An extension of the WeatherBench 2 to binary hydroclimatic forecasts

Tongtiegang Zhao[1], Qiang Li[1], Tongbi Tu[1], and Xiaohong Chen[1]

[1] Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Civil Engineering, Sun Yat-Sen University, Guangzhou 510275, China;

5 *Correspondence to*: Tongtiegang Zhao (zhaottg@mail.sysu.edu.cn) and Qiang Li (liqiang65@mail2.sysu.edu.cn).

Abstract: Binary forecasts on hydroclimatic extremes play a critical part in disaster prevention and risk management. While the recent WeatherBench 2 provides a versatile framework for the verification of deterministic and ensemble forecasts, this paper presents an extension to binary forecasts on the occurrence versus non-occurrence of hydroclimatic extremes. Specifically, sixteen verification metrics on the accuracy and discrimination of binary forecasts are employed and scorecards

10 are generated to showcase the predictive performance. A case study is devised for binary forecasts of wet and warm extremes obtained from both deterministic and ensemble forecasts generated by three data-driven models, i.e., Pangu-Weather, GraphCast and FuXi, and two numerical weather prediction products, i.e., ECMWF's IFS HRES and IFS ENS. The results show that the receiver operating characteristic skill score (ROCSS) serves as a suitable metric due to its relative insensitivity to the rarity of hydroclimatic extremes. For wet extremes, the GraphCast tends to outperform the IFS HRES with the total

15 precipitation of ERA5 data as ground truth. For warm extremes, the Pangu-Weather, GraphCast and FuXi tends to be more skilful than the IFS HRES within 3-day lead time but become less skilful as lead time increases. In the meantime, the IFS ENS tends to provide skilful forecasts of both wet and warm extremes at different lead times and at the global scale. Through diagnostic plots of forecast time series at selected grid cells, it is observed that at longer lead times, forecasts generated by data-driven models tend to be smoother and less skilful compared to those generated by physical models. Overall, the extension

20 of the WeatherBench 2 facilitates more comprehensive comparisons of hydroclimatic forecasts and provides useful information for forecast applications.

Keywords: Binary forecast; forecast verification; warm extreme; wet extreme; forecast skill; scorecard.

25   **1 Introduction**

Accurate numerical weather prediction (NWP) is of great importance to the economy and society (Bi et al., 2023; Lam et al., 2023; Bauer et al., 2015). Conventionally, physical NWP models formulate the governing equations of coupled physical processes in land, ocean and atmosphere and therefore predict weather conditions in the near future based on predetermined initial meteorological fields (Lam et al., 2023; Bauer et al., 2015). Due to advances in remote sensing, data assimilation and

30   computational infrastructure, physical NWP models have witnessed steady improvements and been extensively employed in operational forecasting (Bauer et al., 2020). For example, the European Centre for Medium-range Weather Forecast (ECMWF) operates the Integrated Forecast System (IFS) that has implemented a significant resolution upgrade and methodology for high-resolution forecasts (HRES) and ensemble forecasts (ENS) at the horizontal resolution of 0.1 degrees since January 2016 (Balsamo et al., 2023).

35   Data-driven NWP models have recently gained increasing popularity in hydroclimatic forecasting (Ben Bouallègue et al., 2024; Rasp et al., 2024; de Burgh-Day and Leeuwenburg, 2023; Xu et al., 2024a). Early models, such as the UNet architecture-based cubed sphere projection (Weyn et al., 2020) and deep Resnet architecture-based models (Clare et al., 2021; Rasp and Thuerey, 2021), were of moderate spatial-temporal resolution and forecast skill. Recent deep learning models, such as graph neural network (Keisler, 2022) and FourCastNet (Pathak et al., 2022), began to match operational NWP models in resolution

40   and skills. Pangu-Weather (Bi et al., 2023) and GraphCast (Lam et al., 2023) even outperformed the HRES in terms of some deterministic metrics. The Neural General Circulation Models (NeuralGCM) that integrates data-drive and physical modules is considered to be the first hybrid model obtaining competitive or better scores than the HERS (Kochkov et al., 2024). The GenCast generates global ensemble forecasts that are comparative or even more skilful than the ENS (Price et al., 2025).

There is a growing demand to verify the capability of physical and data-driven models in generating skilful hydroclimatic

45   forecasts (Olivetti and Messori, 2024a; Zhong et al., 2024; Ben Bouallègue et al., 2024). In response to the need of a unified benchmark, the WeatherBench has been established to host a common dataset of forecasts and observations and utilizes popular evaluation metrics for forecast comparisons (Rasp et al., 2020). Owing to rapid advances in data-driven NWP models, the WeatherBench 2 has been developed to support global medium-range forecast verification (Rasp et al., 2024). By following established practices in the World Meteorological Organisation (WMO), the WeatherBench 2 pays attention to both

50   deterministic and ensemble forecasts generated by physical and data-driven NWP models (Jin et al., 2024). Forecast verification is performed by an open-source Python code and publicly available, cloud-optimized ground-truth and baseline datasets (Jin et al., 2024; Olivetti and Messori, 2024b; Rasp et al., 2024).

Besides deterministic and ensemble forecasts, there is a demand of binary forecasts in disaster prevention and risk management (Ben Bouallègue et al., 2024; Larraondo et al., 2020). The importance of binary forecasts arises from the

55   operational need for predicting the occurrence versus non-occurrence of hydroclimatic extremes (Jolliffe and Stephenson,

2012; Larraondo et al., 2020; Rasp et al., 2020). Therefore, this paper aims to extend the WeatherBench 2 to binary forecasts.

The objectives are: 1) to account for verification metrics on binary forecasts derived from global precipitation and temperature

forecasts; 2) to present scorecards to showcase the predictive performance on wet and warm extremes; and 3) to examine the

sensitivity of different metrics to predefined thresholds of hydroclimatic extremes. As will be shown in the methods and results,

60   the extension facilitates an effective intercomparison among binary forecasts of hydroclimatic extremes generated by data-

driven and physical models.

2 Forecasts and metrics in the WeatherBench 2

2.1 Forecast datasets

65   The WeatherBench 2 presents a benchmark for verifying and comparing the performance of data-driven and physical

NWP models (Rasp et al., 2024). On its website (https://weatherbench2.readthedocs.io), there is a database containing past

forecasts in the year 2020:

1) The HRES generated by the ECMWF's IFS is widely regarded as one of the best global deterministic weather forecasts

(Rasp et al., 2024). It offers 10-day forecasts at the horizontal resolution of 0.1 degrees with 137 vertical levels (Balsamo et

70   al., 2023). In the WeatherBench2, the HRES is used as the main baseline for comparing the performance of data-driven models.

2) The ENS generated by the IFS's ensemble version is widely known as one of the best global ensemble weather forecasts.

It consists of 1 control member and 50 perturbed members (Balsamo et al., 2023). In the WeatherBench2, the ENS is also used

as a baseline. The average over the 50 members (ENS Mean) is considered as an additional baseline (Rasp et al., 2024).

3) The ERA5 forecasts are hindcasts generated by the exact IFS that is employed to create the renowned ERA5 of

75   historical global climate conditions (Hersbach et al., 2020). They consist of 10-day hindcasts at the horizontal resolution of

0.25 degrees.

4) The 10-day global forecast generated by graph neural network includes 6 upper-air variables at 13 pressure levels at

the horizontal resolution of 1 degree (Keisler, 2022). This network uses an encoder to map the latitude-longitude grid to an

icosahedron grid and a decoder to map back to the original grid after several rounds of computations. It runs autoregressively

80   to forecast atmospheric states for the next 6-hour based on states from the previous two 6-hour time steps.

5) The two sets of 10-day global forecasts generated by the Pangu-Weather consist of 5 upper-air variables at 13 vertical

levels and 4 surface variables at the horizontal resolution of 0.25 degrees (Bi et al., 2023). The Pangu-Weather is based on the

vision transformer architecture and hierarchical temporal aggregation. Four time steps, i.e., 1, 3, 6 and 24 hours, are chained

autoregressively to generate forecast at any lead time based on the current atmospheric states.

85   6) The two sets of 10-day forecasts generated by the GraphCast includes 6 upper-air variables at 37 vertical levels and 5 surface variables at the horizontal resolution of 0.25 degrees (Lam et al., 2023). The GraphCast is based on the architecture of graph neural network. It runs autoregressively to forecast atmospheric states for the next time step based on states from the previous two time steps at the temporal resolution of 6 hours.

     7) The 10-day global forecasts generated by the spherical convolutional neural networks (CNNs) are composed of 6

90 upper-air variables at 13 vertical levels (Esteves et al., 2023). The Spherical CNNs extend the CNNs to the sphere domains through leveraging the spherical convolutions as the primary linear operation. It produces forecasts at the longitude resolution of 1.4 degrees and the latitude resolution of 0.7 degrees.

     8) The 15-day global forecasts generated by the FuXi consists of 5 upper-air variables at 13 vertical levels and 5 surface variables at the horizontal resolution of 0.25 degrees (Chen et al., 2023). The FuXi is an autoregressively cascading model

95 based on the U-Transformer architecture. It comprises three sub-models fine-tuned for forecasting 0-5, 5-10 and 10-15 days ahead at the temporal resolution of 6 hours. The atmospheric states for the next time step are forecasted based on the states from the previous two time steps.

     9) The 15-day deterministic and ensemble forecasts generated by NeuralGCM are composed of 7 upper-air variables at 37 vertical levels at the temporal resolution of 12 hours (Kochkov et al., 2024). The NeuralGCM integrates the differential dynamical core and the learned physics module. The deterministic version is trained in the horizontal resolution of 0.7 degrees.

100 The ensemble version is trained for the horizontal resolution of 1.4 degrees and is run to produce 50 members.

2.2 Verification metrics

     The WeatherBench 2 takes into consideration in total 6 metrics for deterministic forecasts and 6 metrics for ensemble

105 forecasts, as shown in Table 1. In forecast verification, the ERA5 data is used as the ground truth for verifying the data-driven models. For the sake of fair comparison with the data-driven models, the initial conditions of the IFS HRES is used as the ground truth for verifying IFS forecasts (Lam et al., 2023). As precipitation is not available for IFS HRES's initial conditions, the total precipitation of ERA5 data is used as the ground truth data for all models. In the initial version of the WeatherBence2, the verification is conducted for forecasts initialized at 00 and 12 UTC for the period from 1 January 2020 to 31 December

110 2020. All forecasts, baseline data and ground truth data are resampled to the horizontal resolution of 1.5 degrees that is used as the standard resolution for forecasts verification by the WMO and ECWMF (Rasp et al., 2024).

Table 1 Metrics for deterministic and ensemble forecasts in the WeatherBench 2.

Forecast	Metric	[min, max]	Optimal value
Deterministic	Root mean square error (RMSE)	$[0, +\infty)$	0
	Mean square error	$[0, +\infty)$	0
	Mean absolute error	$[0, +\infty)$	0
	Bias	$(-\infty, +\infty)$	0
	Anomaly correlation coefficient	$[-1, 1]$	1
	Stable Equitable Error in Probability Space (SEEPS)	$[0, 1]$	0
Ensemble	Continuous ranked probability score (CRPS)	$[0, +\infty)$	0
	Ranked probability score (RPS)	$[0, +\infty)$	0
	Spread-Skill Ratio	$[0, 1]$	1
	Energy score	$[0, +\infty)$	0
	Brier score (BS)	$[0, 1]$	0
	Ignorance score	$[0, +\infty)$	0

3 Verification of binary hydroclimatic forecasts

3.1 Conversion to binary forecasts

Binary forecasts on the occurrence versus non-occurrence of target events can be generated from deterministic and ensemble forecasts by using predefined thresholds of hydroclimatic events (Ben Bouallègue et al., 2024). As to precipitation, the 90th percentile of the 24-hour accumulation of total precipitation (TP24h) is considered as the threshold, above which the TP24h is considered as the wet extreme (North et al., 2013). As to temperature, the 90th percentile of the 24-hour maximum of 2m temperature (T2M24h) is set as the threshold, above which the T2M24h is categorized as the warm extreme (Olivetti and Messori, 2024b). Given the pre-defined threshold (q), deterministic forecasts are converted into either 0 or 1:

$$I(f_n > q) = \begin{cases} 1, & f_n > q \\ 0, & otherwise \end{cases} \tag{1}$$

where f_n represents the n-th deterministic forecast. By contrast, ensemble forecasts are converted into forecast probabilities by using the Weibull's plotting position (Makkonen, 2006):

$$p_{f_n} = \frac{\sum_{m=1}^{M} I(f_{n,m} > q)}{M + 1} \tag{2}$$

where $f_{n,m}$ is the m-th member of the n-th ensemble forecasts and M is the number of ensemble members.

In relation to the corresponding observations, binary forecasts can be divided into four categories, i.e., hits (a), false alarms (b), misses (c) and correct rejections (d), as shown in Table 2 (Larraondo et al., 2020). The hits represent events that are successfully forecasted; the false alarms are non-events that are incorrectly forecasted as events; the misses denote events that are incorrectly forecasted as non-events; and the correct rejections represent non-events that are correctly forecasted as non-events. The proportion of the observed events to the total number of events and non-events is the base rate, with lower values often corresponding to events that are more extreme (Ferro and Stephenson, 2011).

Table 2. Contingency table for binary forecasts.

	Observed events	Observed non-events	Total
Forecasted events	$a = \begin{cases} \sum_{n=1}^{N} I(f_n > q \mid p_{o_n} = 1), \; if \; M = 1 \\ \sum_{n=1}^{N} I(p_{f_n} > c \mid p_{o_n} = 1), if \; M \geq 2 \end{cases}$	$b = \begin{cases} \sum_{n=1}^{N} I(f_n > q \mid p_{o_n} = 0), \; if \; M = 1 \\ \sum_{n=1}^{N} I(p_{f_n} > c \mid p_{o_n} = 0), if \; M \geq 2 \end{cases}$	$a + b$
Forecasted non-events	$c = \begin{cases} \sum_{n=1}^{N} I(f_n \leq q \mid p_{o_n} = 1), \; if \; M = 1 \\ \sum_{n=1}^{N} I(p_{f_n} \leq c \mid p_{o_n} = 1), if \; M \geq 2 \end{cases}$	$d = \begin{cases} \sum_{n=1}^{N} I(f_n \leq q \mid p_{o_n} = 0), \; if \; M = 1 \\ \sum_{n=1}^{N} I(p_{f_n} \leq c \mid p_{o_n} = 0), if \; M \geq 2 \end{cases}$	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = N$

Where $M = 1$ and $M \geq 2$ respectively denote the deterministic forecasts and ensemble forecasts; N is the number of pairs of observations

135 *and forecasts for verification; p_{o_n} represents the binary observation with a value of 1 for the event and 0 for non-event; c denotes the*

probability thresholds above which the events are forecasted to occur.

3.2 Verification metrics for binary forecasts

Given the challenges posed by varying hydroclimatic extremes and imbalanced samples, in total 16 metrics are utilized

140 to examine the performance of binary forecasts (Jolliffe and Stephenson, 2012; North et al., 2013). There are 7 base-rate-

dependent metrics and 9 base-rate-independent metrics. The base-rate-dependent metrics provide insights into the performance

in relation to varying frequency of extreme events (Jolliffe and Stephenson, 2012). On the other hand, the base-rate-

independent metrics are suitable for comparing forecasts across different climate regions or time periods, in which the

frequency of extreme events differs substantially (Ferro and Stephenson, 2011; Jacox et al., 2022). Their equations, ranges and

145 optimal values are presented in Table 3.

Table 3. Metrics for binary forecasts.

Metric	Equation	[min, max]	Optimal value	Reference
Base-rate-dependent metrics				
Accuracy (ACC), proportion correct	$ACC = \dfrac{a + d}{N}$	[0, 1]	1	(Finley, 1884)
Success ratio (SR), precision	$SR = \dfrac{a}{a + b}$	[0, 1]	1	(Lagadec et al., 2016)
Critical success index (CSI), threat score, Gilbert score	$CSI = \dfrac{a}{a + b + c}$	[0, 1]	1	(Donaldson et al., 1975; Gilbert, 1884)
Heidke skill score (HSS), Cohen's Kappa	$HSS = \dfrac{a + d - a_r - d_r}{N - a_r - d_r}, d_r = \dfrac{(b + d)(c + d)}{N}$	[−1, 1]	1	(Gomis-Cebolla et al., 2023; Heidke, 1926)

Gilbert skill score (GSS), equitable threat score	$GSS = \dfrac{a - a_r}{a + b + c - a_r}, a_r = \dfrac{(a+b)(a+c)}{N}$	[−1/3, 1]	1	(Gilbert, 1884; Schaefer, 1990)
Extreme dependence score (EDS)	$EDS = \dfrac{\ln p - \ln H}{\ln p + \ln H}, p = \dfrac{a+c}{N}$	[−1, 1]	1	(Primo and Ghelli, 2009; Stephenson et al., 2008)
Symmetric extreme dependence score (SEDS)	$SEDS = \dfrac{\ln q - \ln H}{\ln p + \ln H}, q = \dfrac{a+b}{N}$	[−1, 1]	1	(Orozco López et al., 2010)
Base-rate-independent metrics				
Hit rate (H), sensitivity, recall, probability of detection	$H = \dfrac{a}{a+c}$	[0, 1]	1	(Swets, 1986)
False alarm rate (F), probability of false detection	$F = \dfrac{b}{b+d}$	[0, 1]	0	(Donaldson et al., 1975)
Specificity, true negative rate (TNR)	$TNR = \dfrac{d}{b+d}$	[0, 1]	1	(Agrawal et al., 2023)
Odds ratio skill score (ORSS), Yule's Q	$ORSS = \dfrac{ad - bc}{ad + bc}$	[−1, 1]	1	(Stephenson, 2000)
Peirce's skill score (PSS), Hanssen and Kuipers discriminant	$PSS = \dfrac{ad - bc}{(a+c)(b+d)} = H - F$	[−1, 1]	1	(Peirce, 1884)
Extremal dependence index (EDI)	$EDI = \dfrac{\ln F - \ln H}{\ln F + \ln H}$	[−1, 1]	1	(Ferro and Stephenson, 2011)
Symmetric extremal dependence index (SEDI)	$SEDI = \dfrac{\ln F - \ln H + \ln(1-H) - \ln(1-F)}{\ln F + \ln H + \ln(1-H) + \ln(1-F)}$	[−1, 1]	1	(Ferro and Stephenson, 2011)
Area under receiver operating characteristic (ROC) curve (AUC)	$AUC = \phi\left(\dfrac{\phi^{-1}(H) - \phi^{-1}(F)}{\sqrt{2}}\right)$	[0, 1]	1	(Swets, 1986)
ROC skill score (ROCSS)	$ROCSS = 2(AUC - 0.5)$	[−1, 1]	1	(Swets and Swets, 1986)

The 7 base-rate-dependent metrics in Table 3 are influenced by the underlying distribution of observed events and non-events (Jolliffe and Stephenson, 2012). The accuracy is calculated as the ratio between the number of hits and the total number of events and non-events (Finley, 1884). The success ratio (SR) measures the number of hits divided by the number of forecasted events (Lagadec et al., 2016). The critical success index (CSI) is the number of hits divided by the total number of forecasted and observed events (Chakraborty et al., 2023; Gilbert, 1884; Donaldson et al., 1975). The Heidke skill score (HSS) measures the accuracy relative to that of the random forecasts (Gomis-Cebolla et al., 2023). The Gillert skill score (GSS) evaluates the correctly predicted fraction of the observed and forecasted events after adjusting for the random hits (Chen et al., 2018; Coelho et al., 2022). Converging to a meaningful limit, the extreme dependency score (Stephenson et al., 2008) and the symmetric extreme dependency score (SEDS) (Orozco López et al., 2010) are suitable for binary forecasts of rare events.

The 9 base-rate-independent metrics in Table 3 are valuable for rare events due to their stability to the variation in the proportion of observed events (Ferro and Stephenson, 2011). The hit rate and false alarm rate respectively quantify the proportion of hits in observed events and the proportion of false alarms in observed non-events (Swets, 1986). The specificity

measures the percentage of correct rejections to observed non-events (Agrawal et al., 2023). The odds ratio skill score (ORSS) examines the improvement over the random forecasts (Stephenson, 2000). The Peirce's skill score (PSS) has similar formulation to HSS but does not depend on event frequency (Chakraborty et al., 2023). The extremal dependence index (EDI) and the symmetric extremal dependence index (SEDI) are designed to be nondegenerate (Ferro and Stephenson, 2011). The

165   receiver operating characteristic (ROC) examines the discrimination between hits and false alarms, quantified by the area under the ROC curve (Swets, 1986). A higher ROC skill score (ROCSS) indicates better predictive skill. For probabilistic forecasts, the ROCSS can be calculated by considering the hit rate and false alarm rate for all possible thresholds of probability (Huang and Zhao, 2022).

170   **3.3 Forecast verification**

Considering data availability and forecast setting, the attention is paid to 8 sets of forecasts, i.e., IFS HRES, IFS ENS, IFS ENS Mean, Pangu-Weather (operational), GraphCast (operational), Pangu-Weather, GraphCast and FuXi (Rasp et al., 2024). The ground truth, spatial resolution, initial forecast time and verification period are selected by following the WeatherBench 2. A set of predefined thresholds ranging from the 80th to 99th percentiles of the ground truth data in 2020 are

175   considered for sensitivity analysis (Olivetti and Messori, 2024b; North et al., 2013). To facilitate comparisons at regional to global scales, the 16 metrics are calculated based on the area-weighting method over grid cells (Rasp et al., 2024). The regions are defined by the ECMWF's scorecards, as illustrated in Table 4.

Table 4. Regions that are included in the ECMWF's scorecards.

Region	Range	Region	Range
Northern hemisphere (extra-tropics)	latitude ≥ 20°	Europe	35° ≤ latitude ≤ 75°, -12.5° ≤ longitude ≤ 42.5°
Southern hemisphere (extra-tropics)	latitude ≤ -20°	North America	25° ≤ latitude ≤ 60°, -120° ≤ longitude ≤ -75°
Tropics	-20° ≤ latitude ≤ 20°	North Atlantic	25° ≤ latitude ≤ 60°, -70° ≤ longitude ≤ -20°
Extra-tropics	|latitude| ≥ 20°	North Pacific	25° ≤ latitude ≤ 60°, 145° ≤ longitude ≤ -130°
Arctic	latitude ≥ 60°	East Asia	25° ≤ latitude ≤ 60°, 102.5° ≤ longitude ≤ 150°
Antarctic	latitude ≤ -60°	AusNZ	-45° ≤ latitude ≤ -12.5°, 120° ≤ longitude ≤ 175°

180   *AusNZ: Australia and New Zealand.*

Given the spatial and temporal clustering of hydroclimatic extremes, the two-sided paired t test with cluster-robust standard errors is conducted at the significance level of 0.05 to assess the differences in the performance between data-driven

models and IFS HRES (Olivetti and Messori, 2024b; Liang and Zeger, 1986; Shen et al., 1987). For comparison at the grid

185 scale, the 16 metrics are computed separately for each grid cell. The same paired t test is performed with p value that is

corrected for multiple testing using global false-discovery rates at the significance level of 0.1 (Ouyang et al., 1995; Olivetti

and Messori, 2024b). It approximately corresponds to the significance level of 0.05 for spatially correlated hydroclimatic

extremes (Wilks, 2016).

4 Results

190

4.1 Predictive performance across the globe

Scorecards of the globally area-weighted ROCSS relative to the IFS HRES baseline are shown in Figure 1. As expected,

forecasts become less accurate as lead time increases from 1 day to 10 days. For wet extremes, the IFS ENS, IFS ENS Mean,

GraphCast (operational) and GraphCast tend to outperform the IFS HRES. At the lead times of 3 and 10 days, the ROCSS is

195 respectively 0.59 and 0.16 for the IFS HRES, 0.90 and 0.55 for the IFS ENS, 0.61 and 0.17 for the IFS ENS Mean, 0.65 and

0.20 for the GraphCast (operational), 0.61 and 0.16 for the GraphCast and 0.54 and 0.08 for the FuXi. For warm extremes, the

GraphCast and FuXi tend to be more skilful than the IFS HRES within 3-day lead time. As lead time increases, data-driven

forecasts can be less skilful than the IFS HRES. It is highlighted that the IFS ENS is remarkably more skilful than the IFS

HRES at the lead time from 1 to 10 days. At the lead times of 3 and 10 days, the ROCSS is respectively 0.68 and 0.42 for the

200 IFS HRES, 0.92 and 0.86 for the IFS ENS, 0.62 and 0.32 for the IFS ENS Mean, 0.63 and 0.29 for the Pangu-Weather, 0.68

and 0.39 for the GraphCast and 0.68 and 0.32 for the FuXi.

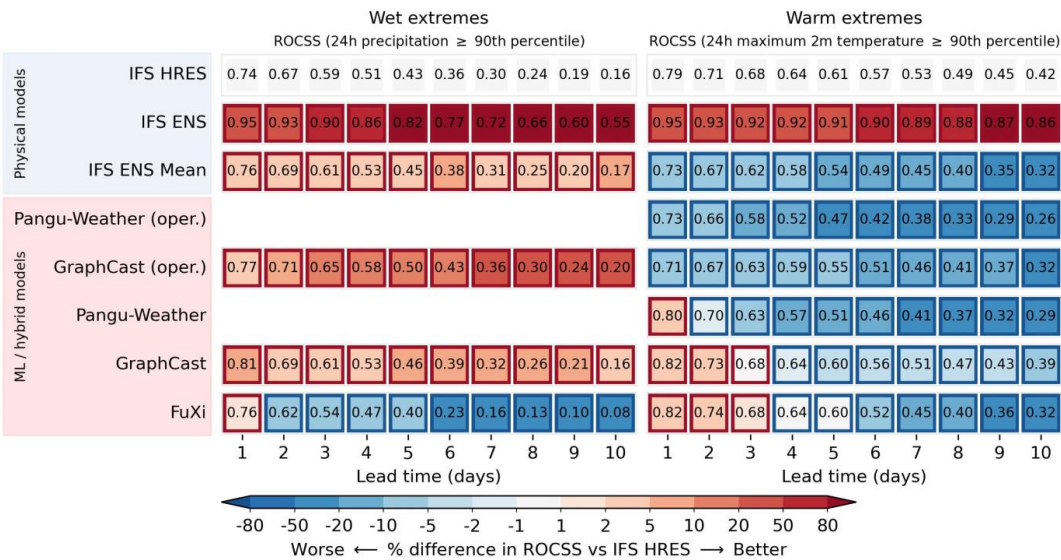



Figure 1. Globally area-weighted ROCSS for wet and warm extremes. The oper. denotes operational version. The red and blue borders indicate significantly different performances compared to the IFS HRES at the significance level of 0.05.

Scorecards of the area-weighted ROCSS for wet extremes relative to the IFS HRES baseline are illustrated by region in Figure 2. Overall, the IFS ENS stands out across different regions and lead times. The GraphCast (operational) tends to outperform the IFS HRES. The GraphCast tend to be better than the IFS HRES in Southern Hemisphere (extra-tropics), Arctic, Antarctic, Europe, North Pacific, East Asia and AusNZ. In Europe, at the lead times of 3 and 10 days, the ROCSS is respectively 0.73 and 0.19 for the IFS HRES, 0.96 and 0.64 for the IFS ENS, 0.76 and 0.23 for the GraphCast (operational), 0.77 and 0.22 for the GraphCast and 0.69 and 0.11 for the FuXi. In the meantime, the FuXi tend to outperform the IFS HRES in the Southern Hemisphere (extra-tropics), tropics, North Atlantic and AusNZ at lead time less than 3 days. Except for the Arctic and Antarctic, the IFS ENS Mean tends to be better than the IFS HRES. The GraphCast (operational) is comparable to the IFS ENS Mean and marginally better in the polar regions. In the Antarctic region, the ROCSS is 0.63 and 0.06 for the IFS HRES, 0.59 and 0.01 for the IFS ENS Mean and 0.66 and 0.06 for the GraphCast (operational) at lead time of 3 and 10 days.

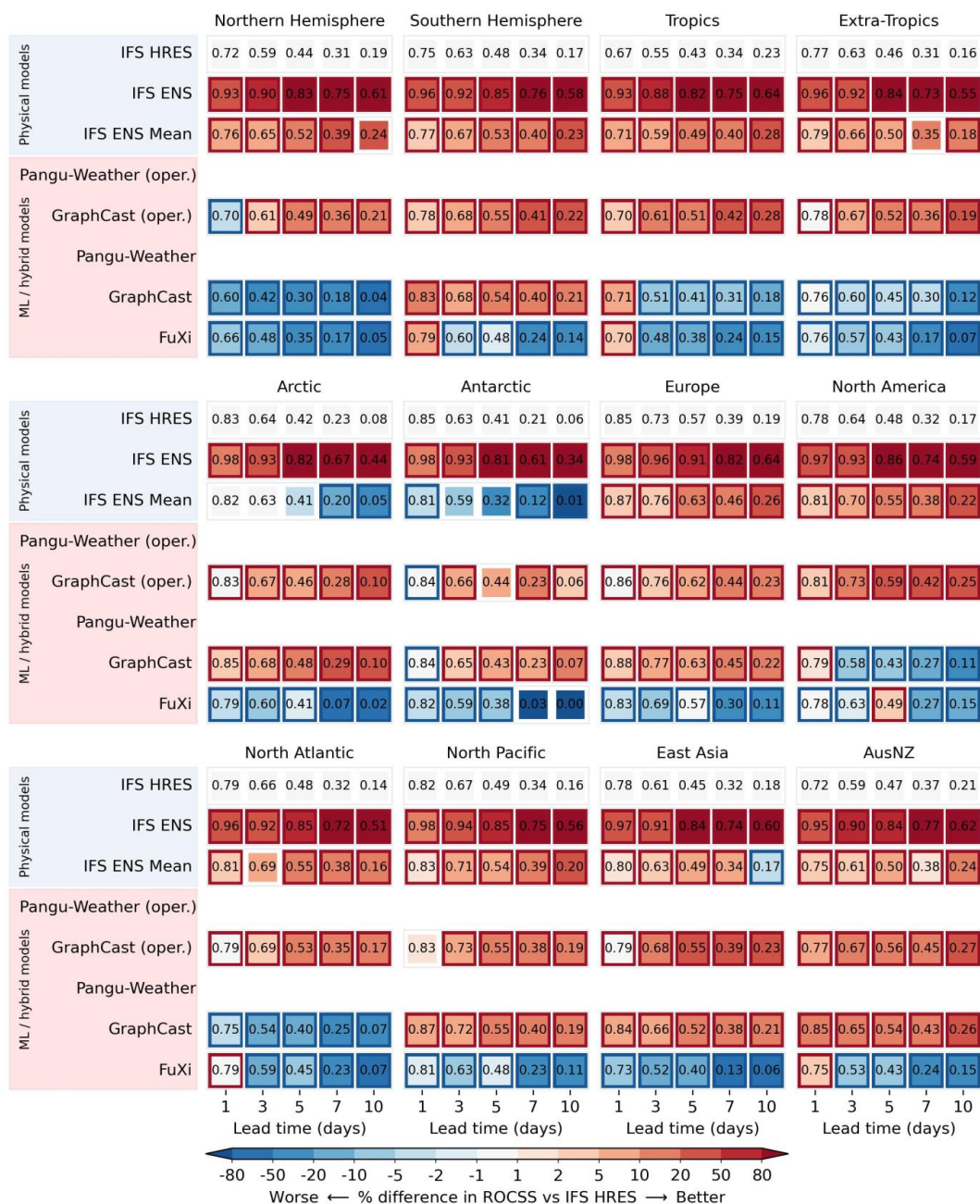



Figure 2. Regionally area-weighted ROCSS of different forecasts for wet extreme. The red and blue borders indicate significantly different performance compared to the IFS HRES at the significance level of 0.05.

Scorecards of the regionally area-weighted ROCSS for warm extremes relative to the IFS HRES baseline are showcased

in Figure 3. The Pangu-Weather, GraphCast and FuXi tend to outperform the IFS HRES within 3-day lead time except for the

Arctic and Antarctic. These results are consistent with the results of a previous study on forecast accuracy of warm extremes

225 (Olivetti and Messori, 2024b). In the North America, North Atlantic, North Pacific, East Asia and AusNZ, the GraphCast and

FuXi tend to outperform the IFS HRES at longer lead time even up to 10 days. The ROCSS in the North Atlantic is respectively

0.39, 0.58 and 0.49 for the IFS HRES, GraphCast and FuXi at the 10-day lead time. On the other hand, the performances of

all data-driven forecasts tend to be worse than that of the IFS HRES in the Arctic and Antarctic. In Europe, the ROCSS is

respectively 0.78, 0.71, 0.76 and 0.75 for the IFS HRES, Pangu-Weather, GraphCast and FuXi at 5-day lead time. As averaging

230 the ensemble members can filter unpredictable features to get smoother forecasts, it is not surprising that the IFS ENS Mean

does not always perform as well as the IFS HRES and IFS ENS for warm extremes (Ben Bouallègue et al., 2024).

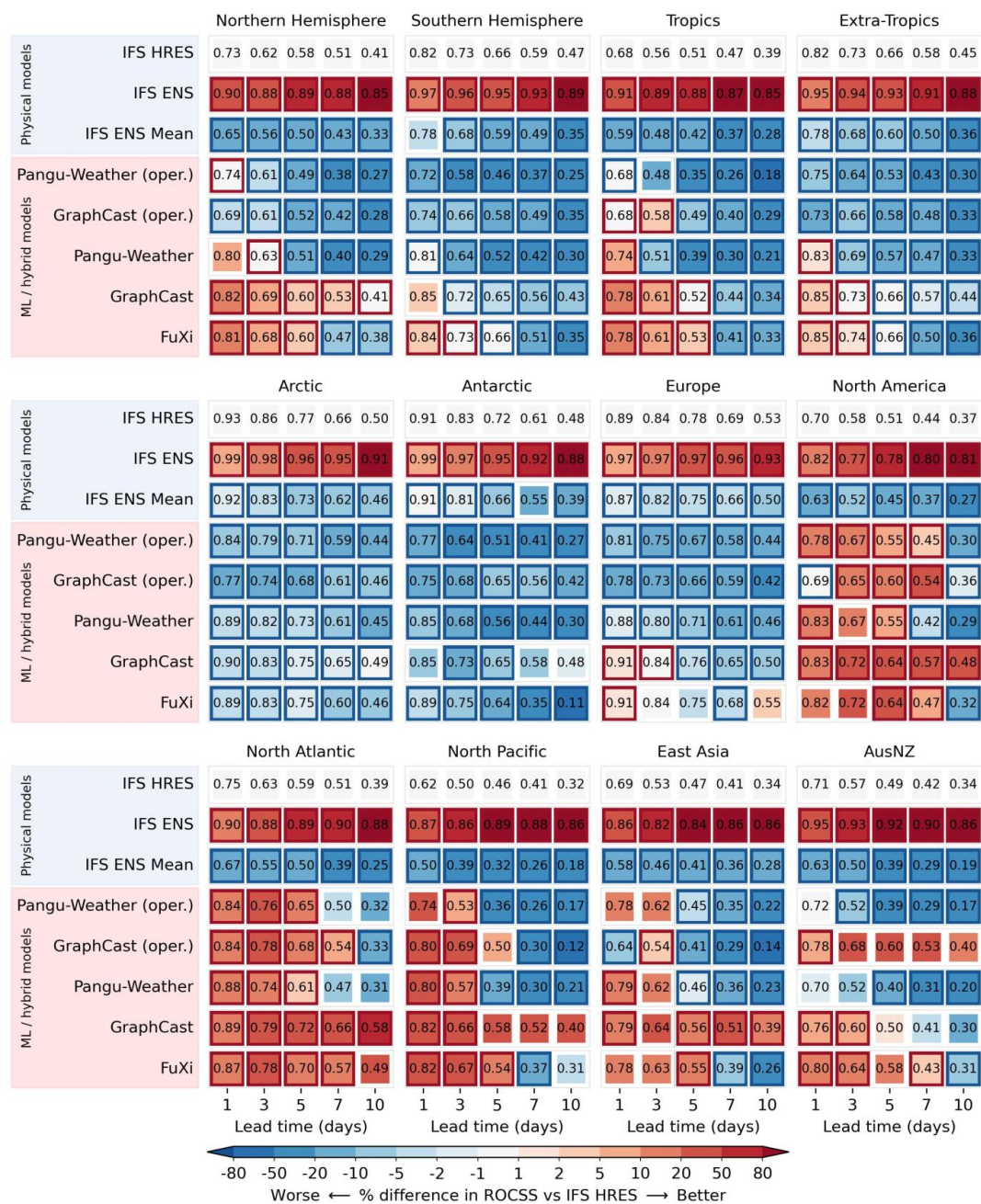



Figure 3. As for Figure 2, but for warm extremes.

235

4.2 Predictive performance of wet extremes

The differences in the ROCSS for wet extremes in comparison with the IFS HRES baseline are illustrated in Figure 4.

Overall, the IFS ENS tends to outperform the IFS HRES at most grid cells across the globe. Except for the Northern Africa

and Arabian Peninsula, the GraphCast's operational forecasts are comparable or more skilful than the IFS HRES. The

240 GraphCast is not as skilful as the IFS HRES in more grid cells, such as Central Australia and Central Asia. The FuXi tends to

be less skilful than the IFS HRES in most grid cells, such as the Atlantic and Pacific. As the lead time increases, the IFS ENS

and GraphCast (operational) are observed to outperform the IFS HRES, while the GraphCast and FuXi underperform. These

results are consistent with the results of Figure 1 and Figure 2. In Northern Africa, forecasts of the three data-driven models

tend to be less skilful than the IFS HRES and IFS ENS. As the GraphCast and FuXi exhibit no hits or false alarms for many

245 of or even almost all the grid cells in this region, the ROCSS is nearly zero so that their forecasts tend to be worse than the IFS

HRES in the Northern Hemisphere (extra-tropics) and Tropics.

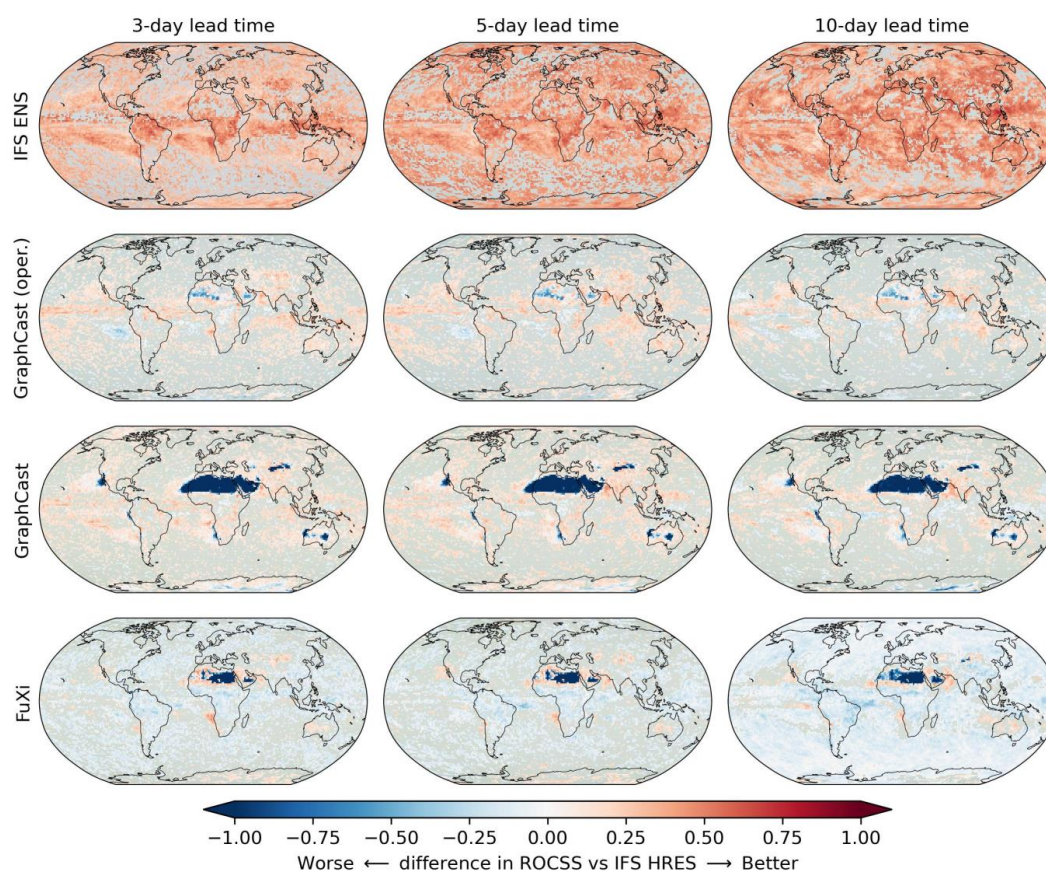



Figure 4. Differences of IFS ENS, GraphCast (operational), GraphCast and FuXi in ROCSS to the IFS HRES for wet extremes at
250 **each grid cell. The grey colour indicates grid with no statistically significant differences at the significance level of 0.1.**

The time series for 24-hour accumulation of total precipitation from different forecasts initialized at 00 UTC are shown

for three selected grid cells in Figure 5. Overall, data-driven models can capture the temporal dynamics of precipitation but their forecasts are smoother than the IFS HRES (Zhong et al., 2024; Xu et al., 2024b). As lead time increases, the skill of IFS

255   ENS and IFS ENS Mean tend to decrease more slowly than that of IFS HRES, while the skill of GraphCast and FuXi reduces more rapidly. For the three grid cells, the five sets of forecasts have close number of correct negatives; the IFS HRES and GraphCast show more hits; the IFS HRES are more capable of capturing the wet extremes but tends to produce more false alarms; the IFS ENS Mean and FuXi tend to underestimate the wet extremes, resulting in more misses and fewer false alarms. For grid cell A, at the lead times of 3 and 10 days, the ROCSS is respectively 0.63 and 0.34 for the IFS HRES, 0.92 and 0.76

260   for the IFS ENS, 0.62 and 0.43 for the IFS ENS Mean, 0.75 and 0.38 for the GraphCast and 0.48 and 0.16 for the FuXi. For grid cells B and C, the numbers of hits and false alarms of FuXi are zero at 10-day lead time, leading to zero values of ROCSS.

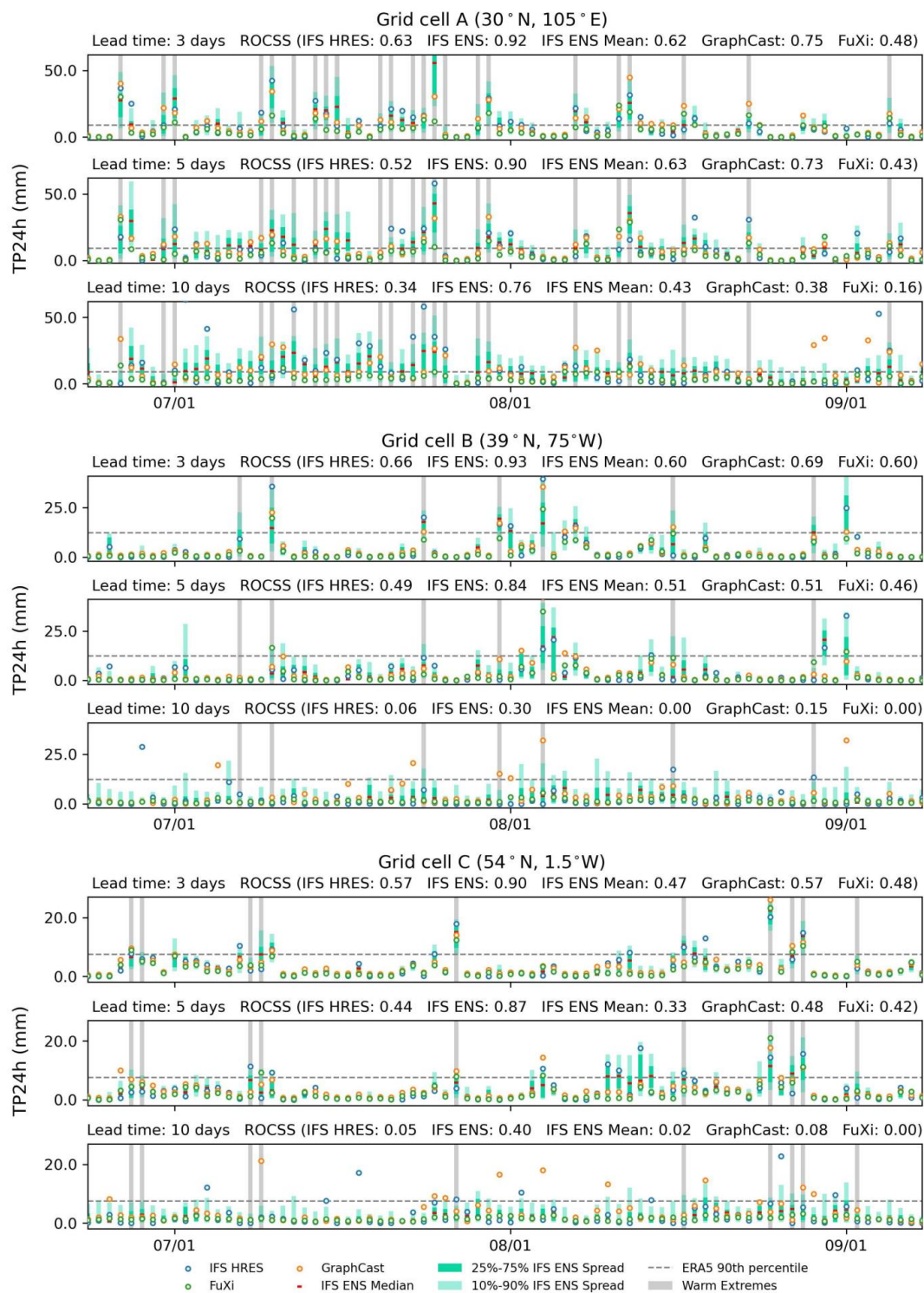



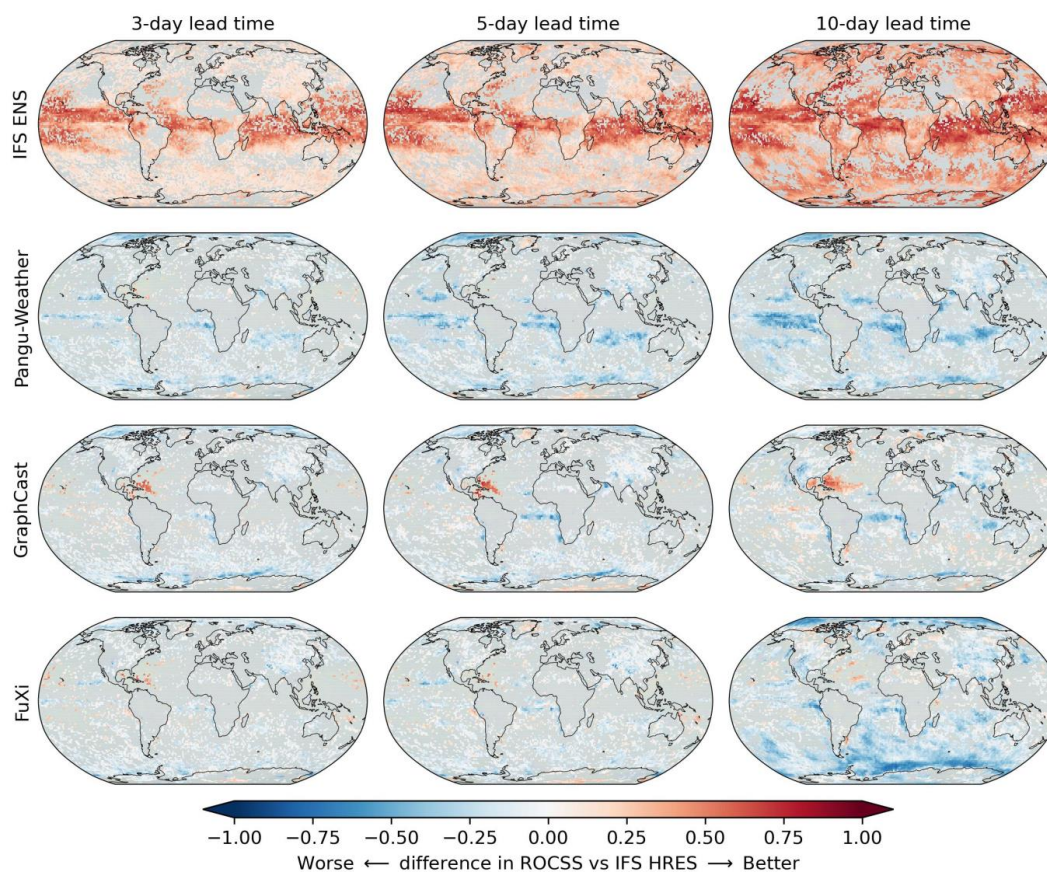
Figure 5. Time series plots of TP24h forecasts initialized at 00 UTC for the IFS HRES, IFS ENS, IFS ENS Mean, GraphCast and
FuXi over three selected grid cells, i.e., A (30 ˚N, 105 ˚E), B (39 ˚N, 75 ˚W) and C (54 ˚N, 1.5 ˚W).

265

4.3 Predictive performance of warm extremes

The differences in ROCSS for warm extremes in comparison with the IFS HRES baseline are illustrated in Figure 6. The IFS ENS tends to outperform the IFS HRES, especially in low-latitude regions. As the lead time increases, the IFS ENS tends

270   to be more skilful than the IFS HRES. The ROCSS of the Pangu-Weather, GraphCast and FuXi is similar to that of the IFS HRES but is lower in most grids of the Pacific, Atlantic and Arctic. The GraphCast tends to outperform the IFS HRES in the Northern Atlantic near the Gulf of Mexico. The spatial patterns of the differences in ROCSS are consistent with the results of Figure 3. As the lead time increases to 10 days, the area where the Pangu-Weather, GraphCast and FuXi are more skilful than the IFS HRES decreases. On the other hand, even for lead time of 10 days, the GraphCast and FuXi continue to outperform

275   the IFS HRES in some regions of the North Atlantic. The different performances of global weather forecasts in different regions emphasize the necessity to verify and calibrate hydroclimatic forecasts before operational application (Ben Bouallègue et al., 2024; Huang et al., 2022).

280 **Figure 6. Differences of IFS ENS, Pangu-Weather, GraphCast and FuXi in ROCSS to the IFS HRES for warm extremes at each grid cell. The grey colour indicates grid with no statistically significant differences at the significance level of 0.1.**

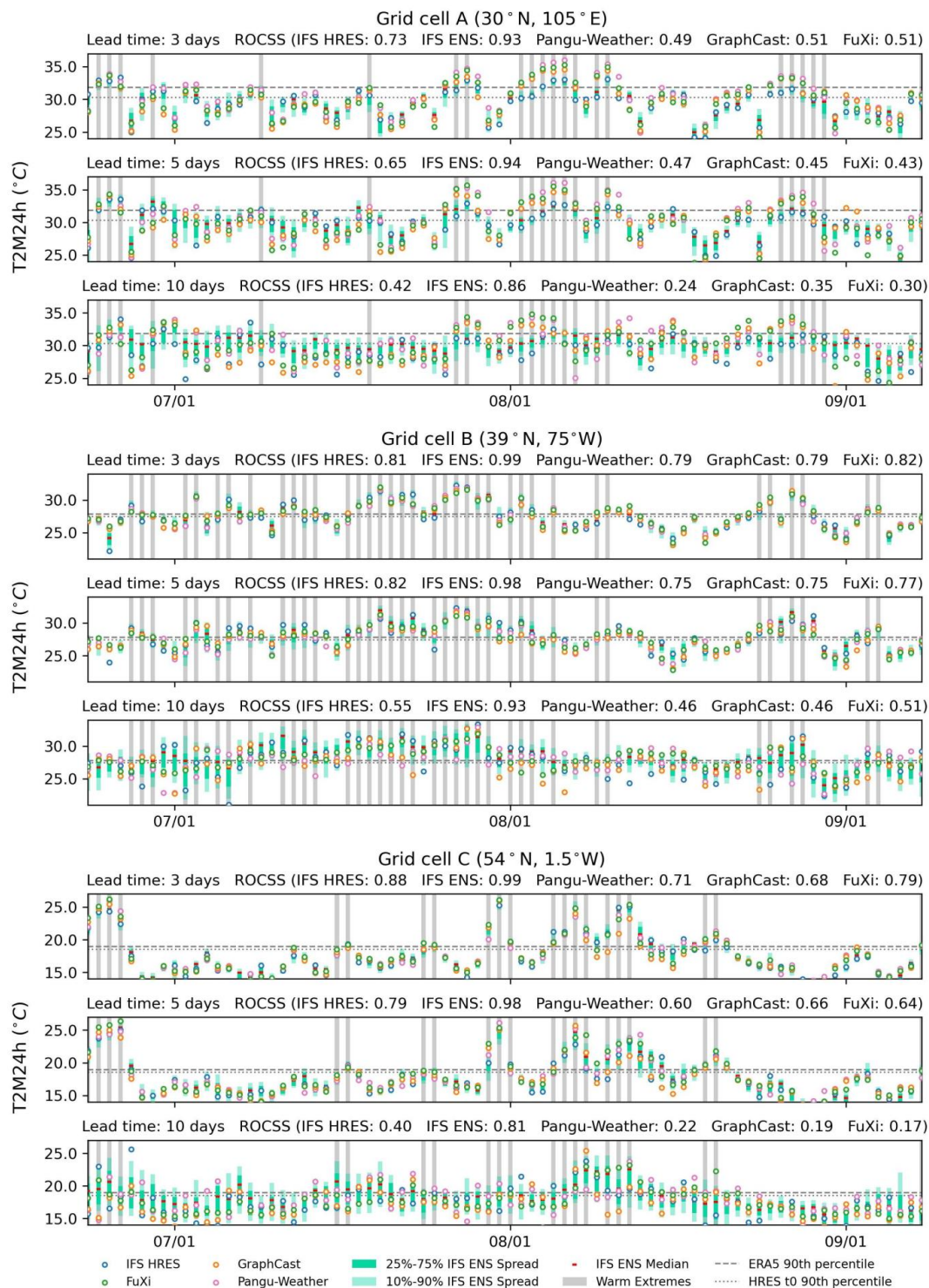
The time series for 24-hour maximum of 2m temperature from different forecasts initialized at 00 UTC are shown for three selected grid cells in Figure 7. Overall, the Pangu-Weather, GraphCast and FuXi exhibit similar temperature dynamics over time to those of the IFS HRES. As lead time increases, these data-driven models tend to produce smoother forecasts with the skill reducing more rapidly than the IFS HRES and IFS ENS (Zhong et al., 2024; Rasp et al., 2024). For grid cells A, B and C, the IFS HRES and IFS ENS tend to outperform the Pangu-Weather, GraphCast and FuXi. At the lead time of 3 and 10 days, the ROCSS for grid cell A is respectively 0.73 and 0.42 for the IFS HRES, 0.93 and 0.86 for the IFS ENS, 0.49 and 0.24 for the Pangu-Weather, 0.51 and 0.35 for the GraphCast and 0.51 and 0.30 for the FuXi. There are more misses for the three data-driven models, indicating that they tend to underestimate the warm extremes. As the lead time increases from 3 to 10 days, the ROCSS for grid cell B reduces from 0.79 to 0.46 for the Pangu-Weather, from 0.79 to 0.46 for the GraphCast and from 0.82 to 0.51 for the FuXi. By contrast, the IFS HRES and IFS ENS change less and the ROCSS decreases from 0.81 to 0.55 for the IFS HRES and from 0.99 to 0.93 for the IFS ENS.

295

Figure 7. Time series plots of T2M24h forecasts initialized at 00 UTC for the IFS HRES, IFS ENS, Pangu-Weather, GraphCast and FuXi over three selected grid cells, i.e., A (30 ̊N, 105 ̊E), B (39 ̊N, 75 ̊W) and C (54 ̊N, 1.5 ̊W).

4.4 Sensitivity to predefined thresholds

300   The globally area-weighted performance under different predefined thresholds is illustrated for 5-day lead time in Figure

8. Among the sixteen metrics, the ROCSS is base-rate-independent and suitable simultaneously for deterministic and

probabilistic forecasts of binary events while other metrics need the predefined probability threshold to convert the

probabilistic forecasts to deterministic forecasts. In the meantime, it is noted that the SEDI is the most applicable to extreme

events because of its base-rate independence and nondegenerate limit (North et al., 2013; Jolliffe and Stephenson, 2012; Brodie

305   et al., 2024). These metrics changes little as the predefined thresholds increase from the 80th to the 99th percentile. Specifically,

as to forecast wet extremes at 5-day lead time, the scores of GraphCast decrease from 0.74 to 0.56 for SEDI and from 0.43 to

0.23 for ROCSS as the thresholds increase from the 80th to the 99th percentile. By contrast, the scores of GraphCast increase

from 0.81 to 0.98 for 1-BS, from 0.87 to 0.95 for ORSS and from 0.51 to 0.52 for SEDS. These metrics are not suitable for

hydroclimatic extremes because it contradicts that rarer events are often more difficult to predict (Ferro and Stephenson, 2011).

310

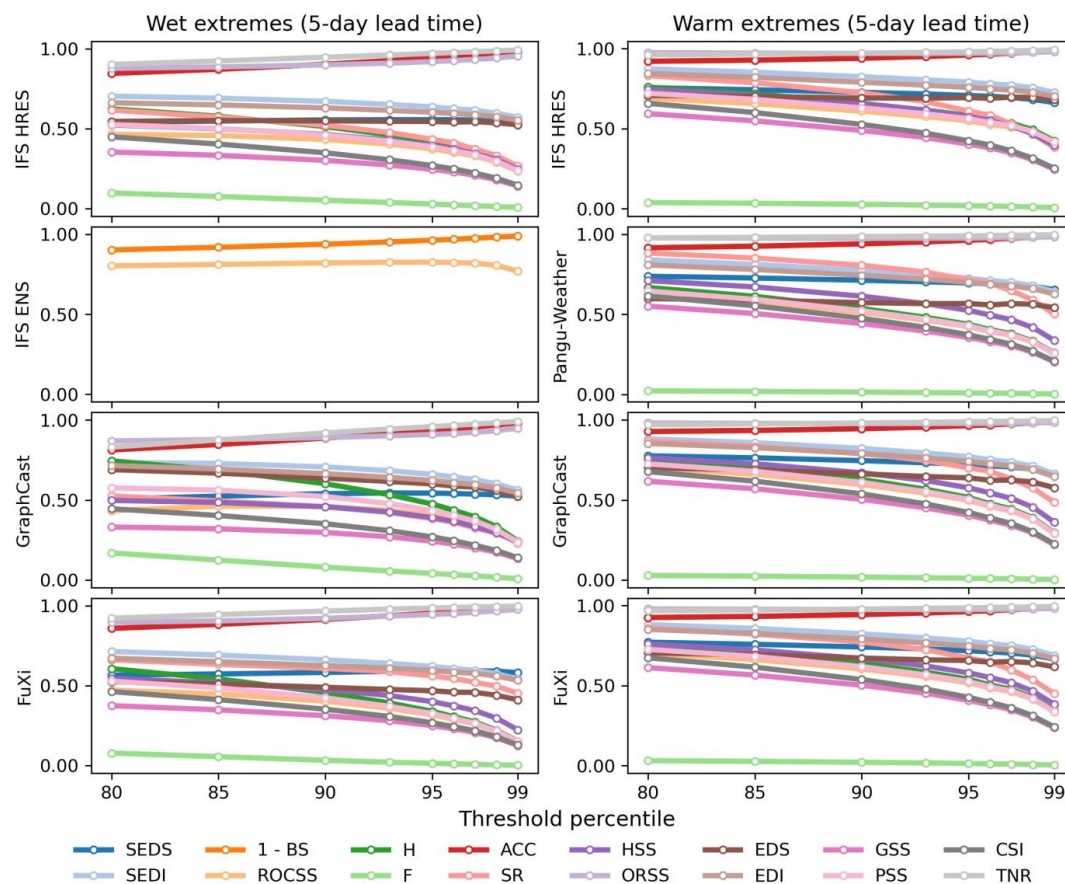



Figure 8. Globally area-weighted performance in forecasting wet extremes and warm extremes with different threshold percentiles at 5-day lead time.

315   The globally area-weighted ROCSS under different predefined thresholds is shown in Figure 9. Overall, the ROCSS decreases for all eight sets of forecasts as the predefined thresholds increase from the 80th to the 99th percentile. The IFS ENS tends to perform better in forecasting wet extremes and warm extremes. Among the available data-driven models, the GraphCast (operational) tends to be more skilful for wet extremes; for warm extremes, the FuXi tends to be more skilful model at lead time less than 5 days and the GraphCast tends to be better at lead time more than 5 days. Specifically, as to forecast wet

320   extremes at 5-day lead time, the ROCSS decreases from 0.46 to 0.24 for IFS HRES, from 0.80 to 0.77 for IFS ENS and from 0.53 to 0.26 for GraphCast (operational). As to forecast warm extremes at 5-day lead time, the ROCSS decreases from 0.69 to 0.41 for IFS HRES, from 0.93 to 0.83 for IFS ENS and from 0.70 to 0.29 for GraphCast. When the lead time is longer than 3 days, the GraphCast, GraphCast (operational) and FuXi tend to be more skilful in predicting warm extremes (Olivetti and Messori, 2024b).

325

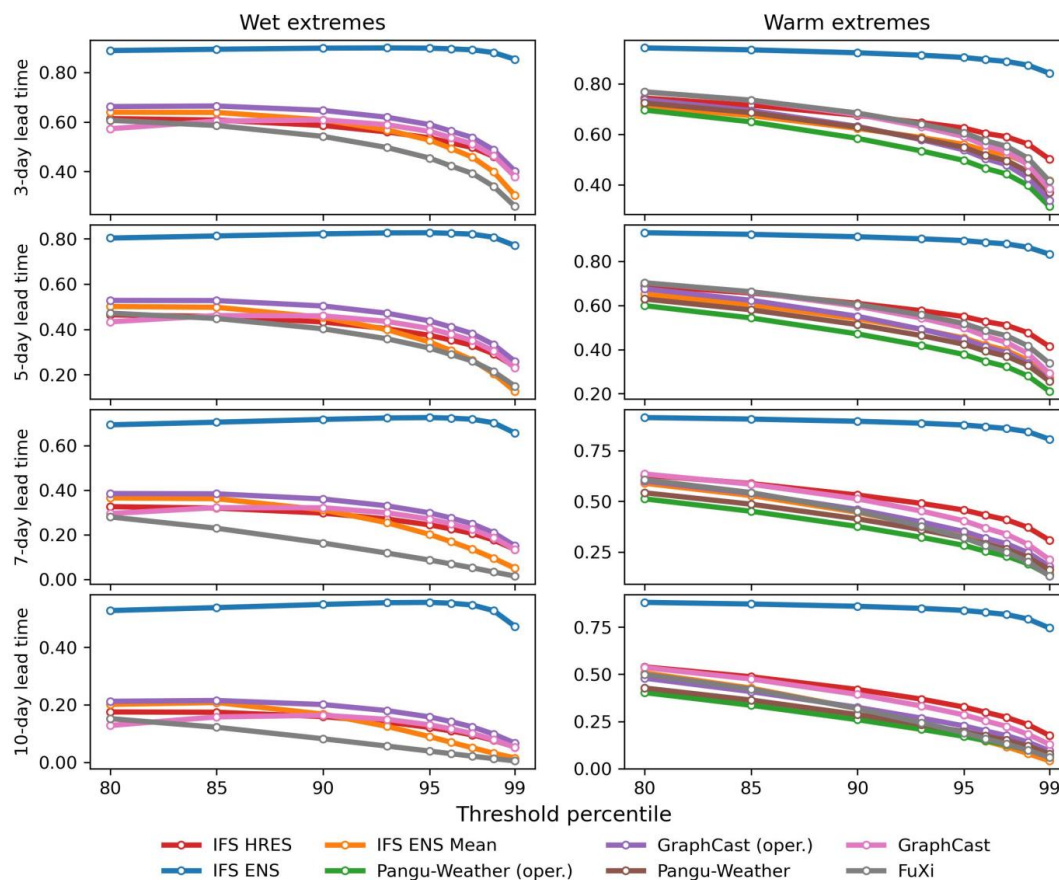



Figure 9. Globally area-weighted ROCSS for wet extremes and warm extremes with different threshold percentiles.

5 Discussion

330   Binary hydroclimatic forecasts provide useful information for disaster prevention and risk mitigation (Ben Bouallègue et al., 2024; Merz et al., 2020). For operational applications such as disaster warning, the focus is usually on the occurrence versus non-occurrence of certain hydroclimatic extremes instead of their precise magnitude (Jolliffe and Stephenson, 2012; Larraondo et al., 2020). In the meantime, binary forecasts emphasize the ability to capture hydroclimatic extremes that contribute little to average verification metrics, ensuring that models are not rewarded for merely minimizing average errors

335   and unrealistically smooth forecasts (Ferro and Stephenson, 2011; Rasp et al., 2020). Binary forecasts are thus more suitable than continuous forecasts in these cases. In this paper, the results show that as lead time increases, forecasts generated by data-driven models tend to be smoother and become less skilful more rapidly than the IFS HRES (Zhong et al., 2024; Rasp et al., 2024). In the meantime, ensemble forecasts can provide a range of potential hydroclimatic states and are important to

quantifying the probability of hydroclimatic extremes (Price et al., 2025; Zhao et al., 2022; Pasche et al., 2025).

340   High skill of data-driven models in forecasting wet extremes can stem from the unfair setting of ground truth data (Rasp et al., 2024; Lam et al., 2023). As is pointed out in the WeatherBench 2, the verification of precipitation using ERA5 data as ground truth data is a compromised setting and should be considered as a placeholder for more accurate precipitation data (Rasp et al., 2024). While this comparison is not fair to the IFS models, the results indicate that using data-driven models to forecast global medium-range precipitation is promising. In addition, the verification is limited to the wet and warm extremes

345   occurring in 2020 due to current data availability. The short verification period can only provide limited information about the model performance and sensitive results to the climate variability (Olivetti and Messori, 2024b). As more forecasts and more accurate baseline data are becoming available, the capability to produce binary forecasts of hydroclimatic extremes warrants further verification. The different roles that the operational IFS analysis and ERA5 data play in the initial conditions to generate forecasts also deserve further verification (Ben Bouallègue et al., 2024; Liu et al., 2024; Xu et al., 2024b).

350

6 Conclusions

This paper has presented an extension of the WeatherBench 2 to binary hydroclimatic forecasts by utilizing sixteen verification metrics. A case study is devised for binary forecasts generated by 3 data-driven models and 2 physical models. Specifically, the TP24h and T2M24h are calculated from the available forecasts and ground truth in the WeatherBench 2; and

355   the 90th percentiles of the ground truth data in 2020 are set as the predefined thresholds above which the wet and warm extremes are respectively detected. The results show that for wet extremes, the GraphCast and its operational version tend to outperform the IFS HRES when the total precipitation of ERA5 data is used as the ground truth. Their globally area-weighted ROCSS is 0.46, 0.50 and 0.43 at 5-day lead time, respectively. For warm extremes, the GraphCast and FuXi tend to be more skilful than the IFS HRES within 3-day lead time while they can be less skilful as the lead time increases. At the lead time of

360   3 and 10 days, the ROCSS is 0.68 and 0.42 for the IFS HRES, 0.92 and 0.86 for IFS ENS, 0.63 and 0.29 for Pangu-Weather, 0.68 and 0.39 for GraphCast and 0.68 and 0.32 for FuXi. When the predefined thresholds of wet extremes increase from the 80th to 99th percentile, the ROCSS decreases from 0.46 to 0.24 for IFS HRES, from 0.80 to 0.77 for IFS ENS and from 0.53 to 0.26 for GraphCast (operational) at 5-day lead time. The extension of the WeatherBench 2 to binary forecasts facilitates more comprehensive comparisons of hydroclimatic forecasts and provides useful information for forecast applications.

365

Code and data availability

The forecasts and ground truth data of the WeatherBench 2 are available from the Google Cloud bucket (https://console.cloud.google.com/storage/browser/weatherbench2) (Rasp et al., 2024). The ERA5 data are available from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/) at https://doi.org/10.24381/cds.bd0915c6 (Anon, 2023a)

370 and https://doi.org/10.24381/cds.adbb2d47 (Anon, 2023b). A subset of the ERA5 data is also available through the WeatherBench 2 (Rasp et al., 2024). The IFS HRES, its initial conditions and the IFS ENS are available from the ECMWF's MARS archive (https://confluence.ecmwf.int/display/UDOC/MARS+user+documentation). The IFS ENS Mean is available through the WeatherBench 2 (Rasp et al., 2024). The training code, pre-trained parameters and access details of the data-driven models are provided in the respective papers of these models (Chen et al., 2023; Bi et al., 2023; Lam et al., 2023).

375 The data and code performing the analysis of binary hydroclimatic forecasts are respectively archived on the Zenodo at https://doi.org/10.5281/zenodo.14691031 and https://doi.org/10.5281/zenodo.14691007 (Li and Zhao, 2025a, 2025b).

Author contributions

TZ: Writing – original draft, Visualization, Software, Methodology, Conceptualization. QL: Validation, Resources, Data 380 curation. TT: Investigation, Formal analysis. XC: Methodology, Conceptualization.

Competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

385

Acknowledgments

This research is supported by the National Natural Science Foundation of China (2023YFF0804900 and 52379033) and the Guangdong Provincial Department of Science and Technology (2019ZT08G090).

390 **References**

Agrawal, N., Nelson, P. V., and Low, R. D.: A Novel Approach for Predicting Large Wildfires Using Machine Learning

towards Environmental Justice via Environmental Remote Sensing and Atmospheric Reanalysis Data across the United States, Remote Sensing, 15, 5501, https://doi.org/10.3390/rs15235501, 2023.

395   Anon: ERA5 hourly data on pressure levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 10, https://doi.org/10.24381/cds.bd0915c6, 2023a.

Anon: ERA5 hourly data on single levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 10, https://doi.org/10.24381/cds.adbb2d47, 2023b.

Balsamo, G., Rabier, F., Balmaseda, M., Bauer, P., Brown, A., Dueben, P., English, S., McNally, T., Pappenberger, F., Sandu, I., Thepaut, J.-N., and Wedi, N.: Recent progress and outlook for the ECMWF Integrated Forecasting System, EGU-

400   13110, https://doi.org/10.5194/egusphere-egu23-13110, 2023.

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, https://doi.org/10.1038/nature14956, 2015.

Bauer, P., Quintino, T., Wedi, N., Bonanni, A., Chrust, M., Deconinck, W., Diamantakis, M., Düben, P., English, S., Flemming, J., and others: The ECMWF scalability programme: Progress and plans, European Centre for Medium Range Weather

405   Forecasts, 2020.

Ben Bouallègue, Z., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, 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: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context, B AM METEOROL SOC, 105, E864–E883, https://doi.org/10.1175/BAMS-

410   D-23-0162.1, 2024.

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

Brodie, S., Pozo Buil, M., Welch, H., Bograd, S. J., Hazen, E. L., Santora, J. A., Seary, R., Schroeder, I. D., and Jacox, M. G.: Ecological forecasts for marine resource management during climate extremes, Nat. Commun., 14, 7701,

415   https://doi.org/10.1038/s41467-023-43188-0, 2024.

de Burgh-Day, C. O. and Leeuwenburg, T.: Machine learning for numerical weather and climate modelling: a review, GEOSCI MODEL DEV, 16, 6433–6477, https://doi.org/10.5194/gmd-16-6433-2023, 2023.

Chakraborty, P., Dube, A., Sarkar, A., Mitra, A. K., Bhatla, R., and Singh, R. S.: How much does a high-resolution global ensemble forecast improve upon deterministic prediction skill for the Indian summer monsoon?, Meteorol. Atmos. Phys.,

420   135, 33, https://doi.org/10.1007/s00703-023-00966-1, 2023.

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 Clim Atmos Sci, 6, 1–11, https://doi.org/10.1038/s41612-023-00512-1, 2023.

Chen, X., Leung, L. R., Gao, Y., Liu, Y., Wigmosta, M., and Richmond, M.: Predictability of Extreme Precipitation in Western U.S. Watersheds Based on Atmospheric River Occurrence, Intensity, and Duration, Geophys. Res. Lett., 45, 11,693-

425   11,701, https://doi.org/10.1029/2018GL079831, 2018.

Clare, M. C. A., Jamil, O., and Morcrette, C. J.: Combining distribution‐based neural networks to predict weather forecast probabilities, Quart J Royal Meteoro Soc, 147, 4337–4357, https://doi.org/10.1002/qj.4180, 2021.

Coelho, G. de A., Ferreira, C. M., and Kinter III, J. L.: Multiscale and multi event evaluation of short-range real-time flood forecasting in large metropolitan areas, J. Hydrol., 612, 128212, https://doi.org/10.1016/j.jhydrol.2022.128212, 2022.

430   Donaldson, R. J., Dyer, R. M., and Kraus, M. J.: An objective evaluator of techniques for predicting severe weather events, in: Preprints, Ninth Conf. on Severe Local Storms, Norman, OK, Amer. Meteor. Soc, 321326, 1975.

Esteves, C., Slotine, J.-J., and Makadia, A.: Scaling Spherical CNNs, https://doi.org/10.48550/arXiv.2306.05420, 8 June 2023.

Ferro, C. A. T. and Stephenson, D. B.: Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events, WEATHER FORECAST, 26, 699–713, https://doi.org/10.1175/WAF-D-10-05030.1,

435   2011.

Finley, J. P.: Tornado predictions., American Meteorological Journal. A Monthly Review of Meteorology and Allied Branches

of Study (1884-1896), 1, 85, 1884.

Gilbert, G. K.: Finley's tornado predictions., American Meteorological Journal. A Monthly Review of Meteorology and Allied Branches of Study (1884-1896), 1, 166, 1884.

440   Gomis-Cebolla, J., Rattayova, V., Salazar-Galán, S., and Francés, F.: Evaluation of ERA5 and ERA5-Land reanalysis precipitation datasets over Spain (1951–2020), Atmos. Res., 284, 106606, https://doi.org/10.1016/j.atmosres.2023.106606, 2023.

Heidke, P.: Berechnung Des Erfolges Und Der Güte Der Windstärkevorhersagen Im Sturmwarnungsdienst, Geografiska Annaler, 8, 301–349, https://doi.org/10.1080/20014422.1926.11881138, 1926.

445   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.: The ERA5 global reanalysis, Quart J Royal
450   Meteoro Soc, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Huang, Z. and Zhao, T.: Predictive performance of ensemble hydroclimatic forecasts: Verification metrics, diagnostic plots and forecast attributes, WIREs Water, 9, e1580, https://doi.org/10.1002/wat2.1580, 2022.

Huang, Z., Zhao, T., Xu, W., Cai, H., Wang, J., Zhang, Y., Liu, Z., Tian, Y., Yan, D., and Chen, X.: A seven-parameter Bernoulli-Gamma-Gaussian model to calibrate subseasonal to seasonal precipitation forecasts, J. Hydrol., 610, 127896,
455   https://doi.org/10.1016/j.jhydrol.2022.127896, 2022.

Jacox, M. G., Alexander, M. A., Amaya, D., Becker, E., Bograd, S. J., Brodie, S., Hazen, E. L., Pozo Buil, M., and Tommasi, D.: Global seasonal forecasts of marine heatwaves, Nature, 604, 486–490, https://doi.org/10.1038/s41586-022-04573-9, 2022.

Jin, W., Weyn, J., Zhao, P., Xiang, S., Bian, J., Fang, Z., Dong, H., Sun, H., Thambiratnam, K., and Zhang, Q.: WeatherReal:
460   A Benchmark Based on In-Situ Observations for Evaluating Weather Models, https://doi.org/10.48550/arXiv.2409.09371, 14 September 2024.

Jolliffe, I. T. and Stephenson, D. B.: Forecast verification: a practitioner's guide in atmospheric science, John Wiley & Sons, 2012.

Keisler, R.: Forecasting Global Weather with Graph Neural Networks, http://arxiv.org/abs/2202.07575, 15 February 2022.

465   Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., and Hoyer, S.: Neural general circulation models for weather and climate, NATURE, 632, 1060–1066, https://doi.org/10.1038/s41586-024-07744-y, 2024.

Lagadec, L.-R., Patrice, P., Braud, I., Chazelle, B., Moulin, L., Dehotin, J., Hauchard, E., and Breil, P.: Description and evaluation of a surface runoff susceptibility mapping method, J. Hydrol., 541, 495–509,
470   https://doi.org/10.1016/j.jhydrol.2016.05.049, 2016.

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, 2023.

475   Larraondo, P. R., Renzullo, L. J., Van Dijk, A. I. J. M., Inza, I., and Lozano, J. A.: Optimization of Deep Learning Precipitation Models Using Categorical Binary Metrics, J Adv Model Earth Syst, 12, e2019MS001909, https://doi.org/10.1029/2019MS001909, 2020.

Li, Q. and Zhao, T.: Code for the extension of the WeatherBench 2 to binary hydroclimatic forecasts (v0.2.0), Zenodo, https://doi.org/10.5281/zenodo.14691007, 2025a.

480   Li, Q. and Zhao, T.: Data for the extension of the WeatherBench 2 to binary hydroclimatic forecasts (v0.1.0), https://doi.org/10.5281/zenodo.14691031, 2025b.

Liang, K.-Y. and Zeger, S. L.: Longitudinal data analysis using generalized linear models, Biometrika, 73, 13–22, https://doi.org/10.1093/biomet/73.1.13, 1986.

485 Liu, C.-C., Hsu, K., Peng, M. S., Chen, D.-S., Chang, P.-L., Hsiao, L.-F., Fong, C.-T., Hong, J.-S., Cheng, C.-P., Lu, K.-C., Chen, C.-R., and Kuo, H.-C.: Evaluation of five global AI models for predicting weather in Eastern Asia and Western Pacific, NPJ CLIM ATMOS SCI, 7, 1–12, https://doi.org/10.1038/s41612-024-00769-0, 2024.

Makkonen, L.: Plotting Positions in Extreme Value Analysis, Journal of Applied Meteorology and Climatology, 45, 334–340, https://doi.org/10.1175/JAM2349.1, 2006.

Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D. N., Domeisen, D. I. V., Feser, F., Koszalka, I., Kreibich, 490 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, Rev. Geophys., 58, e2020RG000704, https://doi.org/10.1029/2020RG000704, 2020.

North, R., Trueman, M., Mittermaier, M., and Rodwell, M. J.: An assessment of the SEEPS and SEDI metrics for the verification of 6 h forecast precipitation accumulations, Meteorol. Appl., 17, 2347–2358, 495 https://doi.org/10.1002/met.1405, 2013.

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

Olivetti, L. and Messori, G.: Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather, and GraphCast, Geosci. Model Dev., 17, 7915–7962, https://doi.org/10.5194/gmd-17-7915- 500 2024, 2024b.

Orozco López, E., Kaplan, D., Linhoss, A., Hogan, R. J., Ferro, C. A. T., Jolliffe, I. T., and Stephenson, D. B.: Equitability Revisited: Why the "Equitable Threat Score" Is Not Equitable, Wea. Forecasting, 25, 710–726, https://doi.org/10.1175/2009WAF2222350.1, 2010.

Ouyang, W., Ye, L., Chai, Y., Ma, H., Chu, J., Peng, Y., Zhang, C., Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. 505 R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., Willson, M., Benjamini, Y., and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society Series B: Statistical Methodology, 57, 289–300, https://doi.org/10.1111/j.2517- 6161.1995.tb02031.x, 1995.

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., and Engelke, S.: Validating Deep Learning Weather Forecast Models on 510 Recent High-Impact Extreme Events, Artificial Intelligence for the Earth Systems, 4, https://doi.org/10.1175/AIES-D- 24-0033.1, 2025.

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, 22 515 February 2022.

Peirce, C. S.: The Numerical Measure of the Success of Predictions, Science, ns-4, 453–454, https://doi.org/10.1126/science.ns-4.93.453.b, 1884.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M.: Probabilistic weather forecasting with machine learning, NATURE, 637, 84–90, 520 https://doi.org/10.1038/s41586-024-08252-9, 2025.

Primo, C. and Ghelli, A.: The affect of the base rate on the extreme dependency score, Meteorol. Appl., 16, 533–535, https://doi.org/10.1002/met.152, 2009.

Rasp, S. and Thuerey, N.: Data‐Driven Medium‐Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench, J. Adv. Model. Earth Syst., 13, e2020MS002405, 525 https://doi.org/10.1029/2020MS002405, 2021.

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, J ADV MODEL EARTH SY, 12, https://doi.org/10.1029/2020MS002203, 2020.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez‐Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F.: 530 WeatherBench 2: A Benchmark for the Next Generation of Data‐Driven Global Weather Models, J ADV MODEL EARTH SY, 16, e2023MS004019, https://doi.org/10.1029/2023MS004019, 2024.

Schaefer, J. T.: The Critical Success Index as an Indicator of Warning Skill, Wea. Forecasting, 5, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2, 1990.

Shen, H., Tolson, B. A., and Mai, J.: PRACTITIONERS' CORNER: Computing Robust Standard Errors for Within‐groups 535 Estimators, Oxford B. Econ. Stat., 49, 431–434, https://doi.org/10.1111/j.1468-0084.1987.mp49004006.x, 1987.

Stephenson, D. B.: Use of the "Odds Ratio" for Diagnosing Forecast Skill, 2000.

Stephenson, D. B., Casati, B., Ferro, C. A. T., and Wilson, C. A.: The extreme dependency score: a non‐vanishing measure for forecasts of rare events, Meteorol. Appl., 15, 41–50, https://doi.org/10.1002/met.53, 2008.

Swets, J. A.: Indices of discrimination or diagnostic accuracy: Their ROCs and implied models., Psychol. Bull., 99, 100–117, 540 https://doi.org/10.1037/0033-2909.99.1.100, 1986.

Swets, J. A. and Swets, J. A.: Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance., Psychol. Bull., 99, 181–198, https://doi.org/10.1037/0033-2909.99.2.181, 1986.

Weyn, J. A., Durran, D. R., and Caruana, R.: Improving Data‐Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere, J. Adv. Model. Earth Syst., 12, e2020MS002109, 545 https://doi.org/10.1029/2020MS002109, 2020.

Wilks, D. S.: "The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It, B AM METEOROL SOC, 97, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1, 2016.

Xu, H., Duan, Y., and Xu, X.: Evaluating AI's capability to reflect physical mechanisms: a case study of tropical cyclone 550 impacts on extreme rainfall, Environ. Res. Lett., 19, 104006, https://doi.org/10.1088/1748-9326/ad6fbb, 2024a.

Xu, H., Zhao, Y., Zhao, D., Duan, Y., and Xu, X.: Improvement of disastrous extreme precipitation forecasting in North China by Pangu-weather AI-driven regional WRF model, Environ. Res. Lett., 19, 054051, https://doi.org/10.1088/1748-9326/ad41f0, 2024b.

Zhao, T., Xiong, S., Wang, J., Liu, Z., Tian, Y., Yan, D., Zhang, Y., Chen, X., and Wang, H.: A Two-Stage Framework for 555 Bias and Reliability Tests of Ensemble Hydroclimatic Forecasts, Water Resources Research, 58, e2022WR032568, https://doi.org/10.1029/2022WR032568, 2022.

Zhong, X., Chen, L., Liu, J., Lin, C., Qi, Y., and Li, H.: FuXi-Extreme: Improving extreme rainfall and wind forecasts with diffusion model, Sci. China Earth Sci., https://doi.org/10.1007/s11430-023-1427-x, 2024.

560