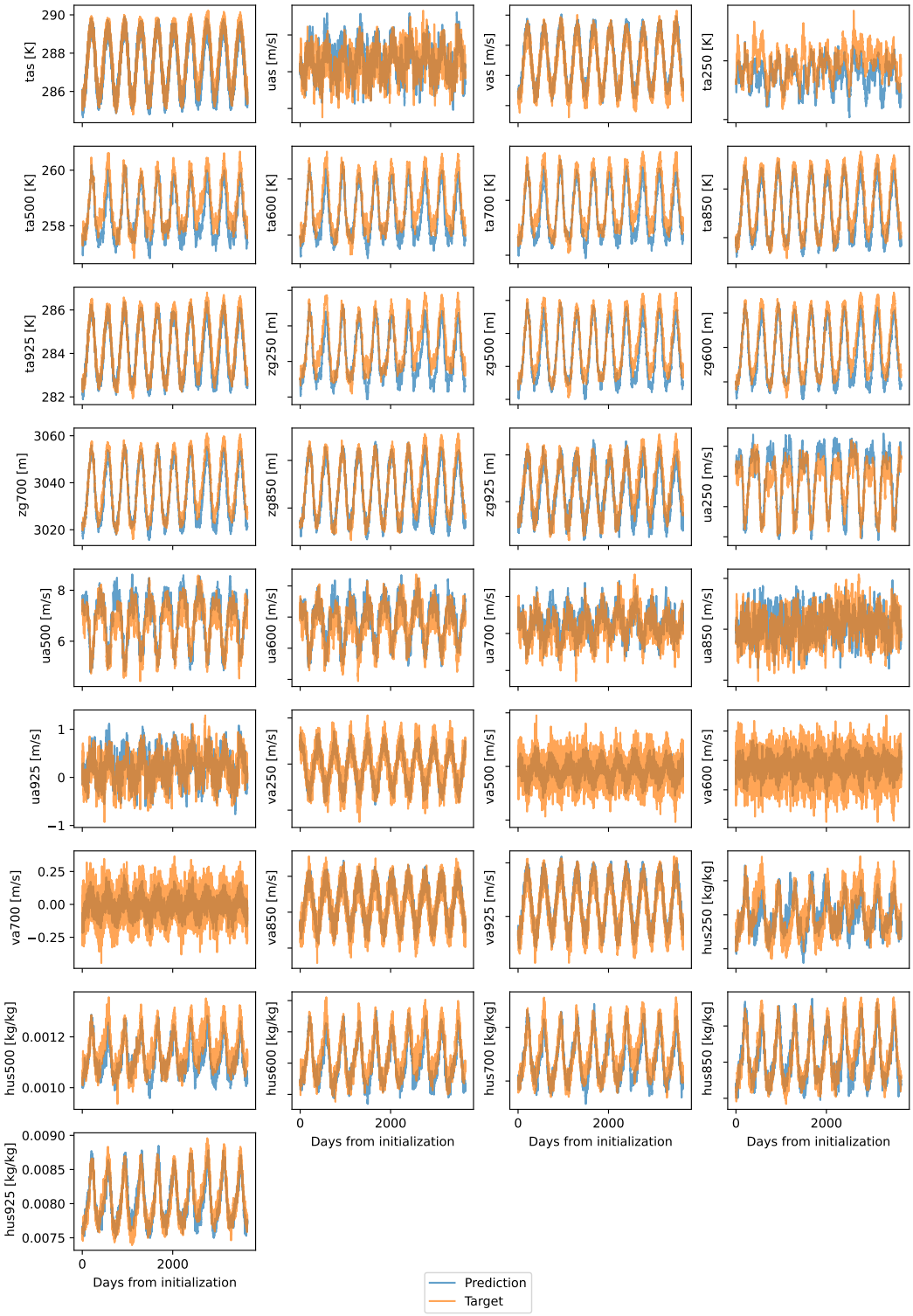

Figure 8: Timeseries of all variables for the SFNO model with 4-steps autoregressive training, 4 layers, hidden dimension of 512 and random seed 597 that show the area-weighted global means for the 10-year rollout.

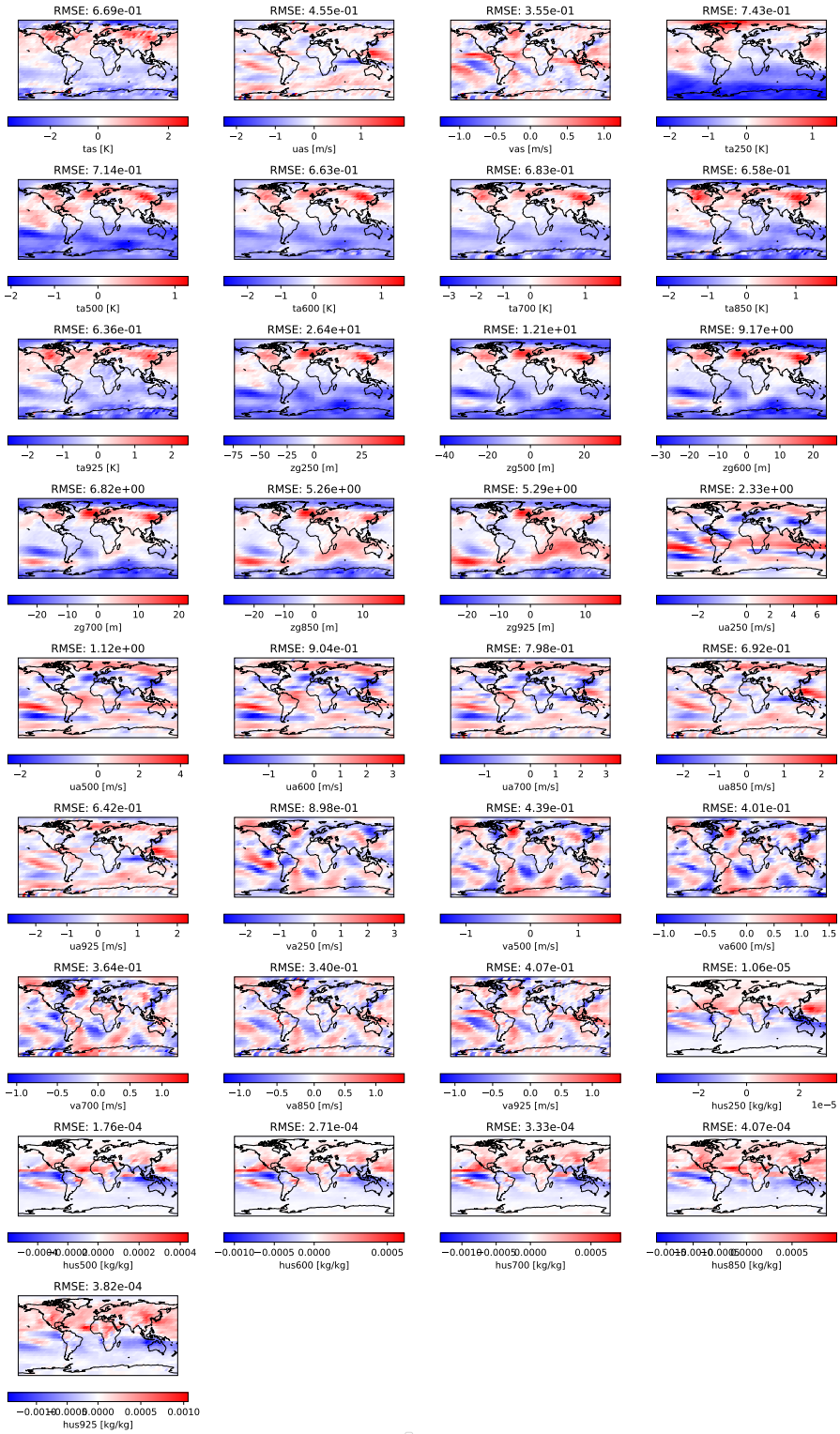

Figure 9: Maps of all variables for the SFNO model with 4-step autoregressive training, 4 layers, hidden dimension of 512 and random seed 597 that show the difference of the temporal means of the 10-year rollout minus the temporal mean of ERA5 data. The RMSE score for the shown variable (not normalized) is indicated above each plot.