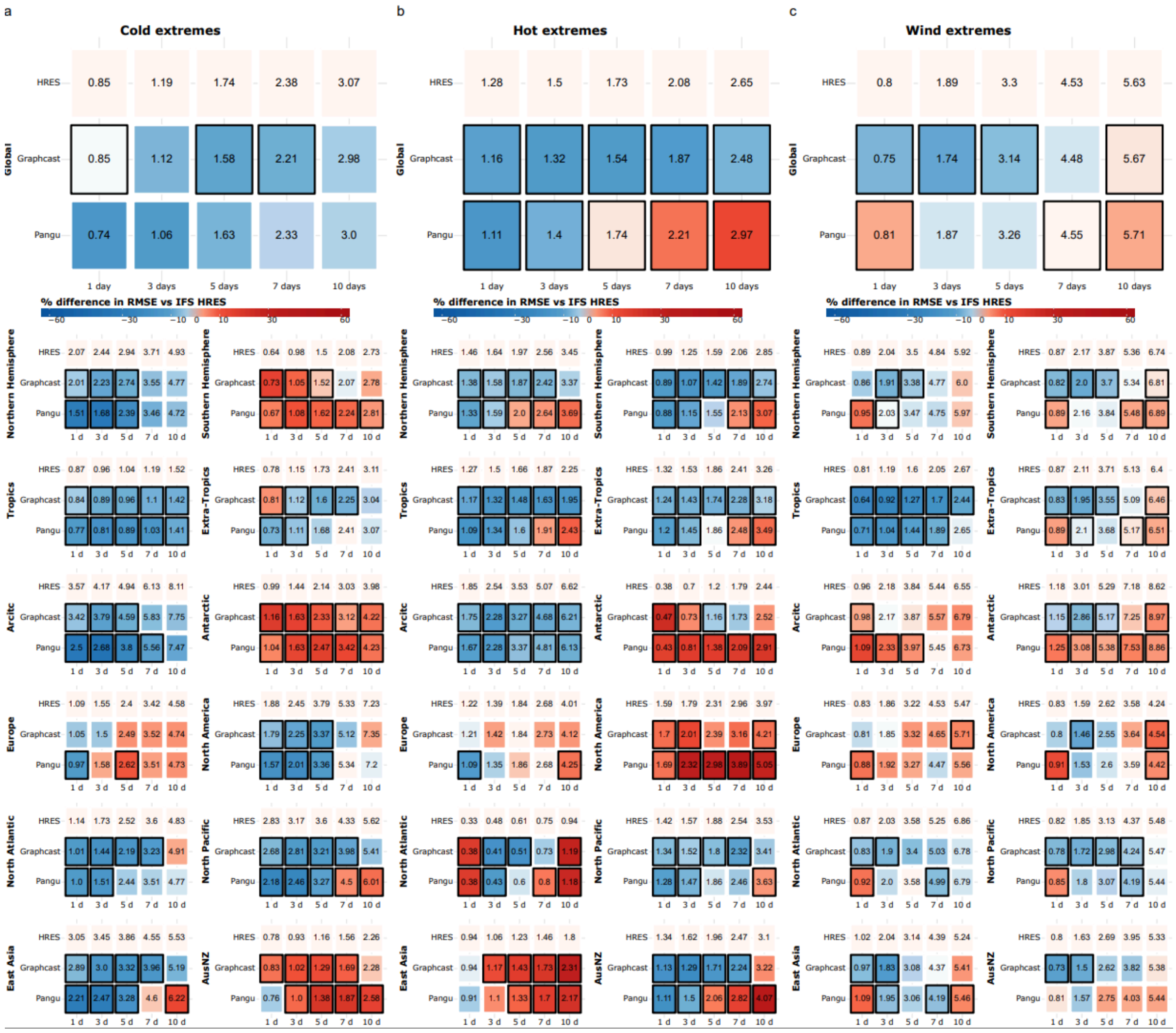
Answer to Review #2

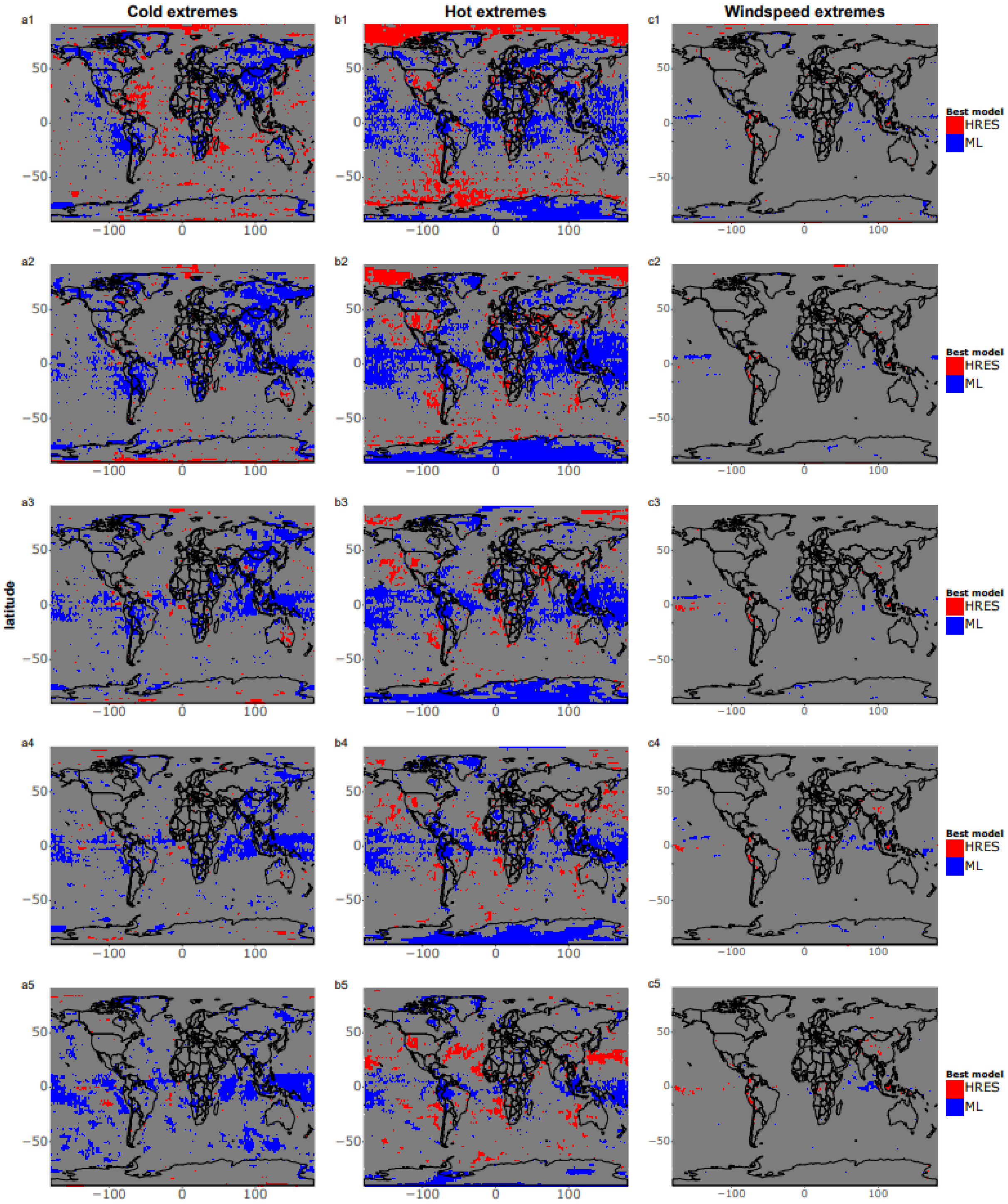
Dear Reviewer,

Thank you for the time you spent reviewing our article and for your constructive feedback. We welcome your suggestion to expand our analysis to explore possible reasons for regional and variable-based differences between models. In the first draft of our manuscript, we limited ourselves to a descriptive analysis of the results due to concerns related to the limited sample size of our test data and the lack of statistical significance of the results. However, following valuable comments from the other reviewer, we have now expanded our analysis to also include robust significant tests of all our metrics, for all regions and individual grid points (Figure 1-6). We believe that these tests may contribute to strengthen the results of our analysis and also help to clarify which of the patterns we identified are supported by robust empirical evidence.

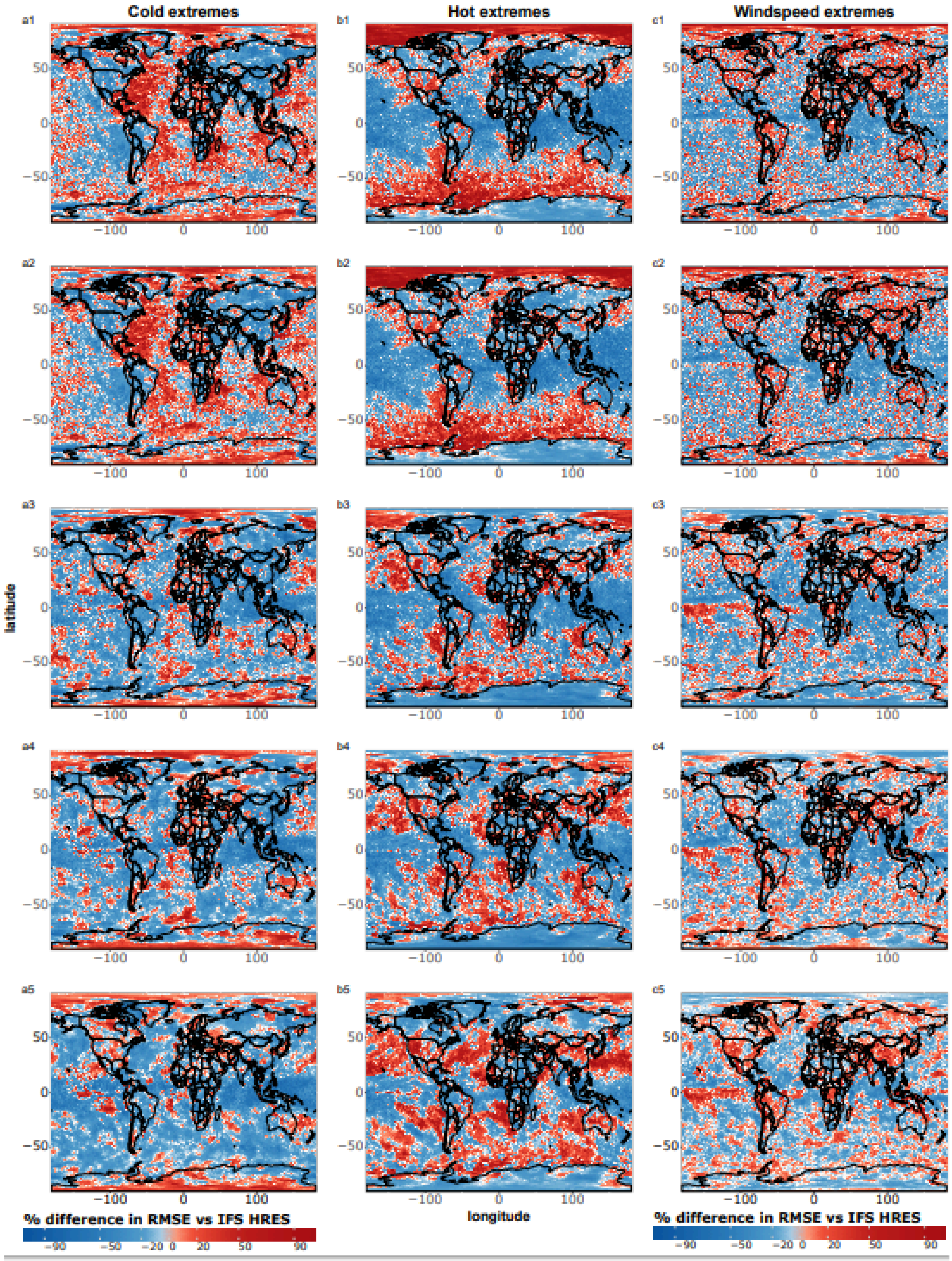
The testing approach we employ is based on clustered standard errors (Liang and Zeger (1986), Arellano (1987), and Cameron and Miller (2015)), a classic econometric approach specifically designed for observations correlated in time and space. The intuition behind the use of clustered standard errors is that since many of our extreme data-points come from adjacent grid points and from events close to each other in time the effective number of degrees of freedom for our tests is much smaller than the total number of available paired forecast differences. Thus, we account for this by inflating the standard errors, by introducing a clustering parameter, which takes into account the clustered nature of our extremes in space and time. Below, we provide an example figure (updated Figure 2) to illustrate our changes. Black borders indicate here that the performance of the model at a given lead time is statistically significantly different from IFS HRES, at the 5% significance level.



Similarly, we performed significance testing with time-clustered standard errors also for grid-pointwise comparisons, exemplified below in a new figure to be added to the manuscript where we evaluate statistically significant differences between the IFS HRES and the best of the machine learning models at each grid point. Here, we also make use of global false discovering rates (Wilks 2016, Benjamini and Hochberg, 1995) to correct for multiple testing and ensure robust statistical inference. As illustrated by Wilks (2016), this approach is also robust for spatially correlated values. As in Figures 1-3, blue shades indicate that the machine learning model is better than IFS HRES, whereas red shades indicate that IFS HRES is better. Gray shades indicate a lack of statistically significant differences.



Additionally, we plan to add corresponding figures showing the magnitude of the differences between the models, see example figure below.


Cold extremes
Hot extremes
Windspeed extremes
a1
b1
c1
a2
b2
c2
a3
b3
c3
a4
b4
c4
a5
b5
c5
50
0
−50
−100
0
100
latitude
longitude
% difference in RMSE vs IFS HRES
−90
−50
−20
0
20
50
90

Besides the above changes, we are planning to specifically address your concerns in a number of additional ways, by thoroughly revising and expanding several sections of our manuscript:

- Figures 1-3 are going to include about a page each of discussion of the results, where we explore possible explanations for regional and lead-time based differences in performance between the models. We identify, in particular, two key drivers for these differences, namely: the 1) the presence of increased blurring for data-driven models in relation to extreme weather forecasts, and 2) a meridional pattern in the quality of data-driven forecasts, with the best performance closest to the Equator, and the worst performance at high latitudes, in most cases. Besides figures 1-6 and the magnitude figure exemplified above, we also find evidence of this behaviour in the additional figure below (Figure R1), where we plot the relative difference in tail RMSE (5% most extreme events) between the ML models and IFS HRES (y-axis) vs latitude (x-axis) for 1-10 days ahead forecasts, as in previous figures. Despite the presence of some noise, we can notice some recurrent convexity in the performance of machine learning models, especially at shorter lead times, with clear spikes of poor performance close to the Poles. We believe that this behaviour may be ascribed to the use of area weighted loss functions, which place greater emphasis on errors closer to the Equator rather than to the Poles in order to maximise the performance in standard area-weighted performance metrics.

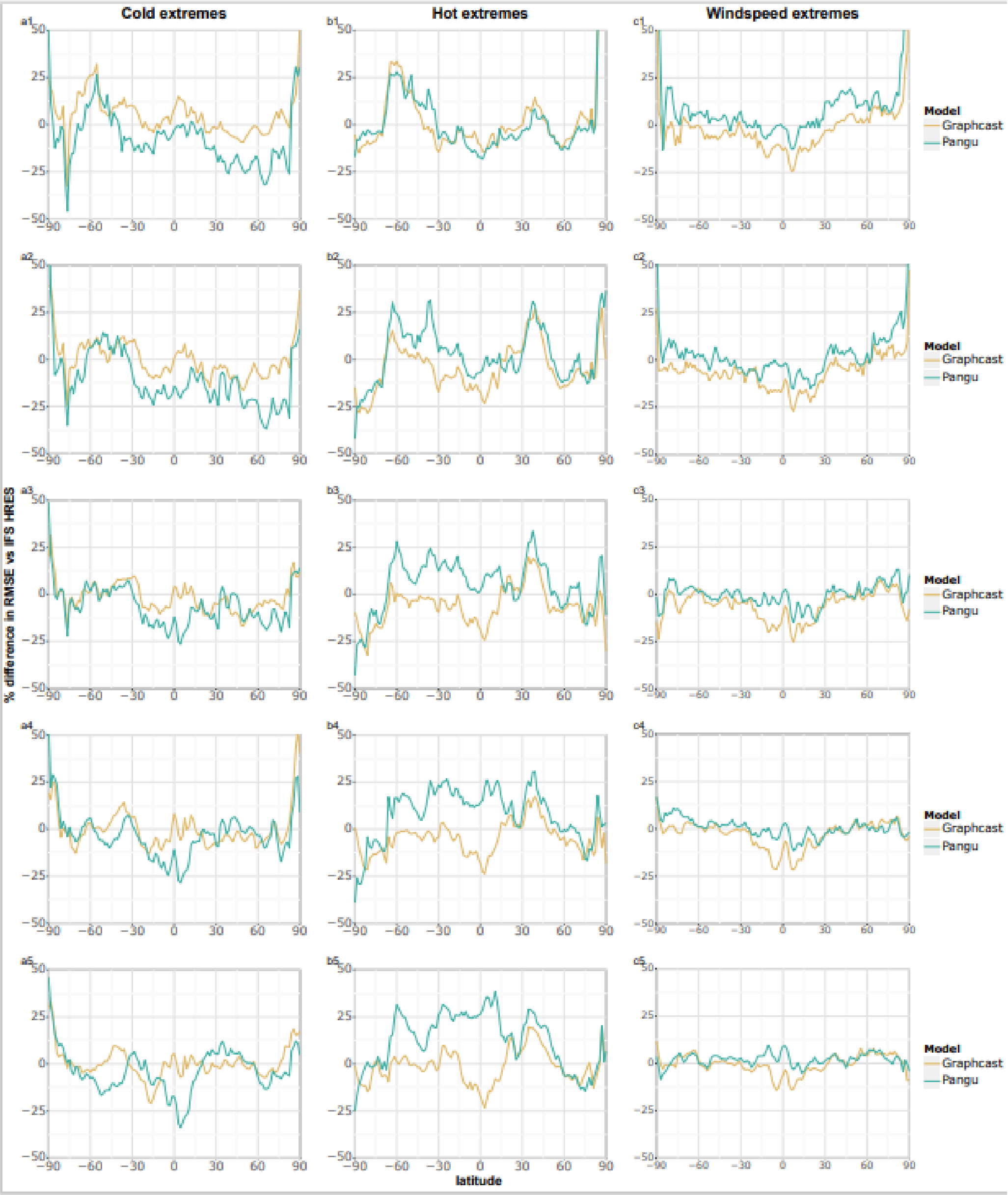


Fig. R1: *Relative difference in tail RMSE (y-axis) vs. latitude (x-axis) for cold (a), hot (b), and windy extremes (c). The data points are computed based on the (a) 5% lowest 2m temperatures, (b) 5% highest 2m temperatures, and (c) 5% highest 10m wind speeds, respectively. Forecasts are shown for X1) 1 day, X2) 3 days, X3) 5 days, X4) 7 days, and X5) 10 days. Negative values of the relative difference indicate better performance than IFS HRES, while positive values indicate worse performance than IFS HRES.*

- We are going to rewrite our description of Figures 5-6 to place greater focus on statistically significant results, and also explore differences in performance between different variables. We link these differences to the previously identified patterns, and also explore alternative explanations for newly identified regional patterns (e.g. lack of key input variables as a driver of subpar performance in continental areas among data-driven models).

- We are going to expand our description of Figures 7-10 to link the calibration results to the rest of the analysis, and provide a more in depth evaluation of the tail reliability of data-driven models in different regions. We now also provide several possible justifications for the results we obtain.
- Lastly, we are going to expand and partially rewrite the discussion and conclusion section to place greater emphasis on statistically significant results. Wherever possible, we also back up possible explanations for these results with findings from previous literature.
- Additionally, we will also expand our discussion to address the limitations of our extreme metrics, emphasising that every metric has weaknesses, and that any attempts to make overarching comparisons between models should account for a range of different metrics simultaneously, as well as look at the performance of the forecasting models for the whole distribution of the variables, and not just at the tails. Specifically, QQ-plots and other reliability checks are key here, since they could easily expose attempts to hedge extreme metrics such as the tail RMSE. We will further add some figures related to this point to a new Appendix B.

In summary, we believe our results may be ascribed to a number of different causes, which we are also going to discuss in our manuscript:

1. The overall worsened performance of data-driven models for extremes compared to standard metrics of average performance is likely linked to the choice of loss functions (Xu et al., 2024, Olivetti and Messori, 2024) and globally smoothed multitask approaches which are explicitly designed to optimise those metrics rather than extreme metrics.
2. The relative decline in performance of data-driven models at longer lead times is tied to the phenomenon of blurring (Bonavita, 2024; Price et al., 2024), which appears to be more prominent for extremes than for the overall distribution of the variables.
3. Other possible explanations for the decreased performance at longer lead times include the use of multi-step approaches in training (e.g. Bi et al., 2023), which may lead to compound errors in initial atmospheric states over time (Bonavita 2024), and the use of only the most recent time steps as inputs for the models.
4. The observed regional pattern of better performance in the Tropics and worse performance at higher latitudes is likely connected to the use of latitude-based area weights (e.g. Lam et al., 2023, Chen et al., 2023), which optimise performance closer to the Equator at the expense of performance at higher latitudes.
5. The weaker performance of data-driven models for windy extremes, and for temperature extremes in some specific regions, may be tied to the lack of key input variables such as snow coverage, precipitation and soil moisture.
6. The overall weaker performance of data-driven models for wind extremes compared to temperature extremes may be related to the separate training of 10m u-and v-wind, whose errors may be magnified when looking specifically at 10m windspeed.

References

Arellano, M. ‘PRACTITIONERS’ CORNER: Computing Robust Standard Errors for Within-Groups Estimators’. *Oxford Bulletin of Economics and Statistics* 49, no. 4 (1987): 431–34. https://doi.org/10.1111/j.1468-0084.1987.mp49004006.x.

Benjamini, Yoav, and Yosef Hochberg. ‘Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing’. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, no. 1 (1995): 289–300.

Bi, Kaifeng, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. ‘Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks’. *Nature*, 5 July 2023, 1–6. https://doi.org/10.1038/s41586-023-06185-3.

Bonavita, Massimo. ‘On Some Limitations of Current Machine Learning Weather Prediction Models’. *Geophysical Research Letters* 51, no. 12 (2024): e2023GL107377. https://doi.org/10.1029/2023GL107377.

Bouallègue, Zied Ben, Mariana C. A. Clare, Linus Magnusson, Estibaliz Gascón, Michael Maier-Gerber, Martin Janoušek, Mark Rodwell, et al. ‘The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning-Based Weather Forecasts in an Operational-like Context’. *Bulletin of the American Meteorological Society* 1, no. aop (29 February 2024). https://doi.org/10.1175/BAMS-D-23-0162.1.

Cameron, A. Colin, and Douglas L. Miller. ‘A Practitioner’s Guide to Cluster-Robust Inference’. *Journal of Human Resources* 50, no. 2 (31 March 2015): 317–72. https://doi.org/10.3368/jhr.50.2.317.

Chen, Lei, 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’. *Npj Climate and Atmospheric Science* 6, no. 1 (16 November 2023): 1–11. https://doi.org/10.1038/s41612-023-00512-1.

Kochkov, Dmitrii, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, et al. ‘Neural General Circulation Models for Weather and Climate’. arXiv, 7 March 2024. https://doi.org/10.48550/arXiv.2311.07222.

Lam, Remi, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, et al. ‘Learning Skillful Medium-Range Global Weather Forecasting’. *Science* 382, no. 6677 (22 December 2023): 1416–21. https://doi.org/10.1126/science.adi2336.

Olivetti, Leonardo, and Gabriele Messori. ‘Advances and Prospects of Deep Learning for Medium-Range Extreme Weather Forecasting’. *Geoscientific Model Development* 17, no. 6 (21 March 2024): 2347–58. https://doi.org/10.5194/gmd-17-2347-2024.

Price, Ilan, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, et al. ‘GenCast: Diffusion-Based Ensemble Forecasting for Medium-Range Weather’. arXiv, 1 May 2024. https://doi.org/10.48550/arXiv.2312.15796.

Rasp, Stephan, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, et al. ‘WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models’. *Journal of Advances in Modeling Earth Systems* 16, no. 6 (2024): e2023MS004019. https://doi.org/10.1029/2023MS004019.

Wilks, D. S. ‘“The Stippling Shows Statistically Significant Grid Points”: How Research Results Are Routinely Overstated and Overinterpreted, and What to Do about It’, 1 December 2016. https://doi.org/10.1175/BAMS-D-15-00267.1.

Xu, Wanghan, Kang Chen, Tao Han, Hao Chen, Wanli Ouyang, and Lei Bai. ‘ExtremeCast: Boosting Extreme Value Prediction for Global Weather Forecast’. arXiv, 2 February 2024. https://doi.org/10.48550/arXiv.2402.01295.