Can Artificial Intelligence Global Weather Forecasting Models Capture Extreme Events? A Case Study of the 2022 Pakistan Floods

Rodrigo Almeida
Applied Machine Learning Group
Fraunhofer Heinrich-Hertz Institute
10587 Berlin, Germany
rodrigo.almeida@hhi.fraunhofer.de

Noelia Otero
Applied Machine Learning Group
Fraunhofer Heinrich-Hertz Institute
10587 Berlin, Germany

Miguel-Ángel Fernández-Torres
Signal Theory and Communications Dept.
Universidad Carlos III de Madrid (UC3M)
28911 Leganés, Madrid, Spain

Jackie Ma
Applied Machine Learning Group
Fraunhofer Heinrich-Hertz Institute
10587 Berlin, Germany

Abstract

Climate change is increasing extreme events. Despite advances in artificial intelligence-based global weather forecasting, most benchmarks remain deterministic, which limits uncertainty representation and extremes prediction skill assessment. This study examines how three state-of-the-art deterministic data-driven models, namely FourCastNet v2/SFNO, GraphCast, and FuXi, respond to input perturbations, evaluating the corresponding ensembles in forecasting extremes. 50-member ensembles are created using perturbation methods based on spherical Gaussian noise, hemispherical centered bred vectors, and huge ensembles. Focusing on August 2022, when devastating floods hit Pakistan, we compare our ensembles against deterministic ERA5-initialized forecasts and the ECMWF Integrated Forecasting System Ensemble for numerical weather prediction. While the huge ensembles method outperforms those based on Gaussian noise and hemispherical centered bred vectors in detecting the associated extreme precipitation event, all models still underperform the numerical weather model, suggesting promising research avenues.

1 Introduction

Climate change is increasing the frequency and severity of extreme weather events [1, 2], making timely, reliable forecasts and earlywarning systems essential for climate adaptation, especially for low-income and developing countries [3]. Recent developments in Artificial Intelligence-based Weather Prediction (AIWP) architectures using transformers [4–7], Fourier neural operators [8, 9] and Graph Neural Networks (GNNs) [10, 11] have pushed weather forecasting skill over what is considered state of the art for Numerical Weather Prediction (NWP) models, which conversely rely on complex equations and require expensive

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

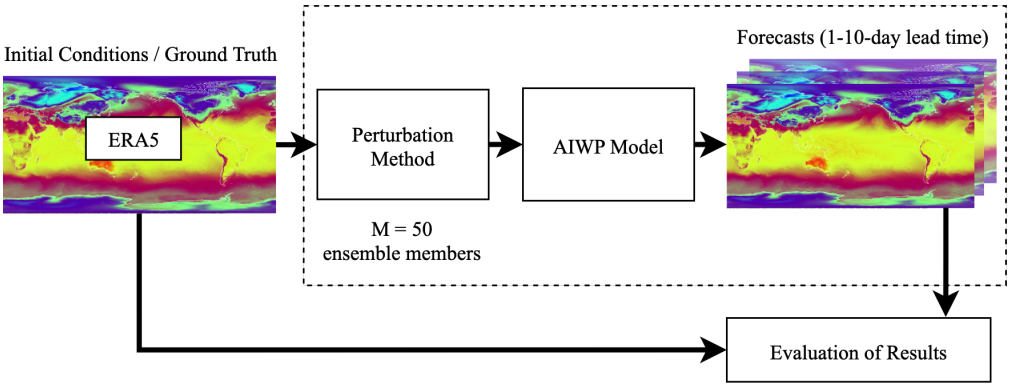

Figure 1: Overview of the methodology followed in this study, including the initial condition perturbation, forecasting with AIWP models, and evaluation of results stages.

computational resources. The Artificial Intelligence/Integrated Forecasting System (AIFS) by ECMWF alongside NWP has marked a significant milestone [12]. In addition, generative approaches, such as GenCast [13], can produce distributions of potential future weather states. Similarly, SEEDS [14] generates large ensembles based on limited NWP forecasts.

Given the inherently chaotic nature of the atmosphere, it is crucial to account for uncertainties in both the initial conditions and the models themselves. While early AIWP research focused on deterministic forecasts, a shift toward probabilistic methods is emerging, incorporating Uncertainty Quantification (UQ) and novel perturbation techniques. Indeed, UQ plays a crucial role in enhancing the reliability and effectiveness of weather forecasting models, particularly for extreme events. Uncertainty is generally categorized into two types: aleatoric (due to intrinsic randomness and noise within the data) and epistemic (arising from model limitations) [15]. Recent efforts have shown the potential of UQ applied to deterministic AIWP to produce probabilistic weather forecasts [16–18]. Specifically, ensemble forecasting addresses aleatoric uncertainty by generating multiple simulations of the initial conditions to represent a range of possible outcomes, forming the basis for probabilistic forecasts [19].

The objective of this study is to explore how different AIWP models respond to input perturbations in extreme weather forecasting, specifically addressing how to create probabilistic forecasts based on ensembles from deterministic models and how these ensembles handle UQ. To the authors' knowledge, no such ensemble comparison has been conducted, despite a prior analysis specifically made for deterministic forecasts [20]. Taking the August 2022 Pakistan floods driven by extreme precipitation in the region as a case study [3, 21], we will investigate the capabilities and limitations of AIWP models in capturing extreme events.

2 Methodology: Forecasting Models and Perturbation Methods

Figure 1 illustrates the methodology followed in this study. First, starting from ERA5 reanalysis data [22] as initial conditions, we apply several perturbation methods, each of them generating $M = 50$ different initial states. Then, these states are provided to diverse 6-hourly resolution AIWP models, resulting in $M = 50$ different predictions for each, which constitute ensemble forecasting members for up to a 10-day lead time. The ensembles for all perturbation methods and AIWP models, along with the ECMWF Integrated Forecasting System Ensemble (ENS) [23], are finally evaluated at daily resolution (accumulated precipitation) against the ground-truth ERA5 to assess their forecasting capabilities. See Appendix A for further details.

For this study, we have selected AIWP models that fulfill the following conditions: i) They achieve state-of-the-art performance for deterministic medium-range weather forecasting; ii) Their code implementation and pre-trained weights are open source; and iii) They all forecast total precipitation. Based on these criteria and our literature review, we identified three architectures: 1) **FuXi**, a cascading U-Transformer and Swin Transformer V2 model [6]; 2) **GraphCast**, a GNN-based architecture [11]; and 3) **SFNO**, a Spherical Fourier Neural Operator-based model [8].

Then, building upon previous weather studies, we have considered three perturbation methods to evaluate their skill concerning extreme event prediction: 1) **Gaussian**, which adds spherical Gaussian noise to the input and is used here as a baseline [16]; 2) Hemispherical Centered Bred Vectors (**HCBV**) [24], which relies on bred vectors scaled separately for each hemisphere; and 3) Huge Ensembles (**HENS**), an initial condition perturbation method presented in [17, 18] similar to HCBV, but considering the prediction error to scale the perturbations. See Appendix B for additional details on these methods.

Last but not least, we have chosen three complementary evaluation metrics: 1) the well-known Root Mean Squared Error (**RMSE**) as a deterministic score; 2) the Continuous Ranked Probability Score (**CRPS**), which constitutes an overall probabilistic score; and 3) the Receiver Operating Characteristic Skill Score (**ROCSS**), a probabilistic score tailored to extremes (see Appendix C for metric definitions). Implementation details are in Appendix D.

3 Experimental Results and Discussion

August 2022 Extreme Precipitation in Pakistan: Consistent with previous studies, the analysis of ERA5 data reveals that the Pakistan region received extreme rainfall on multiple days throughout August 2022 [21, 25]. Deterministic AIWP failed to capture the severity of this event at 3-days lead time (see Appendix E). Considering the entire month of August, the model performances on this extreme event in terms of ROCSS values at the 99th percentile of the ERA5 1990-2020 climatology [26] is shown in Figure 2. ENS showcases the best performance by far, especially at longer lead times. Conversely, deterministic AIWPs (Zero) show the worst performance overall. SFNO stands out as the worst-performing model, potentially due to its precipitation prediction originating from a separate prognostic model. HENS perturbation method seems to best support the forecast skill for this event, especially for GraphCast. While this study mainly addresses aleatoric uncertainty, epistemic uncertainty becomes increasingly dominant over longer lead time horizons, and bred vectors also partly reflect this behavior, which can explain the corresponding performance decrease. Additionally, HENS enables the ROCSS to remain higher up to longer lead times across all models. Looking at the forecast day that

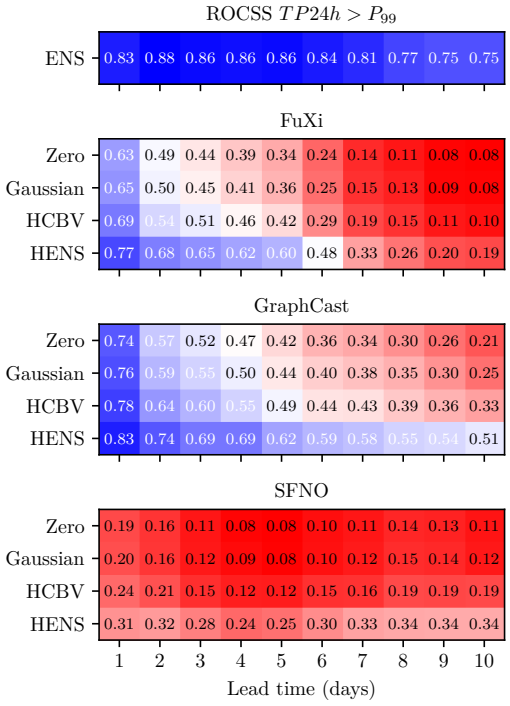

Figure 2: ROCSS values at the 99th percentile of the ERA5 1990-2020 climatology for daily accumulated precipitation, across 10 lead times, given the different AIWP models in our study and the Pakistan region in August 2022. The higher the ROCSS values, the better the performance.

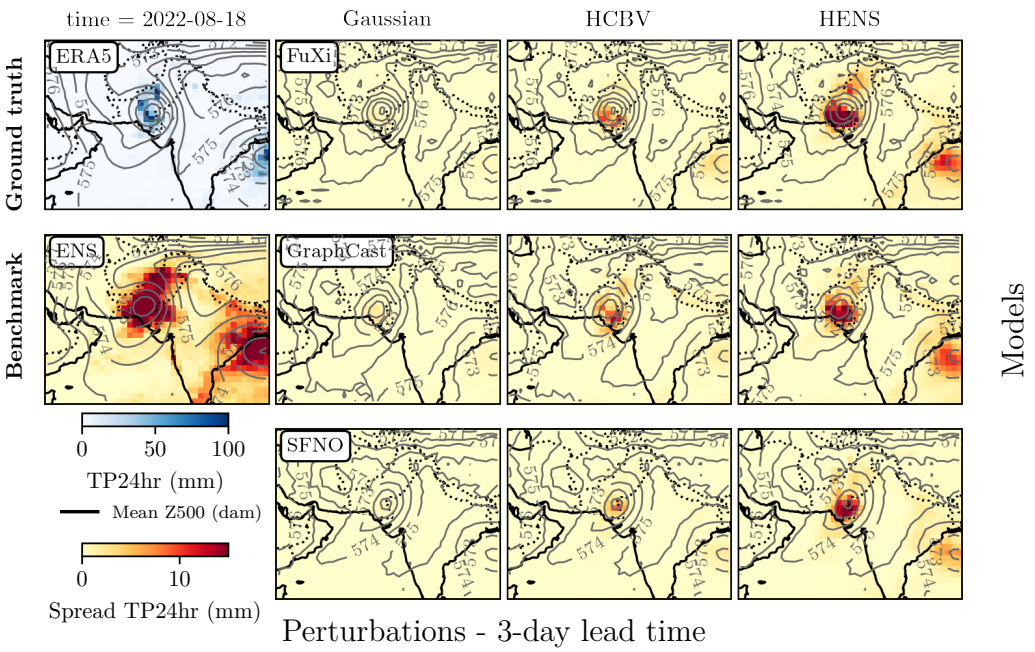

Figure 3: Daily accumulated precipitation spread for the different ensemble models (ENS, AIWPs) over Pakistan on 18th August 2022 for a 3-day lead time forecast. Ground-truth ERA5 spatial distribution for daily precipitation is shown in the top left corner. Daily average geopotential height at 500 hPa is also displayed as contour lines. None of the AIWP ensembles manage to produce a spread spatial distribution similar to the ground truth and the accurate one provided by ENS for this forecast, although FuXi HENS comes closer.

saw the highest amount of precipitation in
August (see Figure 3), FuXi together with
HENS seems to more closely match the ENS spread. Still, none of the AIWP models accurately display the uncertainty for the forecast of this event.

Global Quantitative Results: When looking at the computed metrics worldwide for August 2022 (see Appendix F), Gaussian fails to generate sufficient ensemble spread, leading to poorer performance. HENS and HCBV produce ensembles better aligned with ENS. FuXi consistently performs best in CRPS, with GraphCast closely matching it until around 5-days lead time, after which FuXi benefits from its cascading architecture. SFNO underperforms, particularly for precipitation. At longer lead times, all models converge in skill, with HENS remaining the most effective perturbation strategy.

4 Conclusions and Future Work

Early warning systems for extreme events are key to climate adaptation. The three deterministic AIWP models in this study are not able to capture the intensity and spatial distribution of the extreme precipitation in Pakistan in August 2022. Perturbing the inputs using HENS brings us closer to capturing the uncertainty of the forecasts for this event, but fails to reach the performance level of ENS, especially at longer lead times.

The global results broadly confirm the Pakistan case study: Gaussian perturbations are weakest, HENS is strongest, but none match ENS performance. Interestingly, GraphCast has an advantage over FuXi in the Pakistan region, not seen in the global context. Further

study using explainability methods could identify the root cause for such differences, as exemplified in [27]. In addition, alternative approaches, such as multi-model architecture ensembles [28], model checkpoint ensembles (as proposed in HENS), post-hoc uncertainty estimation methods [16], ensemble member selection and clustering procedures [29], and probabilistic AIWP models [13] are promising avenues to produce skillful data-driven extreme event forecasts.

Acknowledgments

This research has been supported by the European Unions Horizon Europe research and innovation program (EU Horizon Europe) project MedEWSa under grant agreement no. 101121192. M.-Á. F-T thanks the support of the grant for mobility stays from the Research and Transfer Program of Universidad Carlos III de Madrid, Spain, and acknowledges the computer resources provided by Artemisa, funded by the ERDF and Comunitat Valenciana, as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV).

References

[1] Vinod Thomas and Ramón López. Global Increase in Climate-Related Disasters, SSRN Scholarly Paper, Rochester, NY, November 2015. DOI: 10.2139/ssrn.2709331. Social Science Research Network: 2709331. (Visited on 08/13/2025).

[2] S. I. Seneviratne, X. Zhang, M. Adnan, W. Badi, C. Dereczynski, A. Di Luca, S. Ghosh, I. Iskandar, J. Kossin, S. Lewis, F. Otto, I. Pinto, M. Satoh, S. M. Vicente-Serrano, M. Wehner, and B. Zhou. Weather and climate extreme events in a changing climate. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1513–1766. Cambridge University Press, Cambridge, United Kingdom; New York, NY, USA, 2021. DOI: 10.1017/9781009157896.013.

[3] Peng Cui, Nazir Ahmed Bazai, Zou Qiang, Wang Jiao, Wang Yan, Qingsong Xu, Lei Yu, and Zhang Bo. Flood risk assessment with machine learning: insights from the 2022 Pakistan mega-flood and climate adaptation strategies. *npj Natural Hazards*, 2(1):42, May 2025. ISSN: 2948-2100. DOI: 10.1038/s44304-025-00096-1. (Visited on 08/12/2025).

[4] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. en. *Nature*, 619(7970):533–538, July 2023.

[5] Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. Fengwu: pushing the skillful global medium-range weather forecast beyond 10 days lead, 2023. arXiv: 2304.02948 [cs.AI]. URL: https://arxiv.org/abs/2304.02948.

[6] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast, October 2023. DOI: 10.48550/arXiv.2306.12873. arXiv: 2306.12873 [physics]. (Visited on 03/10/2025).

[7] Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, and Richard E Turner. End-to-end data-driven weather prediction. *Nature*, 641(8065):1172–1179, May 2025. DOI: https://doi.org/10.1038/s41586-025-08897-0.

[8] 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, June 2023. DOI: 10.48550/arXiv.2306.03838. arXiv: 2306.03838 [cs]. (Visited on 03/10/2025).

[9] Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '23, Davos, Switzerland. Association for Computing Machinery, 2023. ISBN: 9798400701900. DOI: 10.1145/3592979.3593412. URL: https://doi.org/10.1145/3592979.3593412.

[10] Ryan Keisler. Forecasting global weather with graph neural networks, 2022. arXiv: 2202.07575 [physics.ao-ph]. URL: https://arxiv.org/abs/2202.07575.

[11] 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. GraphCast: Learning skillful medium-range global weather forecasting. *Science*, December 2023. DOI: 10.1126/science.adi2336. (Visited on 02/06/2025).

[12] Simon Lang, Mihai Alexe, Mariana C. A. Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D. Dueben, Sara Hahner, Pedro Maciel, Ana Prieto-Nemesio, Cathal O'Brien, Florian Pinault, Jan Polster, Baudouin Raoult, Steffen Tietsche, and Martin Leutbecher. AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the Continuous Ranked Probability Score, December 2024. DOI: 10.48550/arXiv.2412.15832. arXiv: 2412.15832 [physics]. (Visited on 02/06/2025).

[13] 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. GenCast: Diffusion-based ensemble forecasting for medium-range weather, May 2024. DOI: 10.48550/arXiv.2312.15796. arXiv: 2312.15796 [cs]. (Visited on 06/10/2025).

[14] Lizao Li, Robert Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Generative emulation of weather forecast ensembles with diffusion models. *Sci. Adv.*, 10(13):eadk4489, March 2024. DOI: DOI:10.1126/sciadv.adk4489.

[15] Katherine Haynes, Ryan Lagerquist, Marie McGraw, Kate Musgrave, and Imme Ebert-Uphoff. Creating and Evaluating Uncertainty Estimates with Neural Networks for Environmental-Science Applications. *Artificial Intelligence for the Earth Systems*, 2(2), April 2023. ISSN: 2769-7525. DOI: 10.1175/AIES-D-22-0061.1. (Visited on 02/06/2025).

[16] Christopher Bülte, Nina Horat, Julian Quinting, and Sebastian Lerch. Uncertainty quantification for data-driven weather models, March 2024. DOI: 10.48550/arXiv.2403.13458. arXiv: 2403.13458 [physics]. (Visited on 02/06/2025).

[17] Ankur Mahesh, William Collins, Boris Bonev, Noah Brenowitz, Yair Cohen, Joshua Elms, Peter Harrington, Karthik Kashinath, Thorsten Kurth, Joshua North, Travis OBrien, Michael Pritchard, David Pruitt, Mark Risser, Shashank Subramanian, and Jared Willard. Huge Ensembles Part I: Design of Ensemble Weather Forecasts using Spherical Fourier Neural Operators, February 2025. DOI: 10.48550/arXiv.2408.03100. arXiv: 2408.03100 [physics]. (Visited on 02/19/2025).

[18] J. Baño-Medina, A. Sengupta, D. Watson-Parris, W. Hu, and L. Delle Monache. Toward Calibrated Ensembles of Neural Weather Model Forecasts. *Journal of Advances in Modeling Earth Systems*, 17(4):e2024MS004734, 2025. ISSN: 1942-2466. DOI: 10.1029/2024MS004734. (Visited on 05/14/2025).

[19] M Leutbecher and T N Palmer. Ensemble forecasting. *J. Comput. Phys.*, 227(7):3515–3539, March 2008. DOI: https://doi.org/10.1016/j.jcp.2007.02.014.

[20] Tongtiegang Zhao, Qiang Li, Tongbi Tu, and Xiaohong Chen. An extension of the WeatherBench 2 to binary hydroclimatic forecasts, February 2025. DOI: 10.5194/egusphere-2025-3. (Visited on 08/06/2025).

[21] J S Nanditha, Anuj Prakash Kushwaha, Rajesh Singh, Iqura Malik, Hiren Solanki, Dipesh Singh Chuphal, Swarup Dangar, Shanti Shwarup Mahto, Urmin Vegad, and Vimal Mishra. The pakistan flood of august 2022: causes and implications. *Earths Future*, 11(3), March 2023. DOI: https://doi.org/10.1029/2022EF003230.

[22] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

[23] Franco Molteni, Roberto Buizza, Tim N Palmer, and Thomas Petroliagis. The ecmwf ensemble prediction system: methodology and validation. *Quarterly journal of the royal meteorological society*, 122(529):73–119, 1996.

[24] Zoltan Toth and Eugenia Kalnay. Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review*, 125(12):3297–3319, December 1997. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2. (Visited on 08/12/2025).

[25] Antje Weisheimer, Tim N Palmer, Nicholas J Leach, Myles R Allen, Christopher D Roberts, and Muhammad Adnan Abid. CO2-induced climate change assessment for the extreme 2022 pakistan rainfall using seasonal forecasts. *Npj Clim. Atmos. Sci.*, 8(1):262, July 2025. DOI: https://doi.org/10.1038/s41612-025-01136-3.

[26] Bohar Singh, Muhammad Azhar Ehsan, and Andrew W Robertson. Calibrated probabilistic sub-seasonal forecasting for pakistan's monsoon rainfall in 2022. en. *Clim. Dyn.*, 62(5):3375–3393, May 2024.

[27] Jorge Baño-Medina, Agniv Sengupta, James D. Doyle, Carolyn A. Reynolds, Duncan Watson-Parris, and Luca Delle Monache. Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia. *npj Climate and Atmospheric Science*, 8(1):1–9, March 2025. ISSN: 2397-3722. DOI: 10.1038/s41612-025-00949-6. (Visited on 03/10/2025).

[28] Karan Purohit, Mitali Sinha, and Ravi S Nanjundiah. AI-based Multimodel Superensemble for Improved Weather Prediction. In *2024 IEEE 31st International Conference on High Performance Computing, Data and Analytics Workshop (HiPCW)*, pages 47–51, December 2024. DOI: 10.1109/HiPCW63042.2024.00017. (Visited on 08/13/2025).

[29] William S. Lamberson, Michael J. Bodner, James A. Nelson, and Sara A. Sienkiewicz. The Use of Ensemble Clustering on a Multimodel Ensemble for Medium-Range Forecasting at the Weather Prediction Center. *Weather and Forecasting*, 38(4):539–554, March 2023. ISSN: 1520-0434, 0882-8156. DOI: 10.1175/WAF-D-22-0154.1. (Visited on 08/13/2025).

[30] 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, July 2020. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.3803. (Visited on 07/03/2025).

[31] Robert W. Carver and Alex Merose. ARCO-ERA5: An Analysis-Ready Cloud-Optimized Reanalysis Dataset. In *103rd AMS Annual Meeting*. AMS, January 2023. (Visited on 06/11/2025).

[32] European Centre For Medium-Range Weather Forecasts. ECMWF IFS High-Resolution Operational Forecasts, 2016. DOI: 10 . 5065 / D68050ZV. (Visited on 06/11/2025).

[33] Nicholas Geneva and Dallas Foster. NVIDIA Earth2Studio, April 2024. (Visited on 06/11/2025).

[34] Zoltan Toth and Eugenia Kalnay. Ensemble forecasting at NMC: the generation of perturbations. *Bull. Am. Meteorol. Soc.*, 74(12):2317–2330, December 1993. DOI: https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2..

[35] Ankur Mahesh, William Collins, Boris Bonev, Noah Brenowitz, Yair Cohen, Peter Harrington, Karthik Kashinath, Thorsten Kurth, Joshua North, Travis OBrien, Michael Pritchard, David Pruitt, Mark Risser, Shashank Subramanian, and Jared Willard. Huge Ensembles Part II: Properties of a Huge Ensemble of Hindcasts Generated with Spherical Fourier Neural Operators, February 2025. DOI: 10 . 48550 / arXiv . 2408 . 01581. arXiv: 2408.01581 [cs]. (Visited on 02/19/2025).

[36] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models, January 2024. DOI: 10.48550/arXiv.2308.15560. arXiv: 2308.15560 [physics]. (Visited on 06/11/2025).

[37] W Weibull. Astatistical theory ofthe strength of materials. *Proc. Royal 4cademy Engrg Science*, 15(1):1, 1939.

A Data Description

Two data sources are considered in this study: 1) ERA5 reanalysis data [30, 31], which serves both as initial conditions for our forecasting models and as ground truth to evaluate the forecast performance; and 2) IFS [32], which is taken as a benchmark dataset to compare the forecast skill of the AIWP architectures with respect to the NWP models, represented by the ECMWF Integrated Forecasting System Ensemble (ENS) [23]. First, we use a subset of the ERA5 ARCO [31] variables required for prediction and evaluation, following [16], with 6-hourly temporal resolution (see Table 1 for the complete list). Additionally, we incorporate the IFS, ECMWF's leading NWP probabilistic system, i.e., the ENS [32]. To enable a fair comparison between the AIWP model and IFS outputs, all data used in this study are downscaled to a common 1ř Œ 1ř spatial resolution, using bilinear interpolation [33]. The evaluation is conducted based on the variables described in Table 2.

Table 1: ERA5 variables used for forecasting with AIWP models.

Variable	**Description**	**Unit**	**Pressure levels** (hPa)
z[a]	Geopotential	$m^2\,m^{-2}$	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50 – 13 levels
q	Specific humidity	kg kg	
r[b]	Relative humidity	%	
t[a]	Temperature	K	
u[a]	U component of wind	$m\,s^{-1}$	
v[a]	V component of wind	$m\,s^{-1}$	
w	Vertical velocity	Pa s-1	
msl[a]	Mean sea level pressure	Pa	–
u10m[a]	10m U component of wind	$m\,s^{-1}$	–
v10m[a]	10m V component of wind	$m\,s^{-1}$	–
u100m	100m U component of wind	$m\,s^{-1}$	–
v100m	100m V component of wind	$m\,s^{-1}$	–
t2m[a]	2m Temperature	K	–
sp	Surface pressure	Pa	–
tcwv	Total column water vapour	$kg\,m^{-2}$	–
tp06[c]	Total precipitation 6 hourly accumulation	m	–
z	Geopotencial at surface	$m^2\,s^{-2}$	–
lsm	Land sea mask	(0 - 1)	–

[a]This variable is perturbed when using perturbation methods.
[b]This variable is derived from pressure, **q** and **t**.
[c]This variable is accumulated in 6-hourly intervals from the original 1-hourly data.

Table 2: ERA5 variables used for evaluating and benchmarking the AIWP models.

Variable	**Description**	**Unit**
z500	500 hPa Geopotential	$m^2\,s^{-2}$
t850	850 hPa Temperature	K
t2m	2 m Temperature	K
tp06	Total precipitation 6 hourly accumulation	m

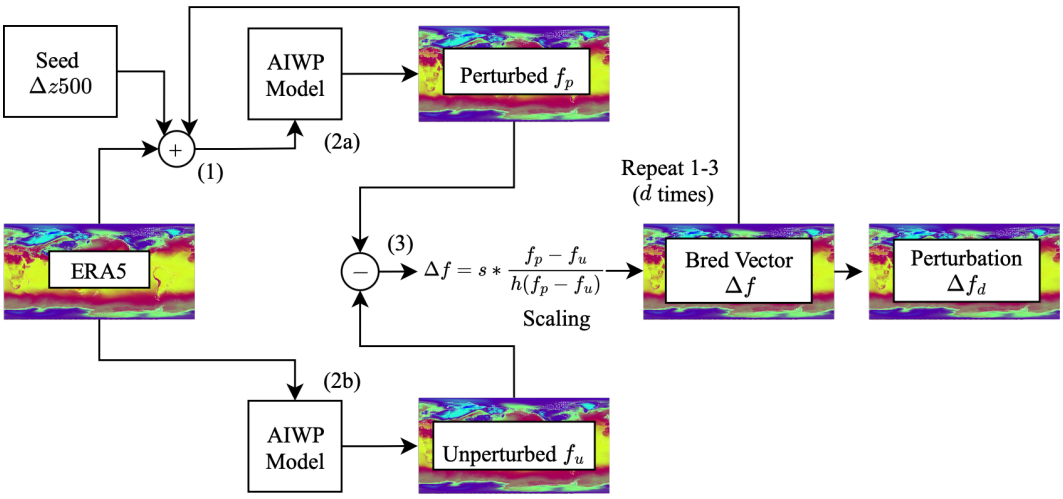

Figure 4: Hemispheric Centered Bred Vector (HCBV) perturbation method. $\Delta z500$ represents a correlated spherical Gaussian noise added to the 500 hPa geopotential variable. h represents the norm or size computed separately for the north and south hemispheres, and interpolated in the tropics. d represents the integration depth, that is the number of recursive cycles which the final perturbation is computed from. This perturbation is additionally centered, so the perturbation vector is alternatively added and subtracted from the initial conditions. Diagram adapted from [18].

B Perturbation Methods

Gaussian: The Gaussian perturbation method applies a random spherical Gaussian field, scaled with a fixed factor, to the input y. We used the Spherical Gaussian implementation in [33]. Formally:

$$y_{\text{pert}}[n] \;=\; y[n] \;+\; s\,\xi[n],$$

where n represents each input spatial location, s is the noise amplitude (scale factor), and $\xi[n]$ is the mean-zero spherical Gaussian random field value sampled for each spatial location n.

Hemispheric Centered Bred Vector (HCBV): The bred vector algorithm, firstly introduced in [34], is a perturbation method that is based on the fact that initial conditions generated by data assimilation/reanalysis processes accumulate growing errors. To estimate this growing error direction and magnitude, i.e., the bred vector Δf, the weather prediction model is used to create a perturbed forecast f_p based on a small change in the initial condition (seed). We use a hemispherical-centered approach [17] where we do not only scale the amplitude of Δf by a factor s but also normalize the perturbation field in extra-tropical regions ($|latitude| > 70°$, $h(\Delta f)$) separately from the north, tropics, and south regions, i.e., $\Delta f = s * \frac{f_p - f_u}{h(f_p - f_u)}$ (f_u is the unperturbed forecast). In this study, we use an integration factor of $d = 3$, following [27], recursively computing Δf three times to better sample the growing errors, which results in Δf_d. This is intended to more accurately reflect the uncertainties in the analysis data and data assimilation process. Moreover, we use a correlated spherical Gaussian noise for the 500 hPa geopotential variable as a seeding perturbation method ($\Delta z500$). The method is illustrated in Figure 4.

Huge Ensembles (HENS): Huge ensembles is a slightly adapted HCBV perturbation method [35]. The method makes use of the average RMSE at 48h for each of the perturbed

variables and AIWP models to further tune the amplitude of the perturbation ($s_{HENS} = \text{RMSE} * s$). Similarly to HCBV, we use $d = 3$ as a parameter for the integration depth [18].

It should be noted that only the common input variables between all models were perturbed (see Table 1). Moreover, we used a scaling factor $s = 0.35$ for all perturbation methods, following [17].

C Evaluation Metrics

For the following metric definitions, y_{nt} is the ERA5 value for each spatial location n (N spatial locations in a study region) at time $t \in \{1, \ldots, T\}$, being T the number of forecast timesteps, and $\hat{y}_{nt}^{(m)}$ is the forecast provided by ensemble member $m \in \{1, \ldots, M\}$, where $M = 50$.

C.1 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a metric that measures the average magnitude of errors between predicted and observed values [36]. It penalizes larger errors more heavily, making it sensitive to outliers. Lower RMSE represents a better forecast. RMSE is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \left(y_{nt} - \frac{1}{M} \sum_{m=1}^{M} \hat{y}_{nt}^{(m)} \right)^2}$$

C.2 Continuous Ranked Probability Score (CRPS)

The Continuous Ranked Probability Score (CRPS) is a scoring rule that measures the accuracy of probabilistic forecasts by quantifying the difference between the predicted cumulative distribution function (CDF) and the observed outcome [36]. It generalizes the mean absolute error to probabilistic forecasts. Lower CRPS points at better forecast skill. CRPS is defined as follows:

$$\text{CRPS} = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \left[\frac{1}{M} \sum_{m=1}^{M} \left| \hat{y}_{nt}^{(m)} - y_{nt} \right| - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{k=1}^{M} \left| \hat{y}_{nt}^{(m)} - \hat{y}_{nt}^{(k)} \right| \right]$$

C.3 Receiver Operating Characteristic Skill Score (ROCSS)

The Receiver Operating Characteristic Skill Score (ROCSS) is a metric derived from the Area Under the ROC Curve (AUC) that quantifies the ability of a forecasting system to discriminate between the occurrence and non-occurrence of an event, relative to random chance [20]. In this study, the event of interest, i.e., extreme precipitation, is defined by a threshold τ, which is set to the 99th climatological percentile from ERA5 data over 1990-2020. We selected this percentile due to the record breaking nature of this event, according to [26]. The computation of the ROCSS proceeds as follows:

1. **Thresholding observations and ensemble forecasts**. Given the threshold τ, observations and ensemble forecasts are converted into binary outcomes. Then, ensemble binary forecasts are converted into forecast probabilities:

$$o_{nt} = \mathbf{1}\{y_{nt} > \tau\}, \quad f_{nt}^{(m)} = \mathbf{1}\{\hat{y}_{nt}^{(m)} > \tau\}, \quad r_{nt} = \sum_{m=1}^{M} f_{nt}^{(m)}, \quad p_{nt} = \frac{r_{nt}}{M+1}.$$

Here, o_{nt} is the binary observation, $f_{nt}^{(m)}$ is the binary forecast of ensemble member m, r_{nt} is the ensemble vote count, and p_{nt} is the resulting forecast probability, using Weibulls plotting position [37].

2. **Applying probability thresholds**. To construct the Receiver Operating Characteristic (ROC) curve, the forecast probability is converted back to binary outcomes for each probability threshold $\theta \in [0, 1]$:

$$\hat{o}_{nt}(\theta) = \mathbf{1}\{p_{nt} \geq \theta\}$$

3. **Calculating Hit Rate (HR) and False Alarm Rate (FAR)**:

$$\mathrm{HR}(\theta) = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T} \mathbf{1}\{\hat{o}_{nt}(\theta) = 1 \wedge o_{nt} = 1\}}{\sum_{n=1}^{N} \sum_{t=1}^{T} o_{nt}},$$

$$\mathrm{FAR}(\theta) = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T} \mathbf{1}\{\hat{o}_{nt}(\theta) = 1 \wedge o_{nt} = 0\}}{\sum_{n=1}^{N} \sum_{t=1}^{T} (1 - o_{nt})}$$

4. **Constructing the ROC curve**. The ROC curve plots the trade-off between hits and false alarms across all probability thresholds:

$$\{(\mathrm{FAR}(\theta), \mathrm{HR}(\theta)) : \theta \in [0, 1]\}$$

5. **Computing the AUC**:

$$AUC_{\text{forecast}} = \int_0^1 \mathrm{HR}(\mathrm{FAR}^{-1}(x))\, dx,$$

where $\mathrm{FAR}^{-1}(x)$ denotes the inverse mapping of the false alarm rate along the ROC curve.

6. **Computing the ROCSS**. Finally, the ROCSS compares the forecast AUC to a reference (random) forecast, which has $AUC_{\text{reference}} = 0.5$:

$$\mathrm{ROCSS} = \frac{AUC_{\text{forecast}} - AUC_{\text{reference}}}{1 - AUC_{\text{reference}}}.$$

A higher ROCSS indicates better discriminatory skill, with 1 representing perfect skill and 0 corresponding to no skill.

D Implementation Details

The NVIDIA-developed Earth2Studio framework [33] has been used for data handling, perturbing the input data and running predictions with various global weather forecasting models. This modular and Python-based framework enables running large-scale inference with multiple models and perturbation methods efficiently. To compute all evaluation metrics, we made use of the WeatherBench framework [36], as well as its binary hydroclimatic extension introduced in [20].

E Deterministic AIWP Forecasts for 18th August 2022

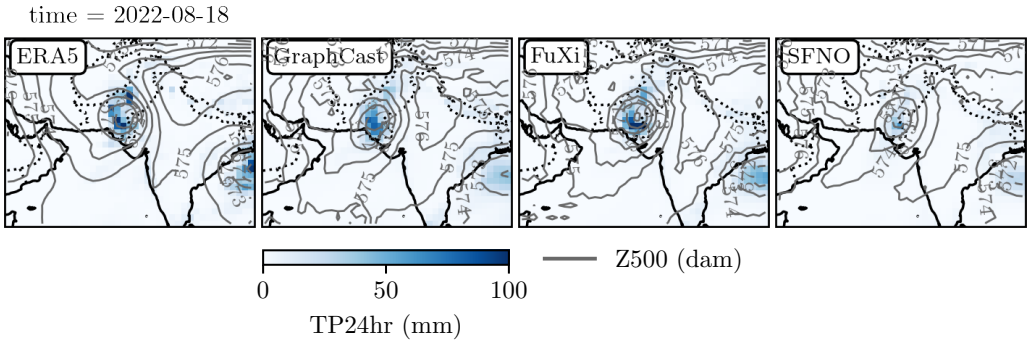

Figure 5: Accumulated daily precipitation on 18th August 2022 in the Pakistan region, both for ERA5 (ground truth, see subfigure on the left side) and the deterministic 3-day lead time forecast provided by GraphCast, FuXi, and SFNO. Daily average geopotential height at 500 hPa is also displayed as contour lines.

F Global RMSE and CRPS for August 2022

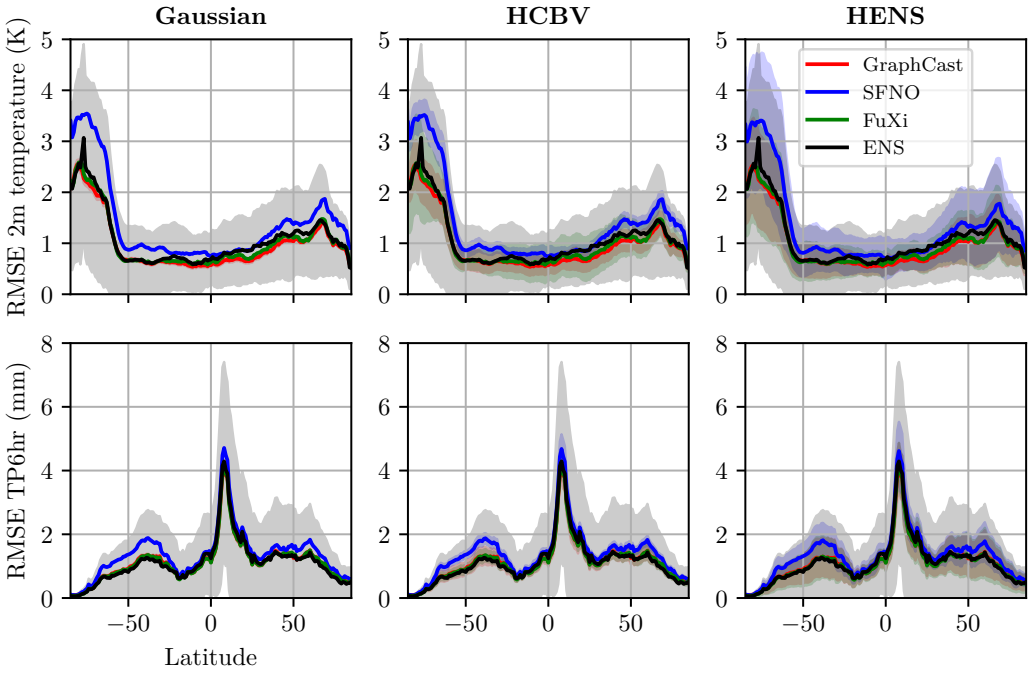

Figure 6: Ensemble mean RMSE per latitude, for a 3-day lead time worldwide forecast in August 2022. The shaded area represents the ensemble mean RMSE variance. The Gaussian perturbation method provides under-dispersive ensembles. HENS method most closely match the ENS spread. SFNO performs worse on all variables and perturbation methods, while GraphCast and FuXi perform at a similar level, but still not reaching the dispersion of ENS.

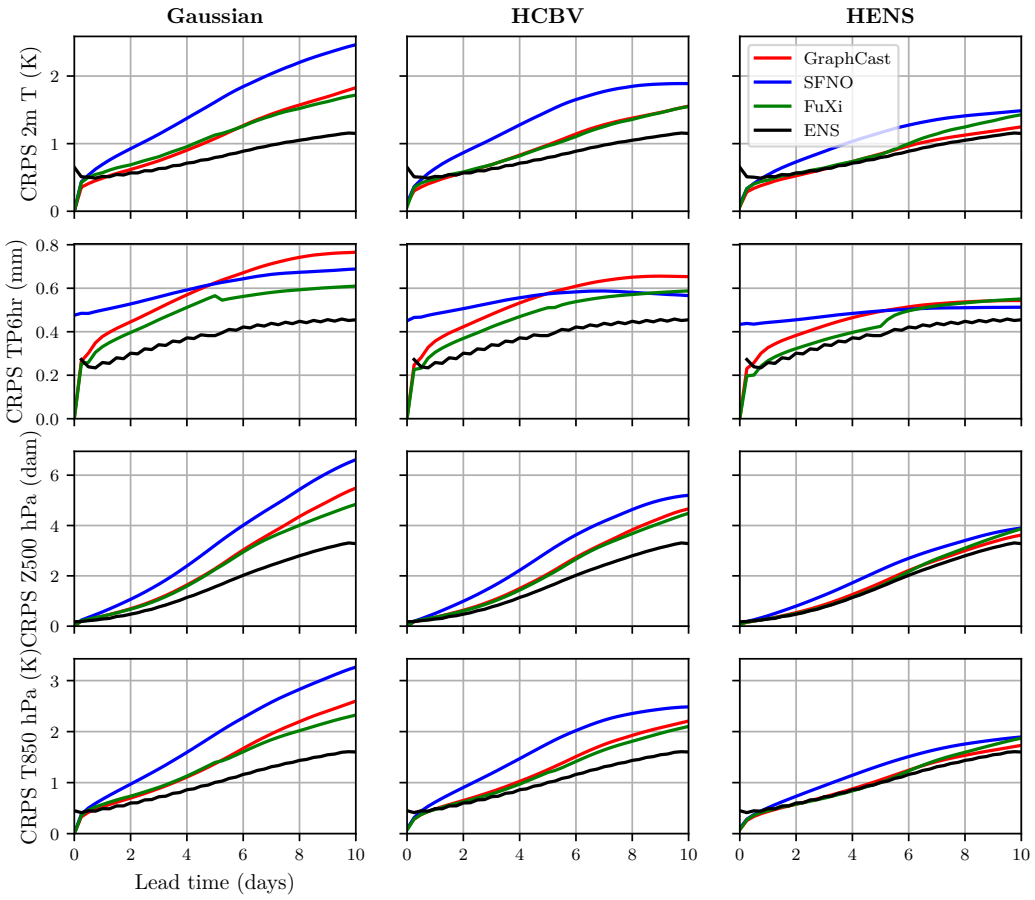

Figure 7: Global average CRPS over 10-day lead time for the different perturbation methods and AIWP models and ENS, in August 2022. ENS showcases the best performance overall for all metrics, except in the very short lead time range, where AIWP models are competitive. HENS perturbation method most closely matches the CRPS of ENS.