In their paper “Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather and GraphCast”, the authors provide an evaluation of the mentioned models plus FuXi for the whole distribution, and for the tails of the distribution (at 5% and 1% levels). It’s really good to see this kind of analysis, as the evaluation of the new generation of data-driven models beyond simple things like global mean RMSE or individual case studies is very limited at present. I really enjoyed reading this paper, and am grateful to the authors for adding this valuable analysis to the body of knowledge on data-driven weather and climate models.

The authors do a good job of describing the data sets used, and their methodology. However, I feel like some additional figures will help to better contextualize their results – specifically where they present maps showing which model performs best at each grid point and lead time. Without corresponding maps showing the magnitude of the difference in scores between each model, it is hard to know how significant the patterns of which model is best are.

My main concern with this paper is that there is no discussion of the statistical significance of their results, which I think is essential given that they are looking at the tails of the distribution, and in some instances some quite small samples. I suggest the authors quote statistical significance thresholds of all their results and add stippling to their figures to indicate statistically significant areas.

More specific comments follow.

Abstract

Line 5-6: “in the average prediction of 10m windspeed and 2m temperature” - averaged over what? At what leadtimes and timescales?

Introduction

Line 49: “comparison in terms of a standard metric (RMSE)” - it would be good to note that RMSE is the objective function of the ML models, so evaluating against RMSE is not a fully independent target (I know you touched on this above – but making it clear you’re taking this into consideration when you choose RMSE for your evaluation metric would be good).

Models and Methodology

Line 71-72: “operational Pangu weather (Bi et al., 2023) and operational GraphCast (Lam et al., 2023)” - I think you need to explain the difference between operational GraphCast/Pangu weather and reanalysis versions. I assume it’s that the operational versions have been fine-tuned on the IFS HRES, but you should make that clear, especially since in the model description sections 2.2 and 2.3 you describe Pangu and GraphCast as being trained on ERA5 and predicting on the ERA5 grid, but then also talk about them predicting on the IFS HRES grid without an explanation of how that works. This is also important to be clear about, since (for example) providing IFS HRES ICs to a model trained on ERA5 then fine-tuned on IFS HRES will give a much better-quantified result than providing some other ICs to the model that it

has not been fine-tuned on – in this case some reduction in performance can be expected (based on my own investigations) but the extent of this would be unknown.

Line 76-79: I'm a little disappointed that you didn't also include an SFNO-based model like FourCastNet, since they exhibit quite different spatial variation in their RMSE scores compared to GraphCast, and are quite a different architectural approach to any of the models you've looked at.

Line 109: "As for Pangu-Weather" - suggest changing to "As with Pangu-Weather"

Section 2.4: Given you only look at FuXi in the Appendix, I'd suggest moving this model description there.

Line 122: "and all comparisons are based on a spatial resolution of 1.5 degrees" - I think it would be good to expand on this a bit. I assume you have regridded the data from 0.1 to 1.5 degree resolution - How did you do this? Any special treatment of the poles? Did you do this before or after calculating metrics (presumably before). Basically, some more procedural details would be appreciated.

Line 123: "sand" -> "and"

Line 126-139: I suggest putting this information in a table for easier reading.

Line 154 and onwards: I would suggest avoiding use of the term 'observations' since you are evaluating against reanalysis. The term observations runs the risk of causing confusion and giving the impression you are evaluating directly against point obs for example.

Line 169: So in case 2, the quantiles are computed using 702 values because that's what there is for each grid point? Might be worth making this clear, and making it clear that the number of points contributing in case 1 is 702 * num_lat * num_lon

Results

Fig 1: It might make the figure a bit too busy, but have you tried drawing borders around the scores? As it is the HRES scores look a bit like headings because their colour is always white. Borders may not look better though – it's hard to tell without seeing it, so please take this as just a loose suggestion.

Line 193: "extremes observations" -> "extreme observations"

Line 193: How many data points does the 5% most extreme cases leave you with? What is the statistical significance of the scores shown in Figure 2 (and Figure 1 as well for completeness I suppose).

Line 200: Ditto previous comment, but for 1%

Figure 4: I would suggest you try color schemes other than red-white-blue for this figure since there's a value judgement (better/worse compared to baseline) associated with those colours from the previous figures. Since this is a straight model comparison, some totally different color scheme would be good in my opinion.

Figure 5: Same comment as for Fig 4

Figure 5: EDIT – never mind – I see that you have addressed this in the discussion! ~~Seems to me like HRES is somewhat better than the ML models at wind speed over land, and worse over the oceans. Do you have any thoughts on why this might be? I feel like there might be a clue in there as to how to make the ML models better (orography as a forcing variable perhaps?).~~

Figure 5: Similarly to the previous comment, it looks to me like there's some tendency for HRES to be better at 2T on the westward side of the continents, and worse on the eastward side. Do you have any thoughts on what this could be due to? Since these are upwelling regions and the feature grows with lead time my intuition is that it could be related to the lack of an ocean in the ML models?

Figures 5 and 6: Some measure of the magnitude of the differences between the models would be very valuable to contextualize what's shown in this figure (maybe maps of the model differences with stippling for statistical significance added to another appendix?) – some of these pixels where one model is shown as best may have a very marginal difference between the models and it would be good to know where this is the case. It would also be good to include an indication of what is statistically significant with this, especially for Fig 6 (where I think your sample is 702 * 0.05 = 35?). This could for example be stippling on Figs 5 and 6 as well as any maps of the magnitudes of the differences.

Figure 7: I find these plots pretty hard to read without leaning right in – perhaps you could increase the marker size, or subtract the y=x line from each set or points and display them ad deviations from perfect calibration to increase the visual distance between the markers?

Figure 7: Some indication of statistical significance or confidence on these plots would be good

Figure 8: Ditto both comments from Fig 7

Line 261-277: While this is interesting, it feels a little out of place since you haven't discussed FuXi anywhere else, and in the next section you once again stop referring to FuXi. I feel like you should either move these paras to the appendix so that the FuXi and other reanalysis initialized models are self-contained in the appendix, or you should add some acknowledgement of their analysis to the opening sentences of the conclusion to make the transition from these paragraphs to the conclusion less jarring.

Discussion and Conclusions

Line 295: It also looks to me like Pangu does better at cold extremes, and for hot extremes GraphCast is better over the oceans while Pangu is better over land (based on Fig 6)

Line 298-299: "the choice of best model depends strongly on region, lead time, type of extreme and in some cases even level of extremeness". I feel like this might not be quite the right way to put it – to me this implies that there is strong variability in the magnitude of the differences between the model's scores with region, lead, extreme type etc., but the magnitude of the differences between the models is not clear form most of your figures. All we know is that there is a lot of variability in which is best with region, lead time etc., but not by how much it is best. I suspect that you meant that with your wording, but I'm a bit worried it could be misinterpreted and suggest you revise this statement.

Line 301-303: ". Ideally, we envisage a hybrid use of physics- and data-driven models to forecast extremes, with physics-based models being supplemented by data-driven models for those areas where

data-driven models have been shown to be superior in terms of tail performance." - are the sizes of the differences in performance enough to justify this approach? I'm taking an operational forecast perspective here, and it feels like the potential gains would have to be more than marginal to justify this.

Appendix

The same comments apply to the appendix figures as to the figures in the main text:

- Some measure of statistical significance would be useful
- For figures A5 and A6, some accompanying figures showing the magnitude of the differences between the models would help give perspective on the significance of the spatial patterns in the figure.
- For figures A4, A5 and A6, a different colour scheme would work better I think – one where each of the models is given a different colour (not just shades of the same colour), and a scheme which is not the same as the one used for showing the magnitude of the differences in scores (since this has a value judgement attached to it)..