Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather and GraphCast

Leonardo Olivetti[1,2] and Gabriele Messori[1,2,3]

[1]Department of Earth Sciences, Uppsala University, 75236 Uppsala, Sweden
[2]Swedish Centre for Impacts of Climate Extremes (climes), Uppsala University, 75236 Uppsala, Sweden
[3]Department of Meteorology and Bolin Centre for Climate Research, Stockholm University, 10691 Stockholm, Sweden

Correspondence: Leonardo Olivetti (leonardo.olivetti@geo.uu.se)

Abstract. The last few years have witnessed the emergence of data-driven weather forecast models able to compete and in some respects outperform physics-based numerical models. However, recent studies question the capability of data-driven models to provide reliable forecasts of extreme events. Here, we aim to evaluate this claim by comparing the performance of leading data-driven models in a semi-operational setting, focusing on the prediction of near-surface temperature and windspeed extremes globally. We find that data-driven models outperform ECMWF's physics-based deterministic model in the average prediction of 10m windspeed and 2m temperature, and can also compete with the physics-based model in terms of extremes in most regions. However, the choice of best model depends strongly on region, type of extreme and sometimes even lead time. Thus, we conclude that data-driven models may already now be a useful complement to physics-based forecasts in those regions where they display superior tail performance, but that some challenges still need to be overcome before widespread operational implementation can take place.

1 Introduction

The first deep learning models for weather applications date back to the 1990s (Schizas et al., 1991; Hall et al., 1999), but it is only in recent years that deep learning models have become competitive as self-standing medium-range forecasting tools. Since 2022, at least seven different research groups (Pathak et al., 2022; Bi et al., 2023; Keisler, 2022; Lam et al., 2023; Chen et al., 2023a; Nguyen et al., 2023; Chen et al., 2023b) claim to have developed deep learning models able to produce more accurate deterministic forecasts than state-of-the-art physics-based models from the European Centre for Medium-Range Weather Forecasts (ECMWF) in a range of atmospheric variables at multiple lead times. Recent independent studies (Rasp et al., 2024; Ben-Bouallegue et al., 2023) support these claims, showing how the data-driven models can outperform physics-based models in a wide range of parameters and metrics. In particular, the WeatherBench 2 (Rasp et al., 2024) provides comprehensive global and regional scorecards for comparing forecast models in terms of test-sample RMSE, while also making all test predictions produced freely available to the public.

However, the studies conducted so far focus on the average skill of the forecasts, without any special treatment of extreme events. Even if some cases studies have been conducted, for instance on cyclone tracking (Bi et al., 2023; Lam et al., 2023; Chen et al., 2023b) and surface temperature extremes (Ben-Bouallegue et al., 2023; Lam et al., 2023), these are too limited to allow for a fair assessment of the capacity of data-driven models to forecast weather extremes globally. Timely and reliable forecasting of weather extremes plays a key role in disaster management and risk mitigation (World Meteorological Organization, 2022; Merz et al., 2020) and in crucial socio-economic functions, such as the energy and insurance sectors (e.g. Kron et al., 2019). We thus argue that greater emphasis should be placed on understanding whether data-driven models can provide reliable forecasts of weather extremes before such models may be implemented operationally (Watson, 2022).

In addition, recent studies (Watson, 2022; Olivetti and Messori, 2024; de Burgh-Day and Leeuwenburg, 2023) problematise the assumption that a strong performance in standard metrics of average skill should translate by default into an equally strong performance in the tails of the distribution. Indeed, there may be several reasons for an asymmetry between average skill and skill for extremes, including the intrinsic sparsity of extreme events in training datasets (Watson, 2022), the use of symmetric loss functions that are inadequate for extremes (Xu et al., 2024; Olivetti and Messori, 2024), and the multitask and multi-step optimisation approaches used in leading deep learning architectures (e.g. Bi et al., 2023; Lam et al., 2023). These issues are further exacerbated by the fact that the current generation of data-driven models provides deterministic predictions, even though a number of promising approaches to provide uncertainty estimates for the predictions are currently being explored (e.g. Hu et al., 2023; Bi et al., 2023; Zhang et al., 2023; Cisneros et al., 2023; Guastavino et al., 2022; Clare et al., 2021; Kashinath et al., 2021).

This article aims to evaluate whether deep learning models can provide skillful forecasts of extreme weather, by providing a pragmatic comparison between physics-based and data-driven models in a semi-operational setting. Specifically, it compares the performance of ECMWF's IFS HRES and leading global deep learning models in the task of forecasting near-surface temperature and windspeed extremes 1-10 days ahead, when provided with the same set of inputs, namely the output of IFS HRES at time 0. To do so, it makes use of the freely available forecast data provided by ECMWF and the WeatherBench 2 dataset (Rasp et al., 2024). The methods for the comparisons between models are largely based on the guidelines for evaluation of tail performance provided by Watson (2022), namely: i) comparison in terms of a standard metric (RMSE) computed on data beyond extreme quantiles only; ii) visual assessment of performance on extremes for specific regions/grid points; and iii) quantile-quantile plots of extreme quantiles to identify possible inconsistencies in tail estimation. All comparisons are performed at multiple time-scales (1-10 days) and for the whole globe, with separate metrics for each region following the ECMWF operational scorecards (ECMWF, 2024).

In the next two sections, we provide an introduction to the models included in the evaluation and the methods employed for the comparison. Then, we outline the results of the comparison for all the variables and regions of interest. Lastly, we reflect on

the results of these comparisons, and on how they may affect the operational implementation of data-driven models. Additional results for models using ERA 5 reanalysis data (Hersbach et al., 2020) as input are included in Appendix A.

2 Models and Methodology

The rationale behind the choice of models and the methodology employed is to make the comparison between data-driven models and physics-based models as fair as possible. For this reason, we include in the main text only those data-driven models included in the WeatherBench 2 that are able to take the same set of initial conditions as IFS HRES, ECMWF's high-resolution deterministic forecasting system. All the models in the main take therefore as an input IFS HRES at time 0, and are able to produce 6 hourly forecasts of 2m temperature and 10m wind, the variables on which the models are evaluated. Those outputs are in turn all compared to the same ground truth, ERA 5 (Hersbach et al., 2020), at 1.5 degrees horizontal resolution, as in the WeatherBench 2 (Rasp et al., 2024). Indeed, models taking as input reanalysis data present a conceptual difference to operational models, as they are based on input data that is available with a considerable time delay and thus cannot be used in an operational setting.

Two-data driven models fit the criteria established above: operational Pangu weather (Bi et al., 2023) and operational GraphCast (Lam et al., 2023). We believe these models may represent reasonably well the performance of data-driven models as a whole, since they display similar performance to other data-driven models in a range of atmospheric and surface variables at multiple lead times (Rasp et al., 2024). Furthermore, these models employ the two leading architectures for data-driven weather forecasting, namely vision transformers (Dosovitskiy et al., 2020) and graph neural networks (Scarselli et al., 2009), respectively. Yet, recognising that some subtle differences may be lost by not including a more diverse range of data-driven models in our comparison, we present in Appendix A a comparison between IFS HRES and the leading reanalysis-based deep learning models, namely reanalysis-based Pangu-Weather, reanalysis-based GraphCast, and FuXi(Chen et al., 2023b), currently regarded as the best data-driven models in terms of RMSE for medium to long range forecasting (Rasp et al., 2024).

In this section, we first provide a brief description of each of the models included in the comparison, and then outline the criteria on which the comparison is based. For a complete description of the models including a full list of inputs and outputs, we refer the reader to Rasp et al. (2024) and Olivetti and Messori (2024), as well as to the original papers introducing the models described in Subsections 2.1-2.4.

2.1 IFS HRES

IFS HRES is ECMWF's flagship deterministic high-resolution model, widely regarded as one of the best physics-based numerical weather forecast models in the world (Rasp et al., 2020, 2024). All the parameters included in the model as well as its regular updates and improvements are thoroughly documented on ECMWF's website (Blanchonnet, 2022). Currently, HRES

takes a much larger set of inputs than any of the data-driven models, and also produces hourly forecasts for a very large set of outputs, at a 0.1° horizontal resolution on 137 pressure levels. The set of inputs forming HRES's initial conditions (HRES at time 0) are a mix of in-situ observations for the three hours surrounding the forecast and model outputs from the previous HRES run. HRES is included here as baseline to which to compare the performance of data-driven models. All IFS HRES forecasts have been generated with the operational version of the model used at the time of the forecast (Rasp et al., 2024), namely model configuration Cy46r1 for forecasts initiated before 2020-06-30, and Cy47r1 for forecasts initiated after that date.

2.2 Pangu-Weather

Pangu-Weather (Bi et al., 2023) is a data-driven, deep learning model using a vision transformer architecture (Dosovitskiy et al., 2020). First developed in 2022 (Bi et al., 2022)) and published in 2023 (Bi et al., 2023), it is the "oldest" data-driven model among those included in the comparison. It is trained on ERA5 reanalaysis data for 1979 to 2017 and uses 2018-2019 as validation. It takes as input five upper-air variables on thirteen atmospheric levels and four surface variables, and it produces forecasts of those same variables for the next atmospheric state 6-hours ahead, in a sequential manner. The output of the model can then be fed again as input, to obtain forecasts at longer lead times. In this way, it is possible to obtain forecasts up to 10 days ahead, at 0.25° resolution. In its operational version, analysed in the main text here, Pangu-Weather takes as input HRES at time 0, while in the version included in Appendix A it takes ERA 5 as initial state.

2.3 GraphCast

GraphCast (Lam et al., 2023) is a deep learning model using a graph-based architecture (Scarselli et al., 2009). First developed in late 2022 (Lam et al., 2022) and published in Lam et al. (2023), it builds on earlier work by Keisler (2022). It is trained on ERA5 reanalysis data for 1979 to 2019. It takes as input six atmospheric variables at 37 atmospheric levels, and numerous surface variables and masks. GraphCast aims to forecast the next state of the atmosphere as a function of its two previous states, in a sequential manner. As for Pangu-Weather, it produces 6-hourly forecasts up to 10 days ahead, at 0.25° resolution. In its operational version, included in the main text, it takes as input HRES at time 0, while in the version included in Appendix A it takes ERA 5 as initial state.

2.4 FuXi

FuXi (Chen et al., 2023b) is the most recent of the data-driven models included here. It builds on the work of Bi et al. (2023) and uses a vision transformer architecture (Dosovitskiy et al., 2020). It is trained on ERA5 reanalysis data for 1979 to 2017. Its main innovation compared to previous models is its cascading optimisation approach, through which different sub-models are developed for different forecasting ranges, with the purpose of improving medium-to-long range forecasts. As of now, data provided by FuXi to the WeatherBench 2 (Rasp et al., 2024) are only for a reanalysis-based version of the model, taking ERA 5 as input. For this reason, FuXi is only included in the comparison shown in Appendix A.

2.5 Criteria for model comparison

The comparison between models is based on their performance in forecasting 2m temperature cold and hot extremes and 10m windspeed extremes globally. Following the WeatherBench 2 (Rasp et al., 2024), the models are tasked with forecasts with a timestep of 6h or less, and all comparisons are based on a spatial resolution of 1.5 degrees. Forecasts are initiated every 12 hours (00:00 sand 12:00) for the period 01-01-2020 to 16-12-2020, thus providing 702 comparable forecasts for each lead time and grid point. Comparisons are performed globally, and for regions included in the ECMWF operational scorecards (ECMWF, 2024), defined as follows:

Northern hemisphere (Extra-tropics): lat $\geq 20°$
Southern hemisphere (Extra-tropics): lat $\leq$ -20°
Tropics: -20° $\leq$ lat $\leq 20°$
Extra-tropics: |lat| $\geq 20°$
Arctic: lat $\geq 60°$
Antarctic: lat $\leq$ -60°
Europe: 35° $\leq$ lat $\leq$ 75°, -12.5° $\leq$ lon $\leq$ 42.5°
North America: 25° $\leq$ lat $\leq$ 60°, -120° $\leq$ lon $\leq$ -75°
North Atlantic: 25° $\leq$ lat $\leq$ 60°, -70° $\leq$ lon $\leq$ -20°
North Pacific: 25° $\leq$ lat $\leq$ 60°, 145° $\leq$ lon $\leq$ -130°
East Asia: 25° $\leq$ lat $\leq$ 60°, 102.5° $\leq$ lon $\leq$ 150°
AusNZ: -45° $\leq$ lat $\leq$ -12.5°, 120° $\leq$ lon $\leq$ 175°

For the sake of conciseness, we focus our comparison here on forecasts for 1, 3, 5, 7 and 10 days ahead. We evaluate the performance of the models based on three different criteria, largely based on the recommendations for evaluation of extreme event forecasts provided by Watson (2022). The criteria are as follows:

1. Accuracy in determining the magnitude of the most extreme observations globally or within a given region. To define the extremes, we pool together all grid-point observations for 2020 for the region of choice, and set a threshold based on a quantile of choice out of all the observations. We then consider as extreme all observations above that threshold. Thus, we allow for any number of global and regional extremes to come from a specific grid point or time. Accuracy is measured in terms of RMSE (lower values are better), as defined below:

- For hot and windspeed extremes:

$$RMSE_t = \sqrt{\frac{1}{TIJ}\sum_{t}^{T}\sum_{i}^{I}\sum_{j}^{J} w(i)\mathbb{1}_{o_t > Q(o)}(\hat{y}_{t,i,j} - o_{t,i,j})^2} \quad (1)$$

- For cold extremes:

$$RMSE_t = \sqrt{\frac{1}{TIJ}\sum_{t}^{T}\sum_{i}^{I}\sum_{j}^{J} w(i)\mathbb{1}_{o_t < Q(o)}(\hat{y}_{t,i,j} - o_{t,i,j})^2}, \quad (2)$$

where,

$1, 2, 3, ..., T$ is the available number of time-points at the given forecast lead time. T is 702 in our case;

$1, 2, 3, ..., I$ is the number of points of latitude included in the region of interest,

$1, 2, 3, ..., J$ is the number of points of longitude included in the region of interest,

$\hat{y}$ is the forecasted value of the variable of interest,

o is the observed value of the variable of interest, in our case from ERA5,

$\mathbb{1}_{o_t > Q(o)}$ is an indicator function taking a value of 1 for observations above the chosen quantile of the variable of interest in the given region, and 0 otherwise. For cold extremes, $\mathbb{1}_{o_t < Q(o)}$ so that the indicator function takes a value of 1 for observations below the chosen quantile, and 0 otherwise.

2. Accuracy in determining the magnitude of grid-point extremes. Extremes are defined as in criterion 1, but at a grid point level, by defining a different threshold and set of extremes for each grid point. The RMSE is computed according to Equations 1 and 2, with a redefined indicator function. For hot extremes, the indicator function is given by $\mathbb{1}_{o_{t,i,j} \geq Q(o_{i,j})}$, taking the value of 1 for observations above or equal to the quantile of interest at the given point of latitude and longitude, and 0 otherwise. For cold extremes, the indicator function becomes $\mathbb{1}_{o_{t,i,j} \leq Q(o_{i,j})}$.

3. Calibration of extreme quantiles, where a quantile behaviour closer to the ground-truth (ERA 5) is considered superior to a quantile behaviour further away from it. We evaluate extreme quantile behaviour by considering quantiles between 90 and 99.9 for hot and wind extremes, and quantiles between 10 and 0.1 for cold extremes. We then produce quantile-quantile plots, where the extreme quantiles in the forecasts are plotted against the corresponding quantiles of ERA 5.

The three criteria jointly provide an overall picture of the performance of the models at forecasting near-surface temperature and wind extremes at global and regional (criterion 1), and local level (criterion 2), as well as of the tail behaviour of the models when faced with values at the edges or beyond the limits of the training distribution (criterion 3).

3 Results

In this section, we report the results of the model comparison performed according to the criteria outlined in Subsection 2.5. The aim here is both to provide a comparison between data-driven and physics-based models as a whole, as well as to identify relevant differences between the data-driven models themselves.

We start by providing an overview of the performance of different models globally and in individual regions when considering all observations, both extremes and not extreme. Figure 1 suggests that data-driven models perform better than ECMWF IFS HRES in virtually all regions and at virtually all lead times, with differences being most evident at shorter lead times. The difference between GraphCast and Pangu-Weather appears to be smaller overall, with GraphCast generally performing best in the tropics and Pangu best in the Extra-Tropics.

Figure 1. RMSE scorecard for 2m temperature (a) and 10m windspeed (b) at a global and regional scale, computed on all test observations. Blue shades indicate better performance than IFS HRES, red shades worse performance.

Figure 2 provides RMSE comparisons for the 5% most extremes observations globally and in each region, in accordance with criterion 1 (Subsection 2.5). Globally, GraphCast outperforms HRES in all variables at most time scales, while Pangu struggles with cold and wind extremes. However, regional comparisons reveal some more complex patterns: for cold extremes, data-driven models prevail in the Northern Hemisphere and in the Tropics, but struggle in the Southern Hemisphere, and in Europe at longer lead times. For hot and wind extremes, data-driven models appear to be overall superior, but they struggle in some densely populated regions, such as East Asia, North America and Europe. Additionally, HRES appears to be the best model for all variables in AusNZ and Antarctic at most lead times.

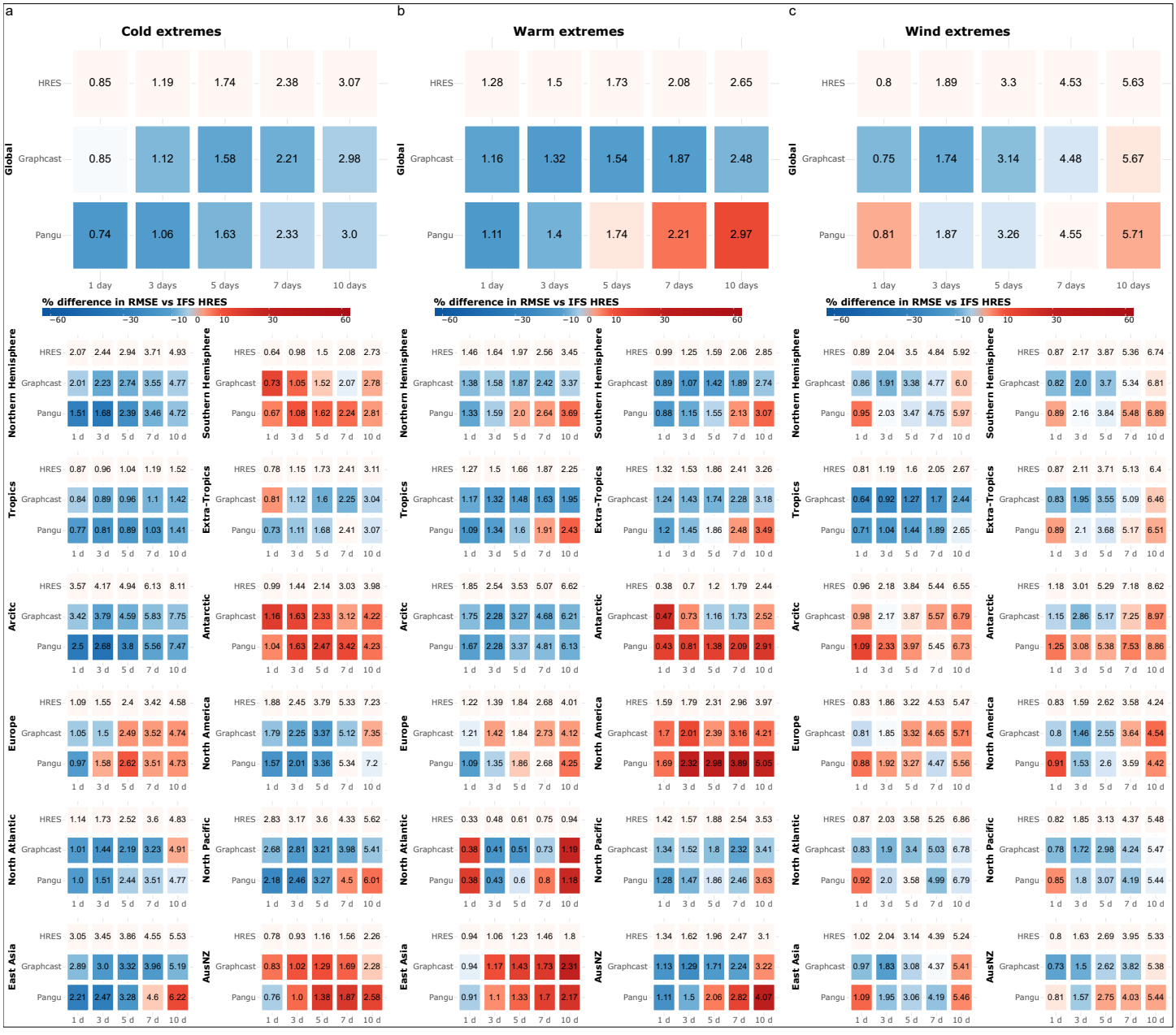

Figure 2. RMSE scorecard for cold (a), hot (b) and wind extremes (c) at a global and regional scale, computed on the (a) 5% lowest 2m temperature, (b) 5% highest 2m temperature and (c) 5% highest 10m windspeed observations, respectively.

Figure 3 repeats the analysis presented in Figure 2, but only on the most extreme events, namely the 1% most extreme observations for the region of interest. We can see how the difference in performance between models becomes very region and variable-dependent, with HRES outperforming the other models in the coldest events globally, mostly taking place in Antarctica, and GraphCast mostly outperforming HRES and Pangu in forecasting the warmest and windiest events. We also note that data-driven models outperform HRES in forecasting cold events in the tropics and Northern Hemisphere extra-tropics, with the exception of Europe, but are outperformed in the Southern Hemisphere, with the exception of East Asia. For hot extremes, data-driven models outperform HRES in several regions, but are outperformed in crucial densely inhabited areas, such as Europe, East-Asia and North America. For windspeed extremes, data-driven models outperform HRES in the extra-tropics, and have otherwise comparable performance to the physics-based model in most densely-inhabited areas, with the exception of Europe where HRES continues to prevail.

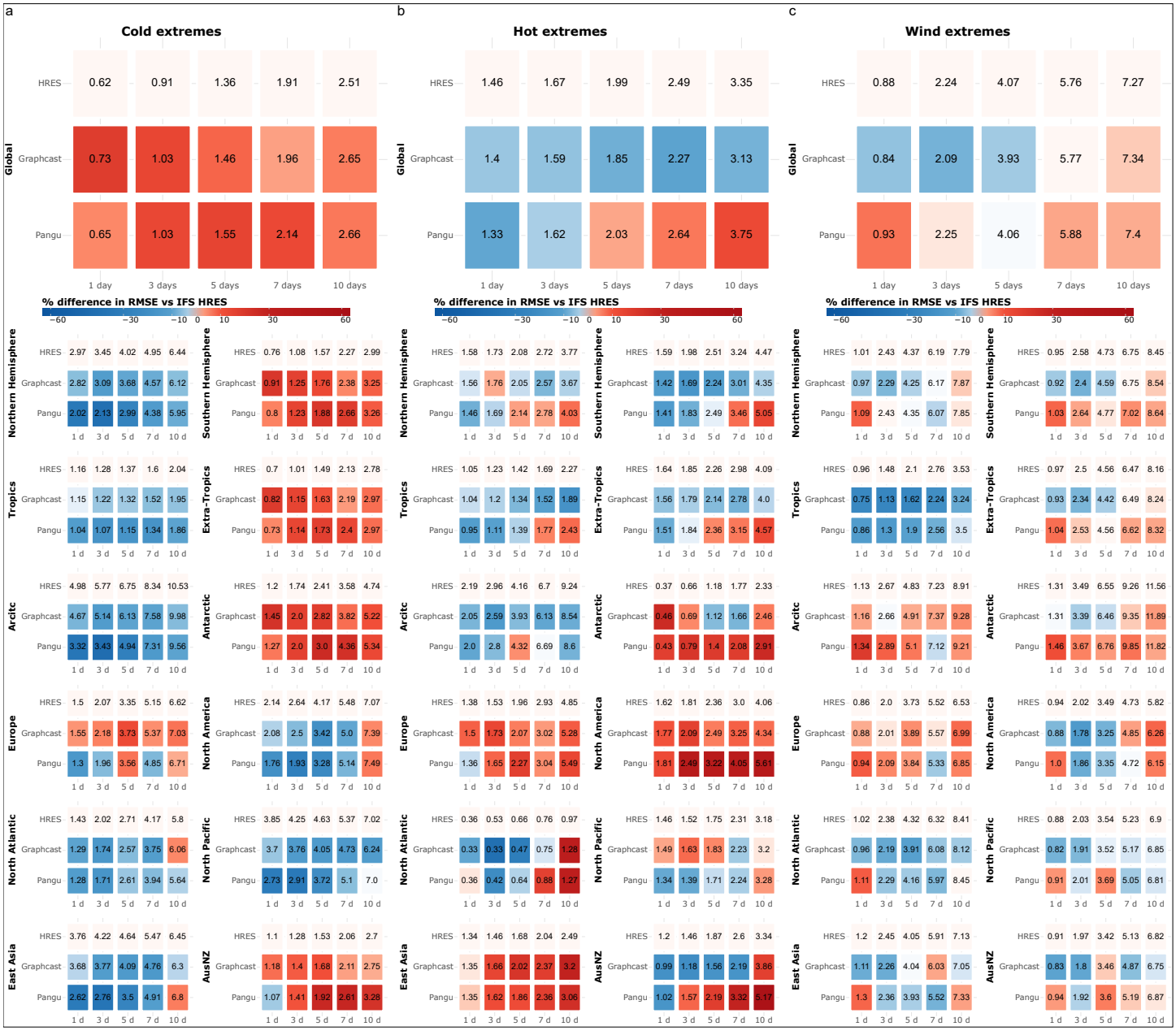

Figure 3. RMSE scorecard for cold (a), hot (b) and wind extremes (c) at a global and regional scale, computed on the (a) 1% lowest 2m temperature, (b) 1% highest 2m temperature and (c) 1% highest 10m windspeed observations, respectively.

A summary scorecard of Figures 1 - 3 is provided in Figure 4, showing which of the three models is best at forecasting cold, hot and windspeed extremes as well as 2m temperature and 10m windspeed overall. The summary scorecards confirm the patterns observed so far, suggesting that data-driven models are generally superior to HRES at forecasting 10m wind and 2m temperature when considering all observations. However, the summary scorecard also shows that the performance of data-driven models degrades relative to HRES when considering extreme quantiles, with HRES being overall superior at forecasting extremes in AusNZ and Antarctic, and mostly outperforming data-driven models in forecasting hot extremes in Europe, North America and East Asia. Nevertheless, HRES and data-driven models display comparable performance in forecasting cold and wind extremes in those regions.

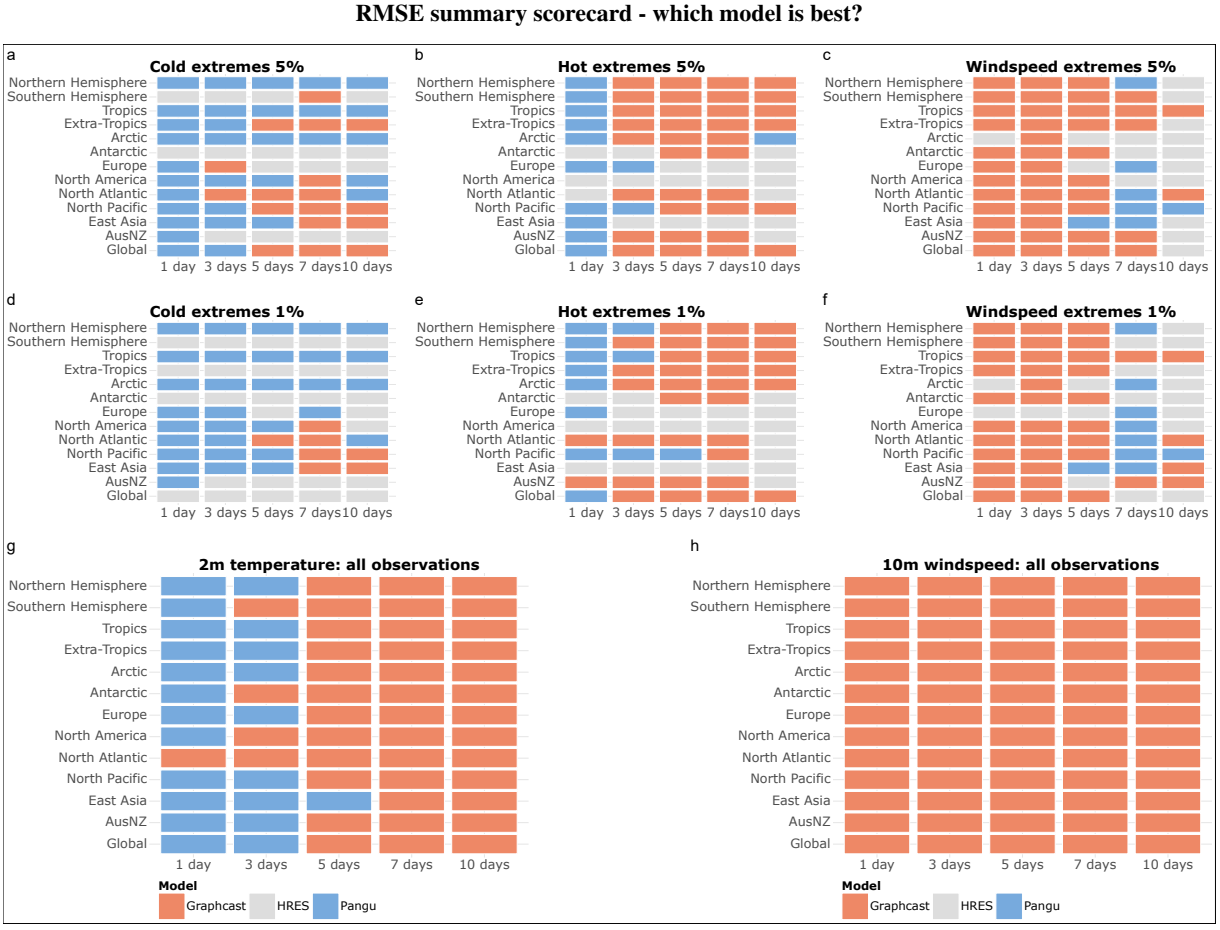

Figure 4. Best model in terms of RMSE computed on the (a) 5% lowest 2m temperature, (b) 5% highest 2m temperature, (c) 5% highest 10m windspeed, (d) 1% lowest 2m temperature, (e) 1% highest 2m temperature, (f) 1% highest 10m windspeed, and on (g) all 2m temperature observations and (h) all 10m windspeed observations.

Figure 5 and Figure 6 apply criterion 2 (Subsection 2.5) to the comparison between models, focusing on grid-point level differences. They display which model is best at forecasting the overall variables (Figure 5) and their extremes (Figure 6) for each pixel and lead time, with extremes being defined here as events in the 5% most extreme quantiles of all events at the given grid point during the test period. GraphCast appears to be the best model at forecasting 10m windspeed (Figure 5, column b) at all lead times, with the exception of few areas. Pangu-Weather, on the hand, performs better on 2m temperatures over land, especially at short lead times (Figure 5, column a), with data-driven models as a whole prevailing in most land areas, bar for 1-day lead times (a1).

When forecasting extremes (Figure 6), local patterns are more complex, and it is hard to distinguish consistent large-scale patterns apart from in a limited number of cases. In the tropics, GraphCast appears to be best at predicting hot and windspeed extremes (columns b and c), whereas Pangu-Weather is best at cold extremes (column a). Outside of the tropics, the situation is less clear, with HRES outperforming data-driven models in predictions of cold extremes in Antarctica, and generally over-performing in highly populated land areas, such as the US and Southeast Asia. Especially in the case of windpseed extremes, HRES is still outperforming data-driven models in most land areas. We also notice that data-driven models become worse relatively to HRES in forecasting hot and windspeed extremes at longer lead times

The main conclusion we can draw is that no model appears to be best overall, and that even at a given grid-point the best model is highly dependent on the variable we aim to forecast and the lead time of interest. When looking at regional scorecards (Figure 4), it was easier to draw some more general conclusions, but those conclusions do not necessary hold when shifting the focus to specific locations.

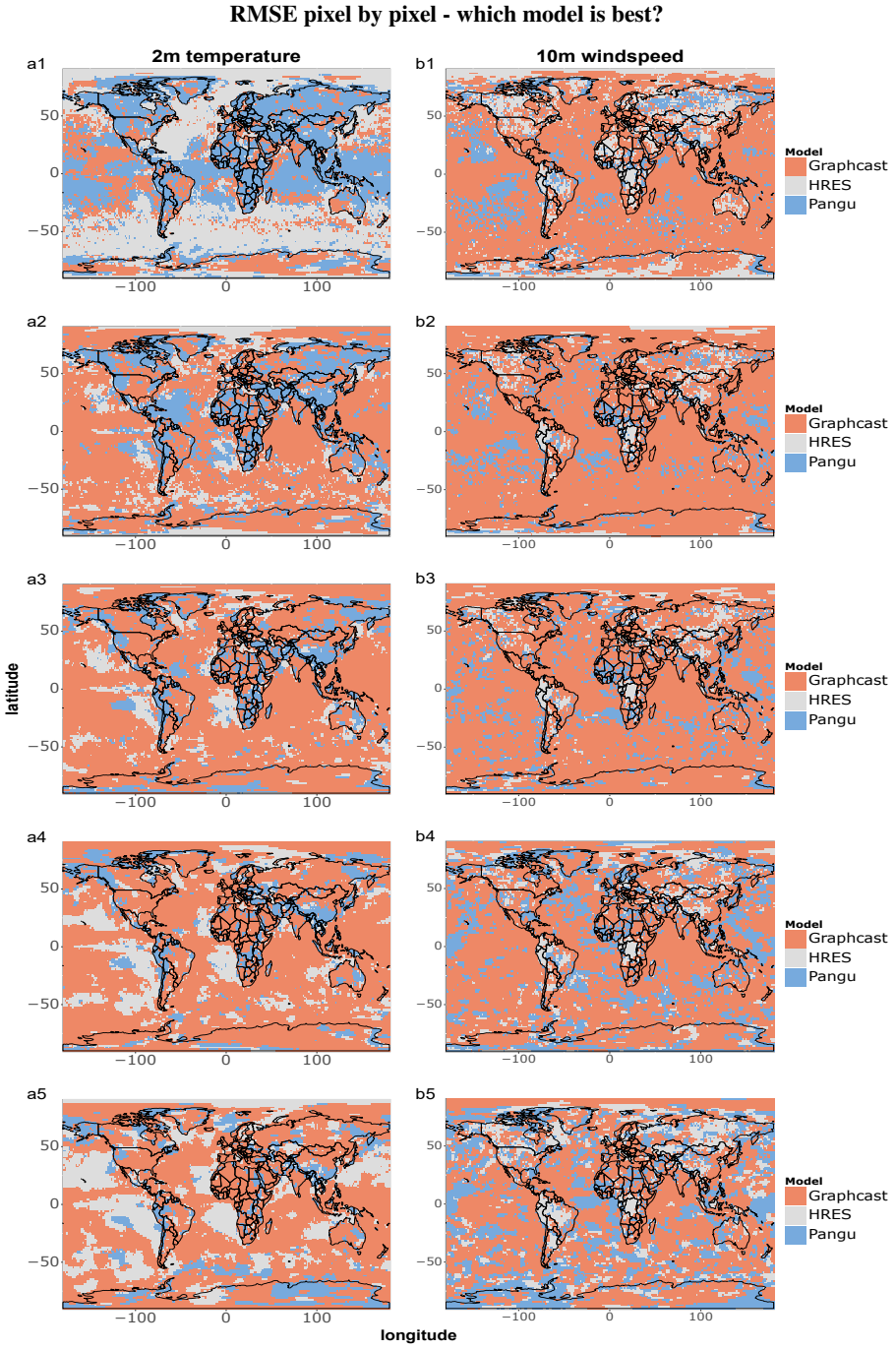

Figure 5. Single-gridpoint RMSE comparison for all observations of 2m temperatures (a) and 10m windspeed (b). X1) 1 day forecasts; X2) 3 days forecasts; X3) 5 days forecasts; X4) 7 days forecasts; X5) 10 days forecasts.

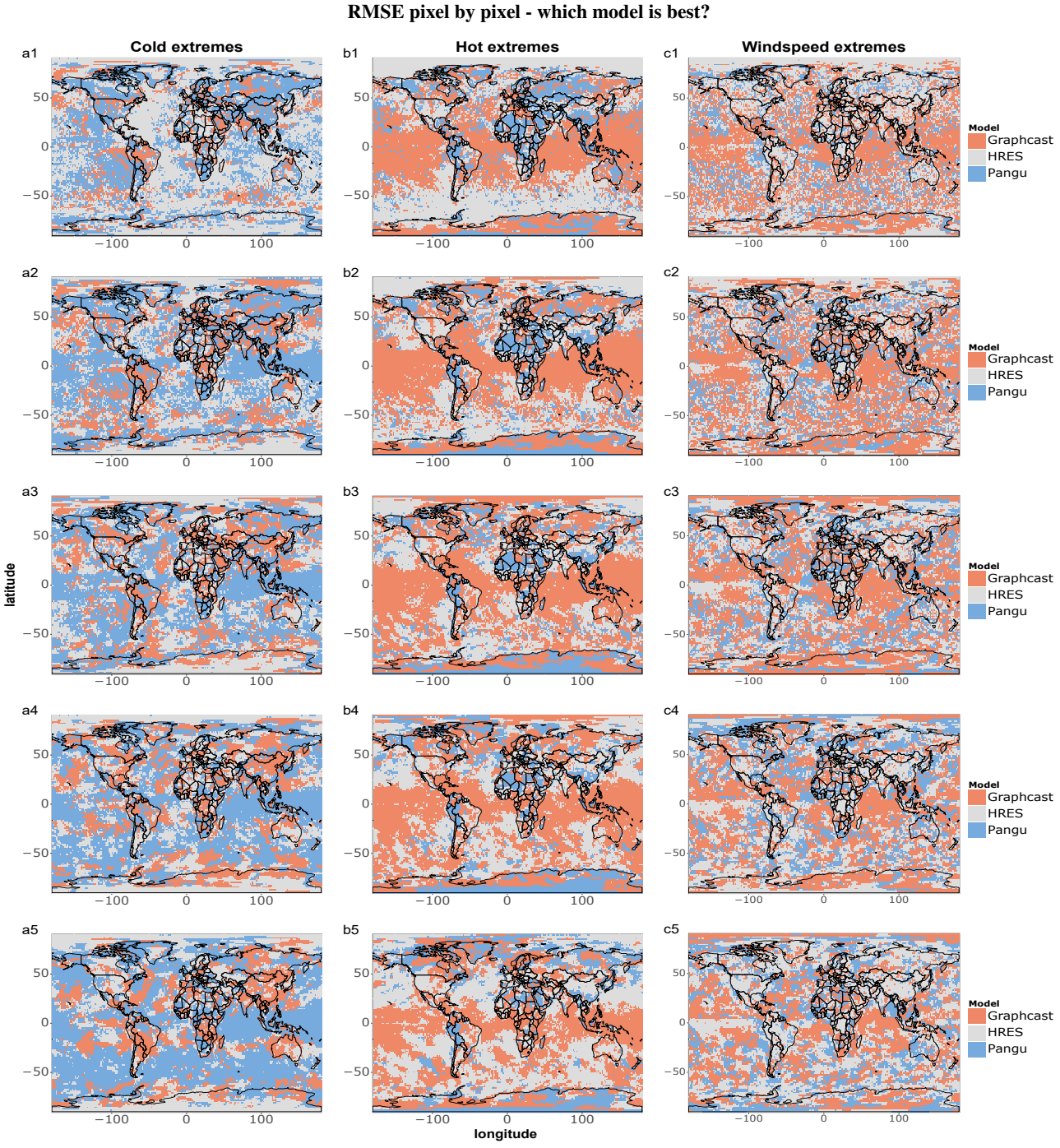

Figure 6. Single-gridpoint RMSE comparison for cold extremes (a), hot extremes (b) and winsdspeed extremes (c). The extremes are defined as in Figure 2. X1) 1 day forecasts; X2) 3 days forecasts; X3) 5 days forecasts; X4) 7 days forecasts; X5) 10 days forecasts.

Lastly, we compare the models on the basis of criterion 3 (Subsection 2.5), namely on the ability of different models to reproduce the tail behaviour of ERA 5. As in the previous cases, we start by looking at global extremes at multiple lead times (Figure 7), in order to assess the tail behaviour of the forecasts within the 10th most extreme quantile. Figure 7 suggests that all models appear to be well calibrated in the forecast of global cold extremes, while data-driven models tend to underestimate the magnitude of hot and windspeed extremes, especially at longer lead times.

Figures 8-10 show qq-plots of extreme quantiles in individual regions for 5-days forecasts. As in previous cases, regional patterns reveal further complexities in the behaviour of the three models. Figure 8 suggests that all models tend to underestimate cold extremes in the Arctic and North Pacific. Additionally, data-driven models tend to underestimate cold extremes in the Antarctic, and HRES and GraphCast also in Europe. The largest underestimation occurs in the Arctic, with the coldest observations being underestimated by 2-3 K by all models, on average.

Figure 9 shows a clearer pattern, with HRES displaying the best tail behaviour overall, and data-driven models underestimating extremes in several regions, including North America, East Asia, Europe, the Tropics and the North Pacific. AusNZ appears to be the only region where some of the models (HRES and GraphCast) overestimate the average magnitude of the extremes. Overall, the largest underestimations occur in North America, where the data-driven models underestimate the warmest observations by around 2 K, on average.

Figure 10 displays a similar pattern to Figure 9, with HRES displaying almost perfect tail behaviour, and data-driven models tending to slightly underestimate windspeed extremes in all regions. Differences between GraphCast and Pangu-Weather are small overall, with the most notable differences being GraphCast outperforming Pangu in the Tropics, and Pangu outperforming GraphCast in Europe. Those differences are largely in line with what observed in Figures 2-6.

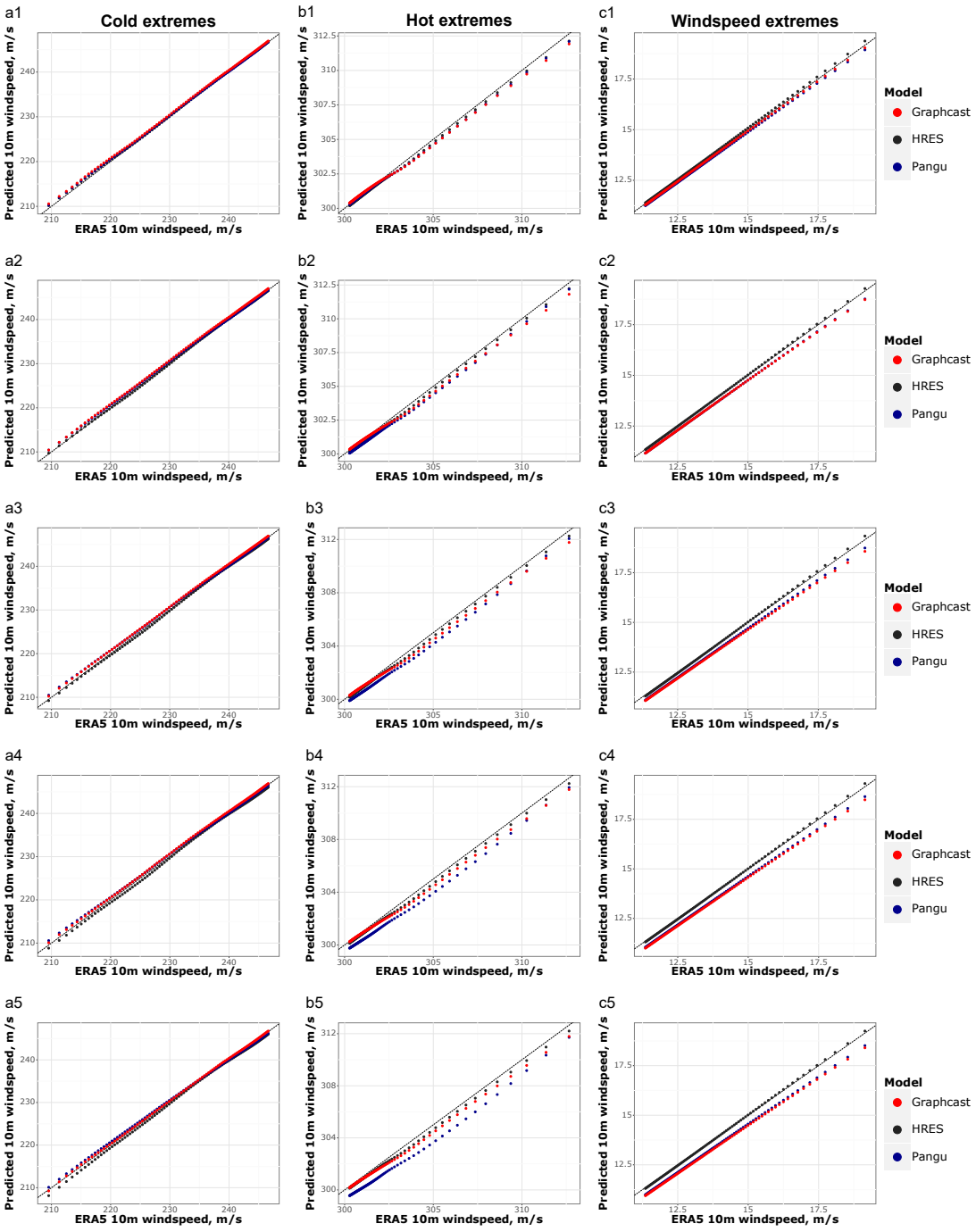

Figure 7. QQ plots of 1 (X1), 3 (X2), 5 (X3), 7 (X4) and 10 (X5) days ahead 10% most extreme 2m cold (a) and hot temperatures (b) and 10m windspeed (c) vs ground truth (ERA 5).

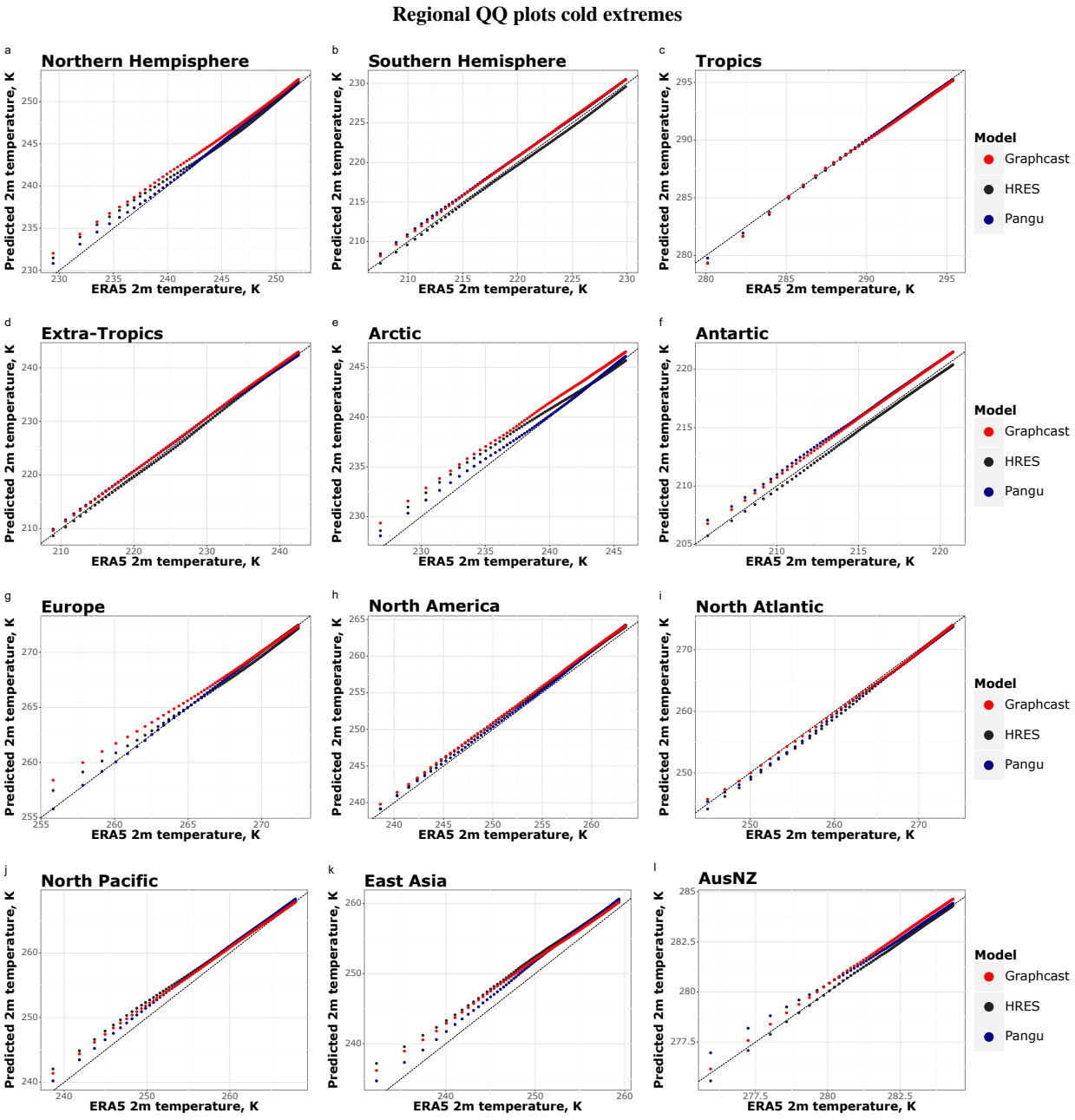

Figure 8. Region-based QQ plots of 5-day forecasts for the 10% coldest days in terms of ERA 5 2m temperatures.

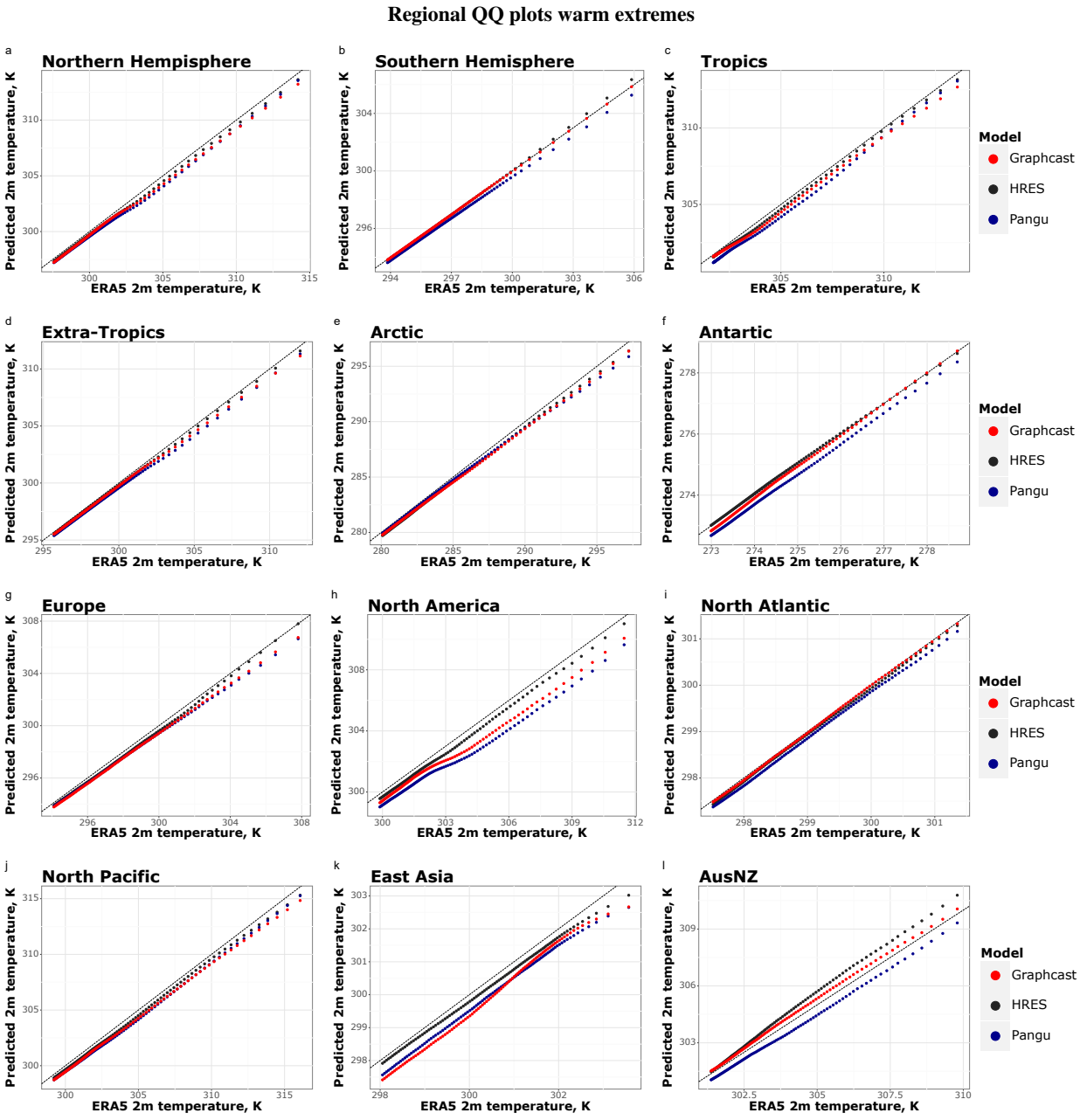

Figure 9. Region-based QQ plots of 5 days forecasts for the 10% hottest days in terms of ERA 5 2m temperatures.

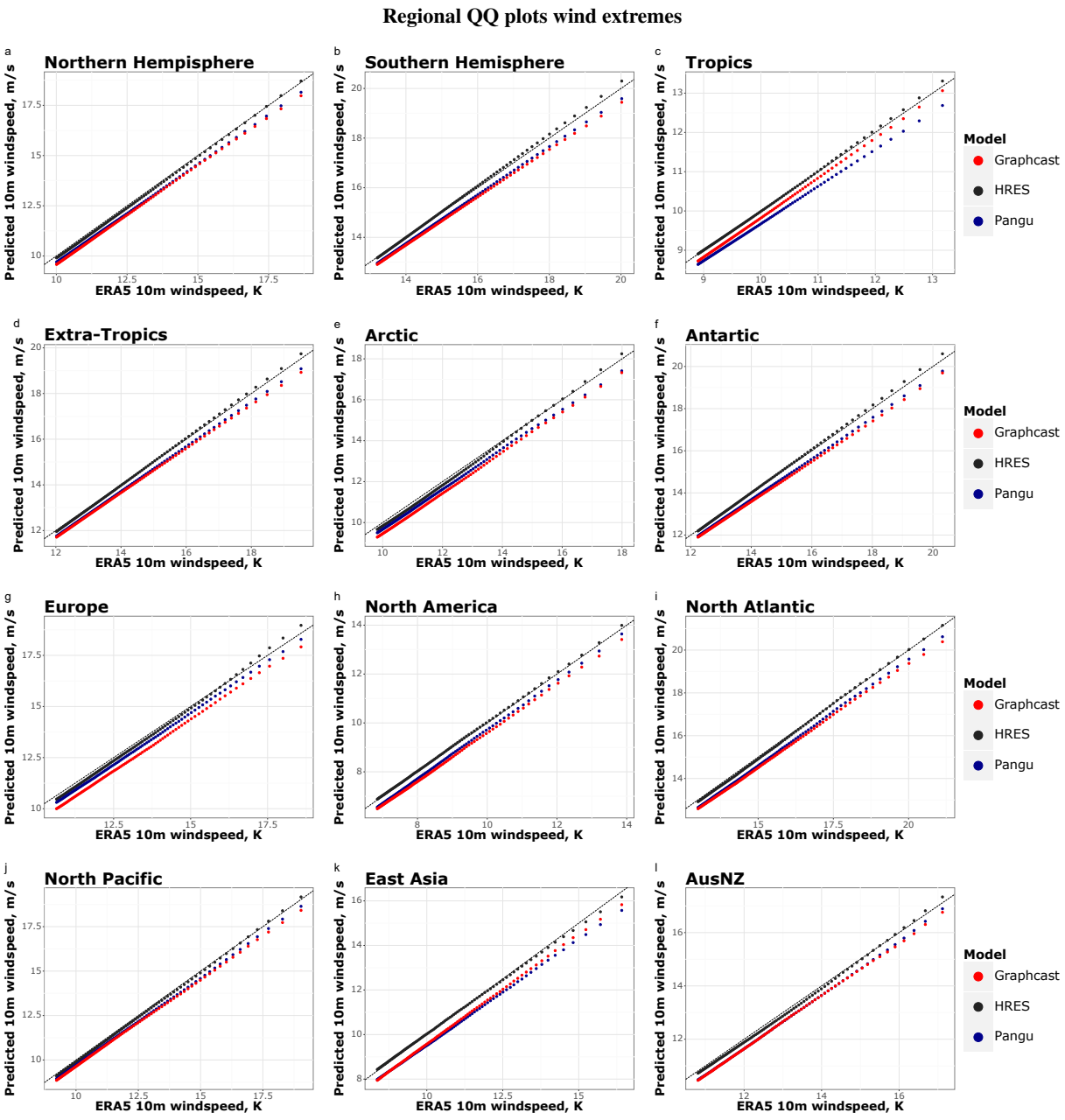

Figure 10. Region-based QQ plots of 5 days forecasts for the 10% windiest days in terms of ERA 5 10m windspeed.

The comparison between HRES and reanalysis-based data driven models, included in Appendix A, mostly supports the findings in the main (Figures A1-A6). Data-driven models as a whole improve on the physics-based model in terms of average skill (bar for short-term 2m temperature forecasts; Figures A1 and A4), and are also competitive in terms of extremes (Figures A2-A4), but not in all regions nor at all lead times. In particular, HRES appears to still outperform all data-driven models in forecasting cold spells at short lead times, but this might also partially be the result of the fact that, differently from in the main, HRES t=0 was used as ground truth for the physics-based model in Appendix A. We also notice that data-driven models display limited skill in 10-days ahead forecasts of windspeed extremes, and that HRES, as in the main, still outperforms data-driven models in forecasting certain types of extremes in some densely populated regions (e.g. windspeed and cold extremes in Europe and hot extremes in East Asia).

For grid-point level performance (Figures A5 and A6), data-driven models are highly competitive in terms of average skill (Figure A5), where we notice the impressive performance of FuXi in forecasting 2m temperature at longer lead times (column a), and the fact that Graphcast and FuXi appear overall superior to Pangu-Weather. In terms of extremes (Figure A6), HRES appears to still be superior to data-driven models in forecasting extremes over land at shorter lead times, with data-driven models gradually catching up in the medium-range. We also note the improved performance of Pangu-Weather relative to other models when forecasting cold and windspeed extremes. Overall, though, the choice of best model appears to be very grid-point, variable and lead time dependent, as for the comparisons in the main.

4 Discussion and Conclusions

This paper has analysed the performance of ECMWF IFS HRES, GraphCast and Pangu-Weather in forecasting near-surface temperature and windspeed extremes up to 10 days ahead in a semi-operational setting. Following Watson (2022), the models have been evaluated with the help of three criteria (Subsection 2.5) assessing the forecast performance (Criteria 1 and 2) and the calibration of the forecasts in the tails of the distribution (criterion 3). The results suggest that data-driven models are superior to HRES in the task of forecasting 2m temperature and 10m windspeed on average in most regions at all time scales (Figures 4, 5), but that the picture is more nuanced when it comes to forecasting weather extremes. HRES is still the best model at forecasting several types of extremes in densely populated areas at multiple times, including hot extremes in North America and East Asia, and windspeed extremes in Europe (Figures 4, 6). Furthermore, HRES appears to generally outperform data-driven models over land, whereas data-driven models often perform best over sea (Figures 5 and 6). We speculate that this may partially depend on the stronger spatial heterogeneity of extremes over land regions, where the larger number of variables and physics-based framework of HRES provide an advantage.

An additional finding is that the data-driven models perform best in relative terms at shorter lead times, whereas HRES performs best in relative terms at longer lead times (Figures 1– 3). HRES appears also to be overall best in terms of tail calibration (Figures 7– 10), even though differences between HRES and data-driven models are small for forecasts of global extremes,

especially at shorter lead times (Figure 7). Differences between the two data-driven models appear to be overall small, with GraphCast oftentimes performing better in the Tropics, and Pangu-Weather in the midlatitudes (Figures 5, 8–10).

We conclude that data-driven models can compete with physics-based models when forecasting near-surface temperature and wind extremes, but that the choice of best model depends strongly on region, lead time, type of extreme and in some cases even level of extremeness. Specifically, we find that physics-based models might still be the best choice in many densely-populated areas of Earth. These results largely hold also for the comparison between reanalysis-based models displayed in Appendix A. 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.

Yet, as suggested by previous literature, some additional challenges need to be addressed before data-driven models may be fully implemented operationally, including the lack of uncertainty information provided by the deterministic forecasts (Molina et al., 2023; de Burgh-Day and Leeuwenburg, 2023; Scher and Messori, 2021; Clare et al., 2021) and the lack of physical constraints in the forecasts generated by the models (Kashinath et al., 2021; Beucler et al., 2020). Moreover, with the exception of GraphCast, none of the data-driven models that we analysed here forecasts precipitation, which when extreme is a key meteorological hazard. Finally, further evaluations of extreme behaviour may be necessary. Our analysis is limited to a narrow range of near-surface extremes and, due to current data availability, to extremes occurring in 2020. This limits our ability to draw conclusions on long-term performance. The short time period considered also exposes our results to sensitivity to low-frequency modes of climate variability, which modulate the occurrence of extreme events and may also affect their predictability (Goddard and Gershunov, 2020; Luo and Lau, 2020; Chartrand and Pausata, 2020). Additionally, as highlighted by Watson (2022), raw measures of performance and qq-plots should also be complemented by a careful study of weather charts of case studies. We would therefore encourage more comprehensive evaluations in the near-future, as more data become available, and deep-learning models are extended to produce forecasts of other relevant variables for weather extremes (e.g. wind gusts and precipitation).

Code and data availability. The forecasts generated by all models are freely available through the WeatherBench 2 (Rasp et al., 2024). All the data-driven models are trained using the ERA 5 reanalysis dataset (Hersbach et al., 2020), which is freely available through the Copernicus Climate Change Service at https://doi.org/10.24381/cds.adbb2d47 and https://doi.org/10.24381/cds.bd0915c6, as well as through the WeatherBench 2 (Rasp et al., 2024). The code used to train the data-driven models included in the comparison are provided by the authors of the models themselves, and details an how to access the code and pre-trained models are provided in their respective papers (Bi et al., 2023; Lam et al., 2023; Chen et al., 2023b). The code developed by the authors of this paper to perform the comparisons and generate the plots included here is available on Zenodo at https://zenodo.org/doi/10.5281/zenodo.1093248 (Olivetti, 2024), as well as on the Github page of the corresponding author, Leonardo Olivetti.

Author contributions. The authors are jointly responsible for the conceptualisation of this work including the visualisations, and all the revision and editing of the submitted manuscript. L. Olivetti has developed the code used for the model comparisons and to generate the visualisations, and written most of the original draft. G.Messori has acquired the funding and other resources necessary to conduct this research and provided extensive supervision.

Competing interests. The authors declare no conflicts of interest relevant to this study.

Appendix A: Comparison of Reanalysis-based Data-driven Models

Here we provide global and regional scorecards and grid-point level comparisons for data-driven models using ERA5 reanalysis data as input. Following the WeatherBench 2 (Rasp et al., 2024), we attempt to make the comparison between reanalysis-based data driven models and HRES as fair as possible by using IFS HRES t=0 as ground truth for IFS HRES, instead of ERA 5.

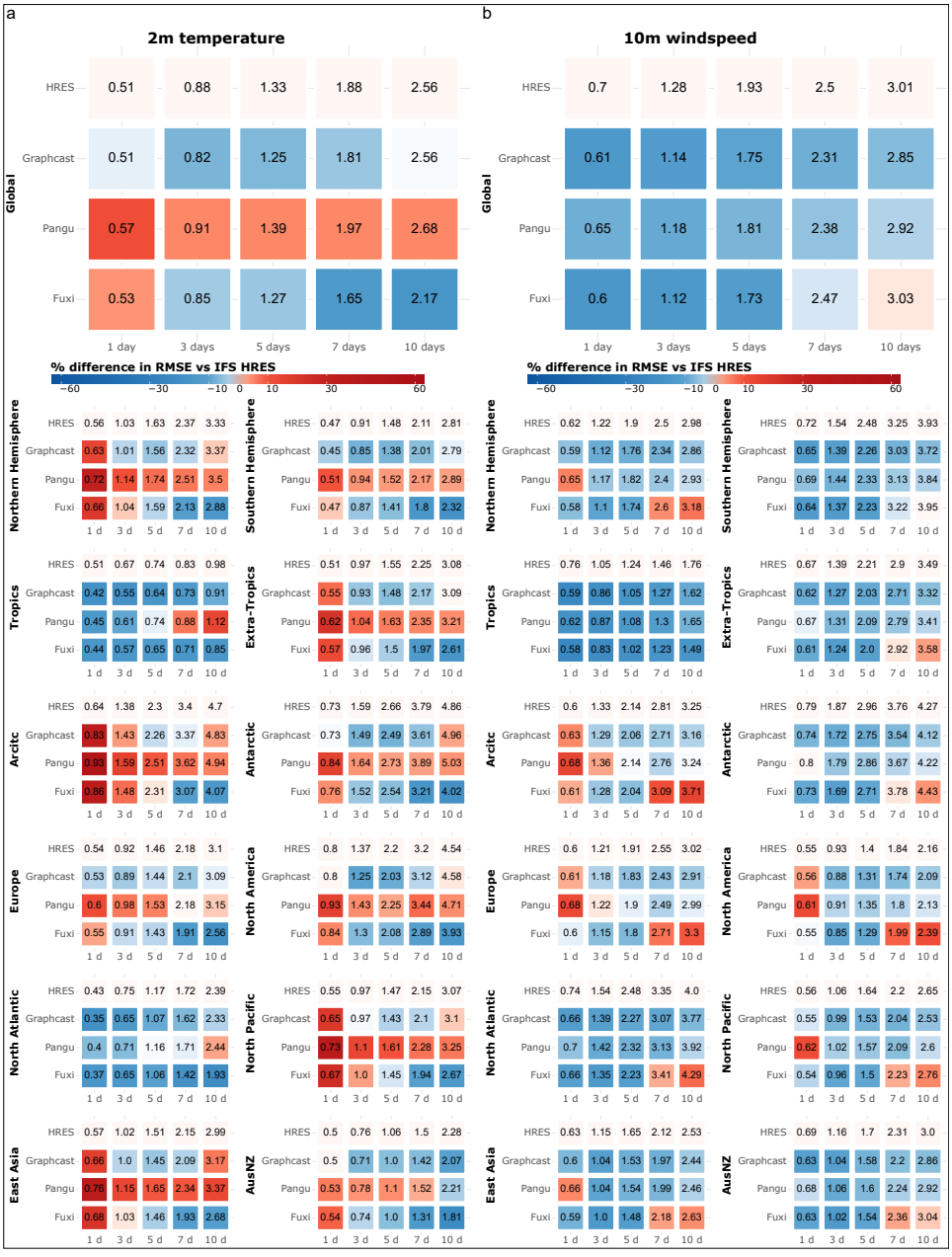

Figure A1. RMSE scorecard for 2m temperature (a) and 10m windspeed (b) at a global and regional scale, computed on all test observations. Blue shades indicate better performance than IFS HRES, red shades worse performance.

Figure A2. RMSE scorecard for cold (a), hot (b) and wind extremes (c) at a global and regional scale, computed on the (a) 5% lowest 2m temperature, (b) 5% highest 2m temperature and (c) 5% highest 10m windspeed observations, respectively.

Figure A3. RMSE scorecard for cold (a), hot (b) and wind extremes (c) at a global and regional scale, computed on the (a) 1% lowest 2m temperature, (b) 1% highest 2m temperature and (c) 1% highest 10m windspeed observations, respectively.

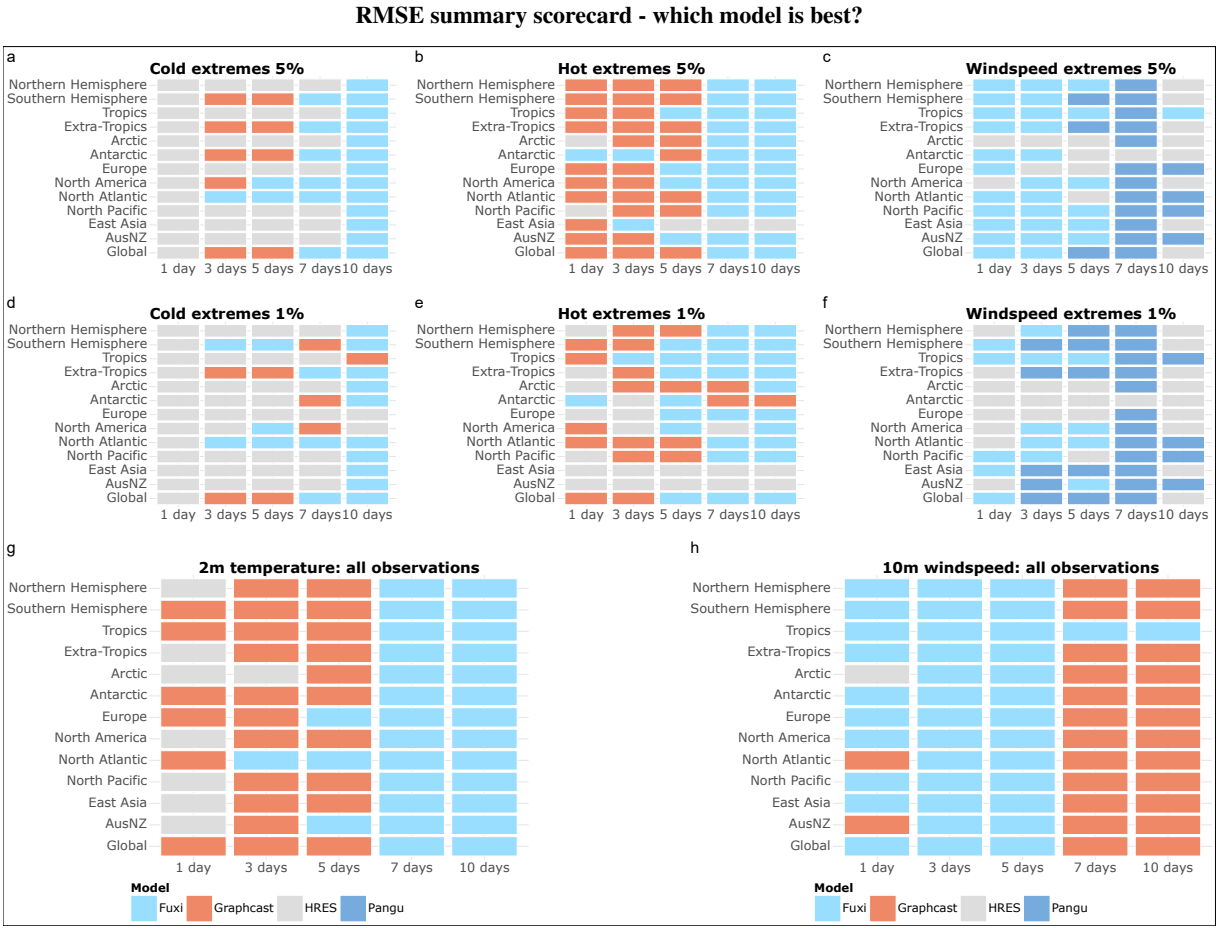

Figure A4. Best model in terms of RMSE computed on the (a) 5% lowest 2m temperature, (b) 5% highest 2m temperature, (c) 5% highest 10m windspeed, (d) 1% lowest 2m temperature, (e) 1% highest 2m temperature, (f) 1% highest 10m windspeed, and on (g) all 2m temperature observations and (h) all 10m windspeed observations.

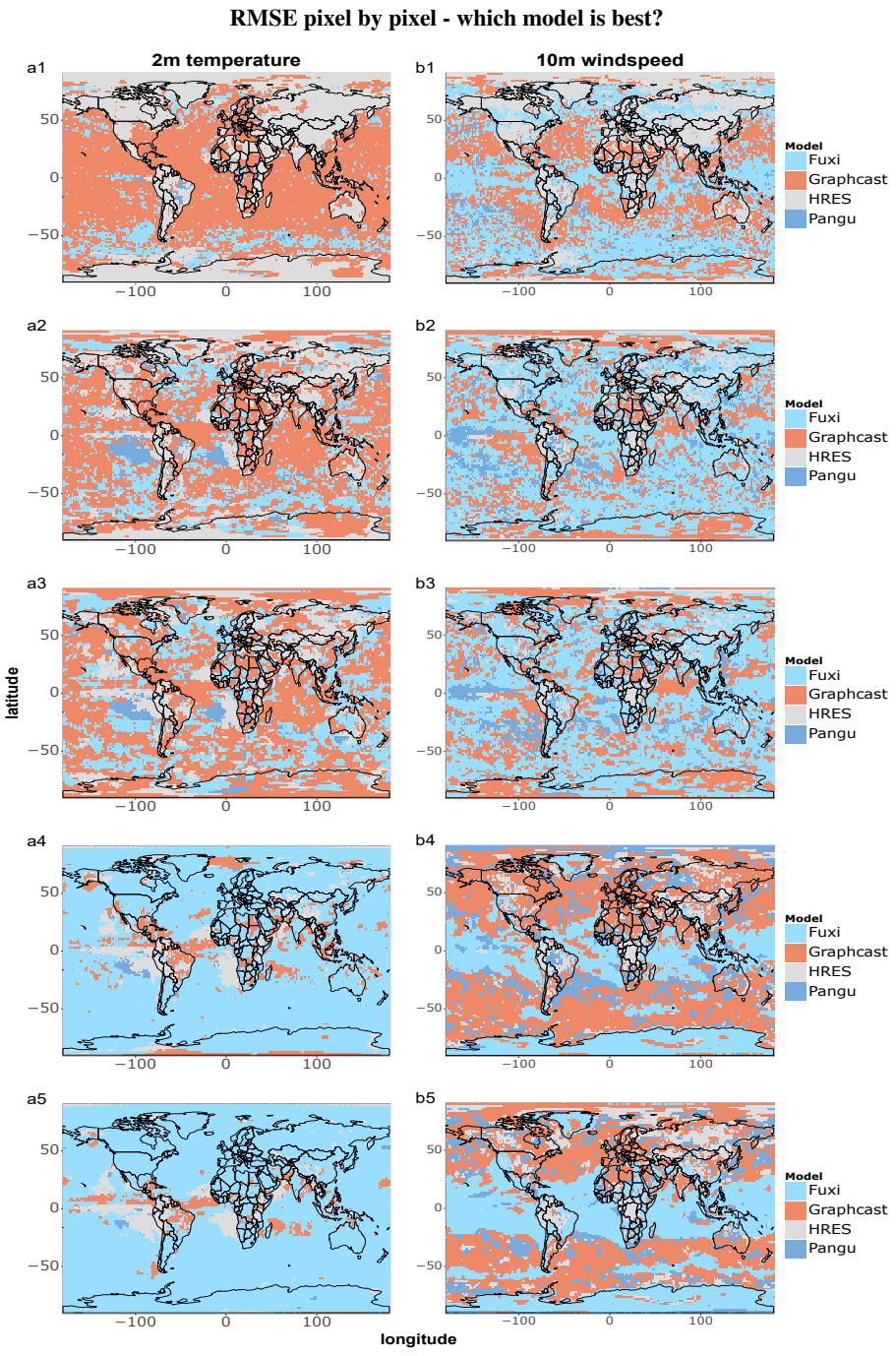

Figure A5. Single-gridpoint RMSE comparison for all observations of 2m temperatures (a) and 10m windspeed (b). X1) 1 day forecasts; X2) 3 days forecasts; X3) 5 days forecasts; X4) 7 days forecasts; X5) 10 days forecasts.

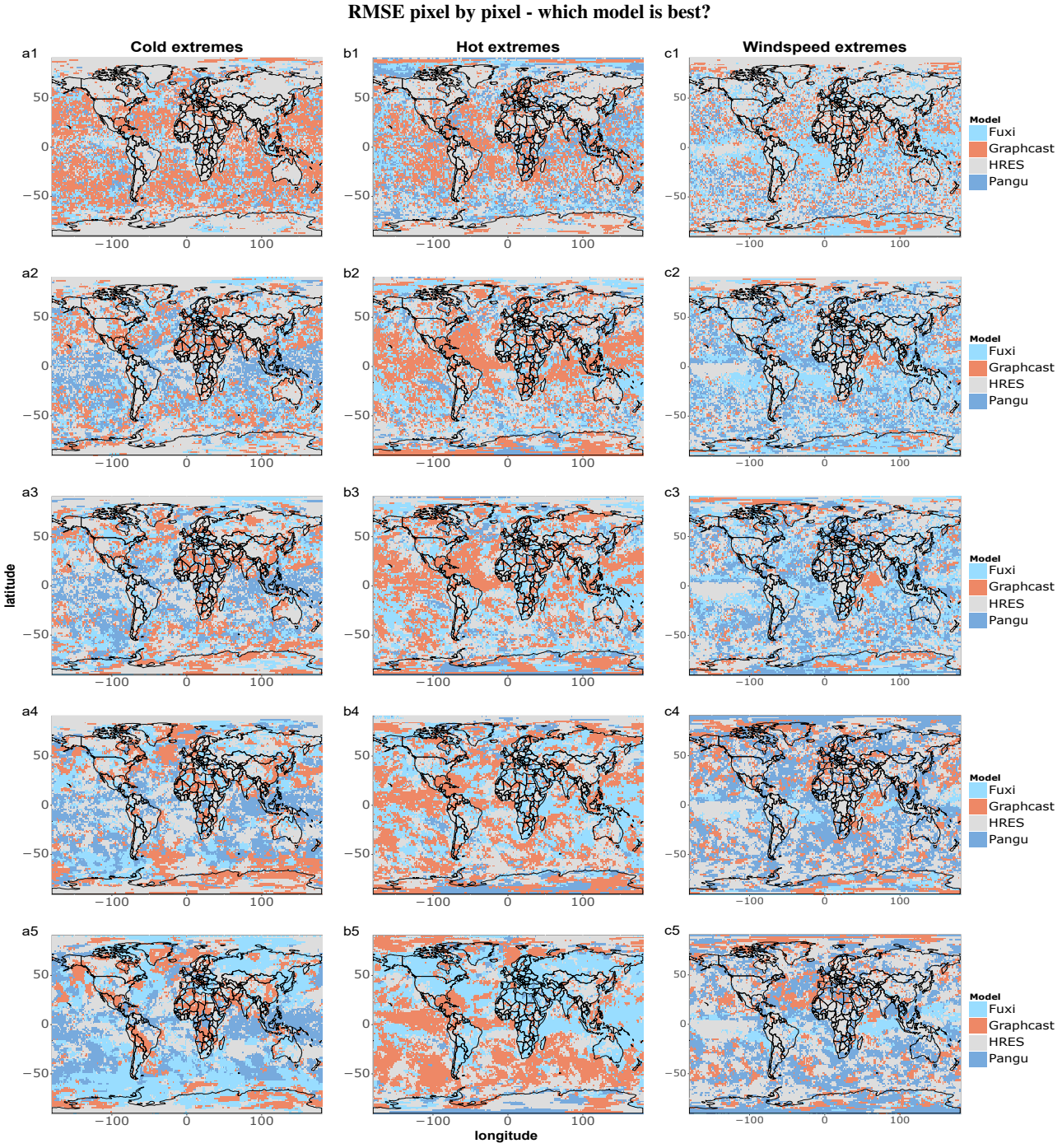

Figure A6. Single-gridpoint RMSE comparison for cold extremes (a), hot extremes (b) and winsdspeed extremes (c). The extremes are defined as in Figure 2. X1) 1 day forecasts; X2) 3 days forecasts; X3) 5 days forecasts; X4) 7 days forecasts; X5) 10 days forecasts.

Acknowledgements. The authors thankfully acknowledge the support of the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project CENAE: compound Climate Extremes in North America and Europe: from dynamics to predictability, Grant Agreement No. 948309). The computations and storage were aided by resources in project NAISS NAISS 2023/22-1356B and NAISS 2023/23-665, provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at C3SE, partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Ben-Bouallegue, Z., Clare, M. C. A., Magnusson, L., Gascon, E., Maier-Gerber, M., Janousek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F.: The rise of data-driven weather forecasting, https://doi.org/10.48550/arXiv.2307.10128, preprint at arXiv:2307.10128, 2023.

Beucler, T., Pritchard, M., Gentine, P., and Rasp, S.: Towards Physically-Consistent, Data-Driven Models of Convection, in: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, pp. 3987–3990, https://doi.org/10.1109/IGARSS39084.2020.9324569, iSSN: 2153-7003, 2020.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast, https://doi.org/10.48550/arXiv.2211.02556, preprint at arXiv:2211.02556, 2022.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, pp. 1–6, https://doi.org/10.1038/s41586-023-06185-3, publisher: Nature Publishing Group, 2023.

Blanchonnet, H.: IFS documentation, https://www.ecmwf.int/en/publications/ifs-documentation, 2022.

Chartrand, J. and Pausata, F. S. R.: Impacts of the North Atlantic Oscillation on winter precipitations and storm track variability in southeast Canada and the northeast United States, Weather and Climate Dynamics, 1, 731–744, https://doi.org/10.5194/wcd-1-731-2020, publisher: Copernicus GmbH, 2020.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., and Ouyang, W.: FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead, https://doi.org/10.48550/arXiv.2304.02948, preprint at 10.48550/arXiv.2304.02948, 2023a.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H.: FuXi: a cascade machine learning forecasting system for 15-day global weather forecast, npj Climate and Atmospheric Science, 6, 1–11, https://doi.org/10.1038/s41612-023-00512-1, number: 1 Publisher: Nature Publishing Group, 2023b.

Cisneros, D., Richards, J., Dahal, A., Lombardo, L., and Huser, R.: Deep graphical regression for jointly moderate and extreme Australian wildfires, https://doi.org/10.48550/arXiv.2308.14547, preprint at 10.48550/arXiv.2308.14547, 2023.

Clare, M. C., Jamil, O., and Morcrette, C. J.: Combining distribution-based neural networks to predict weather forecast probabilities, Quarterly Journal of the Royal Meteorological Society, 147, 4337–4357, https://doi.org/10.1002/qj.4180, 2021.

de Burgh-Day, C. O. and Leeuwenburg, T.: Machine Learning for numerical weather and climate modelling: a review, EGUsphere, pp. 1–48, https://doi.org/10.5194/egusphere-2023-350, preprint at https://egusphere.copernicus.org/preprints/2023/egusphere-2023-350/, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations, https://openreview.net/forum?id=YicbFdNTTy, 2020.

ECMWF: ECMWF | HRES Scorecard, https://sites.ecmwf.int/ifs/scorecards/scorecards-47r3HRES.html, 2024.

Goddard, L. and Gershunov, A.: Impact of El Niño on Weather and Climate Extremes, in: El Niño Southern Oscillation in a Changing Climate, pp. 361–375, American Geophysical Union (AGU), ISBN 978-1-119-54816-4, https://doi.org/10.1002/9781119548164.ch16, section: 16 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119548164.ch16, 2020.

Guastavino, S., Piana, M., Tizzi, M., Cassola, F., Iengo, A., Sacchetti, D., Solazzo, E., and Benvenuto, F.: Prediction of severe thunderstorm events with ensemble deep learning and radar data, Scientific Reports, 12, 20 049, https://doi.org/10.1038/s41598-022-23306-6, number: 1 Publisher: Nature Publishing Group, 2022.

Hall, T., Brooks, H. E., and Doswell, C. A.: Precipitation Forecasting Using a Neural Network, Weather and Forecasting, 14, 338–345, https://doi.org/10.1175/1520-0434(1999)014<0338:PFUANN>2.0.CO;2, publisher: American Meteorological Society Section: Weather and Forecasting, 1999.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, [Dataset], 2020.

Hu, Y., Chen, L., Wang, Z., and Li, H.: SwinVRNN: A Data-Driven Ensemble Forecasting Model via Learned Distribution Perturbation, Journal of Advances in Modeling Earth Systems, 15, e2022MS003 211, https://doi.org/10.1029/2022MS003211, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003211, 2023.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., Hassanzadeh, P., and Prabhat, n.: Physics-informed machine learning: case studies for weather and climate modelling, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20200 093, https://doi.org/10.1098/rsta.2020.0093, publisher: Royal Society, 2021.

Keisler, R.: Forecasting Global Weather with Graph Neural Networks, https://doi.org/10.48550/arXiv.2202.07575, preprint at http://arxiv.org/abs/2202.07575, 2022.

Kron, W., Löw, P., and Kundzewicz, Z. W.: Changes in risk of extreme weather events in Europe, Environmental Science & Policy, 100, 74–83, https://doi.org/10.1016/j.envsci.2019.06.007, 2019.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., Ravuri, S., Ewalds, T., Alet, F., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Stott, J., Vinyals, O., Mohamed, S., and Battaglia, P.: GraphCast: Learning skillful medium-range global weather forecasting, https://doi.org/10.48550/arXiv.2212.12794, preprint at 10.48550/arXiv.2212.12794, 2022.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, publisher: American Association for the Advancement of Science, 2023.

Luo, M. and Lau, N.-C.: Summer heat extremes in northern continents linked to developing ENSO events, Environmental Research Letters, 15, 074 042, https://doi.org/10.1088/1748-9326/ab7d07, publisher: IOP Publishing, 2020.

Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D. N., Domeisen, D. I. V., Feser, F., Koszalka, I., Kreibich, H., Pantillon, F., Parolai, S., Pinto, J. G., Punge, H. J., Rivalta, E., Schröter, K., Strehlow, K., Weisse, R., and Wurpts, A.: Impact Forecasting to Support Emergency Management of Natural Hazards, Reviews of Geophysics, 58, https://doi.org/10.1029/2020RG000704, publisher: John Wiley & Sons, Ltd, 2020.

Molina, M. J., O'Brien, T. A., Anderson, G., Ashfaq, M., Bennett, K. E., Collins, W. D., Dagon, K., Restrepo, J. M., and Ullrich, P. A.: A Review of Recent and Emerging Machine Learning Applications for Climate Variability and Weather Phenomena, Artificial Intelligence for the Earth Systems, 2, https://doi.org/10.1175/AIES-D-22-0086.1, publisher: American Meteorological Society Section: Artificial Intelligence for the Earth Systems, 2023.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A.: ClimaX: A foundation model for weather and climate, https://doi.org/10.48550/arXiv.2301.10343, preprint at 10.48550/arXiv.2301.10343, 2023.

Olivetti, L.: LeonardoOlivetti/Do-data-driven-models-beat-numerical-models-in-forecasting-weather-extremes-: First developmental version, https://doi.org/10.5281/zenodo.10932486 [software], 2024.

Olivetti, L. and Messori, G.: Advances and prospects of deep learning for medium-range extreme weather forecasting, Geoscientific Model Development, 17, 2347–2358, https://doi.org/10.5194/gmd-17-2347-2024, publisher: Copernicus GmbH, 2024.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, https://doi.org/10.48550/arXiv.2202.11214, preprint at 10.48550/arXiv.2202.11214, 2022.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 203, https://doi.org/10.1029/2020MS002203, 2020.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Bouallegue, Z. B., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F.: WeatherBench 2: A benchmark for the next generation of data-driven global weather models, https://doi.org/10.48550/arXiv.2308.15560, preprint at arXiv:2308.15560, 2024.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G.: The Graph Neural Network Model, IEEE Transactions on Neural Networks, 20, 61–80, https://doi.org/10.1109/TNN.2008.2005605, 2009.

Scher, S. and Messori, G.: Ensemble Methods for Neural Network-Based Weather Forecasts, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2020MS002331, 2021.

Schizas, C., Michaelides, S., Pattichis, C., and Livesay, R.: Artificial neural networks in forecasting minimum temperature (weather), in: 1991 Second International Conference on Artificial Neural Networks, pp. 112–114, 1991.

Watson, P. A. G.: Machine learning applications for weather and climate need greater focus on extremes, Environmental Research Letters, 17, 111 004, https://doi.org/10.1088/1748-9326/ac9d4e, publisher: IOP Publishing, 2022.

World Meteorological Organization: Early warnings for all: Executive action plan 2023-2027, https://www.preventionweb.net/publication/early-warnings-all-executive-action-plan-2023-2027, https://www.preventionweb.net/publication/early-warnings-all-executive-action-plan-2023-2027, 2022.

Xu, W., Chen, K., Han, T., Chen, H., Ouyang, W., and Bai, L.: ExtremeCast: Boosting Extreme Value Prediction for Global Weather Forecast, https://doi.org/10.48550/arXiv.2402.01295, preprint at arXiv:2402.01295 [cs], 2024.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J.: Skilful nowcasting of extreme precipitation with NowcastNet, Nature, 619, 526–532, https://doi.org/10.1038/s41586-023-06184-4, number: 7970 Publisher: Nature Publishing Group, 2023.