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;

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, seventeen verification metrics on the accuracy and discrimination of binary forecasts are employed and scorecards 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 precipitation of ERA5 reanalysis data as the ground truth. For warm extremes, the Pangu-Weather, GraphCast and FuXi tend 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 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.

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 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 remarkable 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).

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 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 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 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). Operational applications usually pay attention to the

occurrence versus non-occurrence of certain hydroclimatic extremes instead of their precise magnitude (Larraondo et al., 2020; Rasp et al., 2020). Binary forecasts meet this demand by emphasizing the ability to capture hydroclimatic extremes, ensuring that models are not rewarded for merely minimizing average errors and unrealistically smooth forecasts (Ferro and Stephenson, 2011; 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, the extension facilitates an effective intercomparison among binary forecasts of hydroclimatic extremes generated by both data-driven and physical models.

2 Forecasts and metrics in the WeatherBench 2

2.1 Forecast datasets

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 al., 2023). In the WeatherBench 2, the HRES is primarily used as the 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 WeatherBench 2, the ENS also serves as an important baseline, with the mean value of the 50 members, i.e., ENS Mean, being extensively used (Rasp et al., 2024).

3) The 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. It is noted that two sets of Pangu-Weather forecasts, which are respectively based on the ERA5 and HRES initializations, are generated (Rasp et al., 2024).

4) The 10-day forecasts generated by the GraphCast includes 6 upper-air variables at a maximum of 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 2 time steps at the temporal resolution of 6 hours. Similarly, there are two sets of GraphCast forecasts generated from the ERA5 and HRES initializations (Rasp et al., 2024).

5) 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 based on the U-Transformer architecture. It consists of three sub-models fine-tuned for forecasting 0-5, 5-10 and 10-15 days ahead at the temporal resolution of 6 hours. Atmospheric states are forecasted based on states from the previous 2 time steps.

2.2 Verification metrics

The WeatherBench 2 takes into consideration in total 6 metrics for deterministic forecasts and 6 metrics for ensemble forecasts, as shown in Table 1. The ERA5 reanalysis 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 the verification of IFS forecasts (Lam et al., 2023). As precipitation is not available for the IFS HRES's initial conditions, the total precipitation of ERA5 reanalysis data is used as the ground truth data for all models. Following the initial version of the WeatherBench 2, the verification is conducted for forecasts initialized at 00 and 12 UTC for the period from 1 January 2020 to 31 December 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. Verification metrics of 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). In operational applications, binary forecasts of extreme precipitation events and heatwaves can respectively be derived from precipitation and temperature forecasts (Huang and Zhao, 2022; Lang et al., 2014; Zhao et al., 2022; Slater et al., 2023). 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; Xiong et al., 2024). 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 (Xiong et al., 2024; Zhao et al., 2024). It is noted that the thresholds at each grid cell are separately calculated. Given the pre-defined threshold q, deterministic forecasts are converted into either 0 or 1:

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

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

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

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

The comparison of binary forecasts against the corresponding observations facilitate four categories, i.e., true positives (a), false positives (b), false negatives (c) and true negatives (d), as shown in Table 2 (Larraondo et al., 2020). Specifically, the true positives indicate that target occurrences are successfully forecasted; the false positives indicate non-occurrences incorrectly forecasted as occurrences; the false negatives indicate target occurrences incorrectly forecasted as non-occurrences; and the true negatives indicate non-occurrences that are correctly forecasted as non-occurrences. The proportion of the observed occurrences to the total number of occurrences and non-occurrences is the base rate ($(a+c)/N$), with lower values often corresponding to events that are more extreme (Ferro and Stephenson, 2011).

Table 2. Contingency table for binary forecasts.

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

Forecasted non-occurrences	$c = \begin{cases} \sum_{n=1}^{N} I(f_n \le q \mid p_{o_n} = 1), & if \ \ M = 1 \\ \sum_{n=1}^{N} I(p_{f_n} \le p \mid p_{o_n} = 1), & if \ \ M \ge 2 \end{cases}$	$d = \begin{cases} \sum_{n=1}^{N} I(f_n \le q \mid p_{o_n} = 0), & if \ \ M = 1 \\ \sum_{n=1}^{N} I(p_{f_n} \le p \mid p_{o_n} = 0), & if \ \ M \ge 2 \end{cases}$	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = N$

Where $M = 1$ and $M \geq 2$ respectively represent deterministic and ensemble forecasts; N is the number of pairs of observations and forecasts for verification; p_{o_n} represents the binary observation which is 1 for occurrences and 0 for non-occurrences; p denotes the probability thresholds above which occurrences are forecasted to occur for ensemble forecasts.

3.2 Verification metrics for binary forecasts

Given the challenges posed by varying hydroclimatic extremes and imbalanced samples, in total 17 metrics are utilized to examine the performance of binary forecasts (Jolliffe and Stephenson, 2012; North et al., 2013). Notably, there are 8 base-rate-dependent metrics and 9 base-rate-independent metrics. On the one hand, the base-rate-dependent metrics facilitate 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 applicable to 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 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	$ACC = \frac{a+d}{N}$	[0, 1]	1	(Finley, 1884)
Success ratio (SR), precision	$SR = \frac{a}{a+b}$	[0, 1]	1	(Lagadec et al., 2016)
Critical success index (CSI), threat score, Gilbert score	$CSI = \frac{a}{a+b+c}$	[0, 1]	1	(Donaldson et al., 1975; Gilbert, 1884)
Gilbert skill score (GSS), equitable threat score	$GSS = \frac{a-a_r}{a+b+c-a_r}, a_r = \frac{(a+b)(a+c)}{N}$	[−1/3, 1]	1	(Gilbert, 1884; Schaefer, 1990)
Heidke skill score (HSS), Cohen's Kappa	$HSS = \frac{a+d-a_r-d_r}{N-a_r-d_r}, d_r = \frac{(b+d)(c+d)}{N}$	[−1, 1]	1	(Gomis-Cebolla et al., 2023; Heidke, 1926)
Extreme dependence score (EDS)	$EDS = \frac{\ln[(a+c)/N] - \ln H}{\ln[(a+c)/N] + \ln H}$	[−1, 1]	1	(Primo and Ghelli, 2009; Stephenson et al., 2008)
Symmetric extreme dependence score (SEDS)	$SEDS = \frac{\ln[(a+b)/N] - \ln H}{\ln[(a+c)/N] + \ln H}$	[−1, 1]	1	(Orozco López et al., 2010)
Potential relative economic value (REV)	$REV = \max_{0 \le p \le 1} \frac{\min\{a+c, r\} - [(a+b)r + c]}{\min\{a+c, r\} - (a+c)r}$	[0, 1]	1	(Richardson, 2006, 2000; Wilks, 2001)

Base-rate-independent metrics				
Hit rate (H), sensitivity, recall, probability of detection	$H = \frac{a}{a+c}$	[0, 1]	1	(Swets, 1986)
False alarm rate (F), probability of false detection	$F = \frac{b}{b+d}$	[0, 1]	0	(Donaldson et al., 1975)
Specificity, true negative rate (TNR)	$TNR = \frac{d}{b+d}$	[0, 1]	1	(Agrawal et al., 2023)
Odds ratio skill score (ORSS), Yule's Q	$ORSS = \frac{ad-bc}{ad+bc}$	[−1, 1]	1	(Stephenson, 2000)
Peirce's skill score (PSS), Hanssen and Kuipers discriminant	$PSS = \frac{ad-bc}{(a+c)(b+d)} = H - F$	[−1, 1]	1	(Peirce, 1884)
Extremal dependence index (EDI)	$EDI = \frac{\ln F - \ln H}{\ln F + \ln H}$	[−1, 1]	1	(Ferro and Stephenson, 2011)
Symmetric extremal dependence index (SEDI)	$SEDI = \frac{\ln F - \ln H + \ln(1-H) - \ln(1-F)}{\ln F + \ln H + \ln(1-H) + \ln(1-F)}$	[−1, 1]	1	(Ferro and Stephenson, 2011)
Area under receiver operating characteristic (ROC) curve (AUC)	$AUC = \int_0^1 HdF$	[0, 1]	1	(Swets, 1986)
ROC skill score (ROCSS)	$ROCSS = 2(AUC - 0.5)$	[−1, 1]	1	(Swets and Swets, 1986)

Where a, b, c and d respectively denote the number of true positives, false positives, false negatives and true negatives, with the equations shown in Table 2; N is the number of pairs of observations and forecasts; p denotes the probability thresholds above which the events are forecasted to occur for ensemble forecasts; r represents the cost-loss ratio for calculating the relative economic value; all calculation equations of other variables can be found in this table.

The 8 base-rate-dependent metrics in Table 3 are influenced by the underlying distribution of observed occurrences and non-occurrences (Jolliffe and Stephenson, 2012). The accuracy is calculated as the ratio between the number of true positives and the total number of occurrences and non-occurrences (Finley, 1884). The success ratio (SR) measures the number of true positives divided by the number of forecasted occurrences (Lagadec et al., 2016). The critical success index (CSI) is the number of true positives divided by the total number of forecasted and observed occurrences (Chakraborty et al., 2023; Gilbert, 1884; Donaldson et al., 1975). The Gillert skill score (GSS) evaluates the fraction of true positives over the observed and forecasted occurrences after adjusting for the random true positives (Chen et al., 2018; Coelho et al., 2022). The Heidke skill score (HSS) measures the accuracy relative to that of the random forecasts (Gomis-Cebolla et al., 2023). The extreme dependency score (EDS) (Stephenson et al., 2008) and the symmetric extreme dependency score (SEDS) (Orozco López et al., 2010) can measure the general performance of binary forecasts for rare events. The potential relative economic value (REV) quantifies the potential value of a forecast over a range of different probability thresholds (p) to make decision (Richardson, 2006, 2000; Wilks, 2001). It compares the saved expense using the forecasts instead of climatology relative to the saved expense using the

perfect forecast (Price et al., 2025).

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 occurrences (Ferro and Stephenson, 2011). The hit rate and false alarm rate respectively quantify the proportion of true positives in observed occurrences and the proportion of false positives in observed non-occurrences (Swets, 1986). The specificity measures the percentage of true negatives to observed non-occurrences (Agrawal et al., 2023). The odds ratio skill score (ORSS) examines the improvement over the random forecasts, emphasizing the balance between positive and negative samples (Stephenson, 2000). The Peirce's skill score (PSS) has similar formulation to HSS but does not depend on event frequency (Chakraborty et al., 2023). For deterministic forecasts, the PSS equals to the maximum value of REV when the cost-loss ratio equals to the base rate (Richardson, 2006). The extremal dependence index (EDI) and the symmetric extremal dependence index (SEDI) are designed to be nondegenerate to measure the predictive performance for rare events. (Ferro and Stephenson, 2011). The receiver operating characteristic (ROC) examines the discrimination between true positives and false positives, quantified by the area under the ROC curve (AUC) (Swets, 1986). The ROC skill score (ROCSS) compares the discriminative ability over random forecasts.

Among the 17 metrics, the ROCSS is base-rate-independent and suitable for both deterministic and probabilistic forecasts of binary events. By contrast, the other metrics need some predefined probability thresholds to convert probabilistic forecasts into deterministic forecasts. Therefore, the ROCSS is selected as the primary verification metric in the analysis. For probabilistic forecasts, the ROCSS is calculated by considering the hit rate and false alarm rate for all possible thresholds of probability (Huang and Zhao, 2022). It is noted that higher ROCSS values indicate better forecast skill.

3.3 Forecast verification

Considering data availability and forecast settings, the verification focuses on 8 sets of forecasts: IFS's HRES, ENS and ENS Mean; operational forecasts from Pangu-Weather, GraphCast; and hindcasts from Pangu-Weather, GraphCast and FuXi. 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 considered for sensitivity analysis (Olivetti and Messori, 2024b; North et al., 2013). For the comparison at individual grid cells, the 17 metrics are one by one calculated. Furthermore, the 17 metrics are calculated using the area-weighting method for the regions pre-determined by the ECMWF's scorecards, as shown in Table 4 (Rasp et al., 2024).

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°

AusNZ: Australia and New Zealand.

Considering that hydroclimatic observations are subject to heteroscedasticity and autocorrelation due to spatial and temporal clustering of hydroclimatic extremes (Olivetti and Messori, 2024b), the cluster-robust standard errors are used to correct the paired t test (Liang and Zeger, 1986; Shen et al., 1987). Specifically, the corrected two-sided paired t test is performed 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). For comparison at individual grid cells, 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 (Benjamini and Hochberg, 1995; Olivetti and Messori, 2024b). This setting corresponds to the significance level of 0.05 for spatially correlated hydroclimatic extremes (Wilks, 2016).

4 Results

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 skilful as lead time increases from 1 day to 10 days. This outcome is in general due to the accumulation of forecast errors over time caused by the autoregressive architecture of these models (Olivetti and Messori, 2024b; Bonavita, 2024). 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 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 are generally less skilful than the IFS HRES. This result is not surprising since the over-smoothing is observed to be more prominent among data-driven models than physical models (Bonavita, 2024; Lam et al., 2023). 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 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.

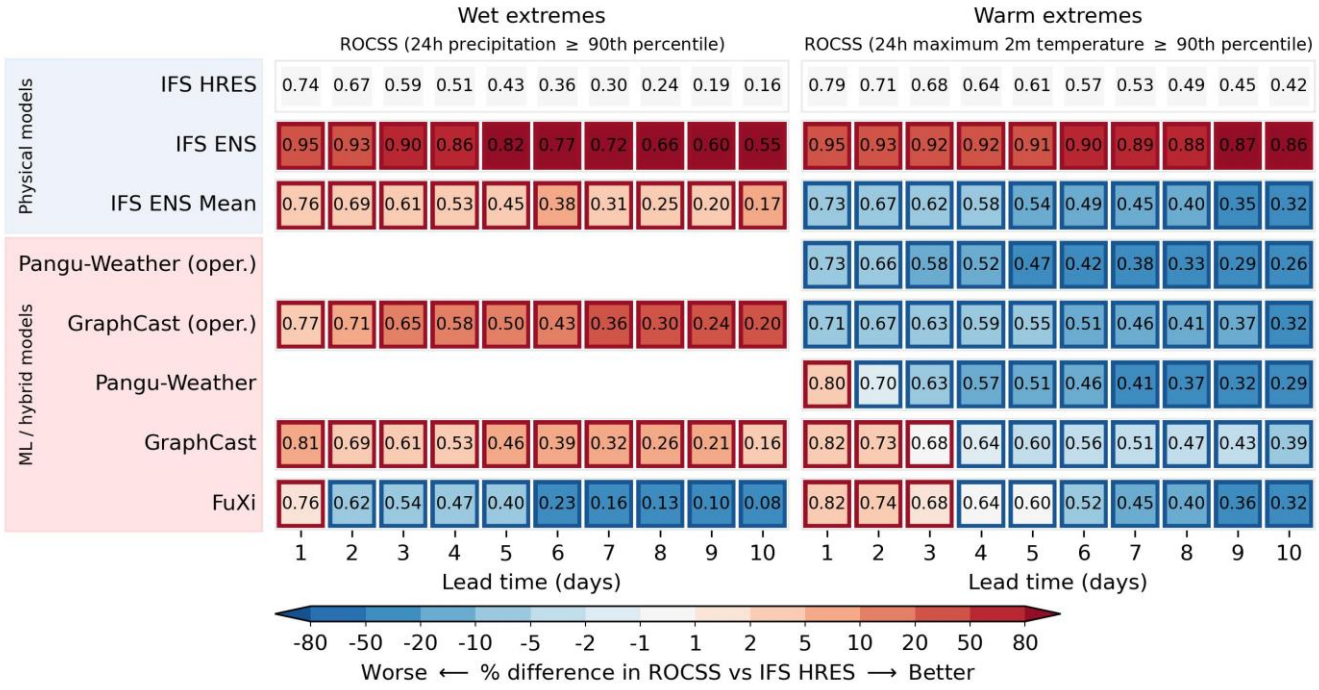

Figure 1. Globally area-weighted ROCSS for wet and warm extremes. The oper. denotes the operational version. The red and blue borders indicate significantly different performances compared to the IFS HRES at the significance level of 0.05.

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 tends 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 times of 3 and 10 days.

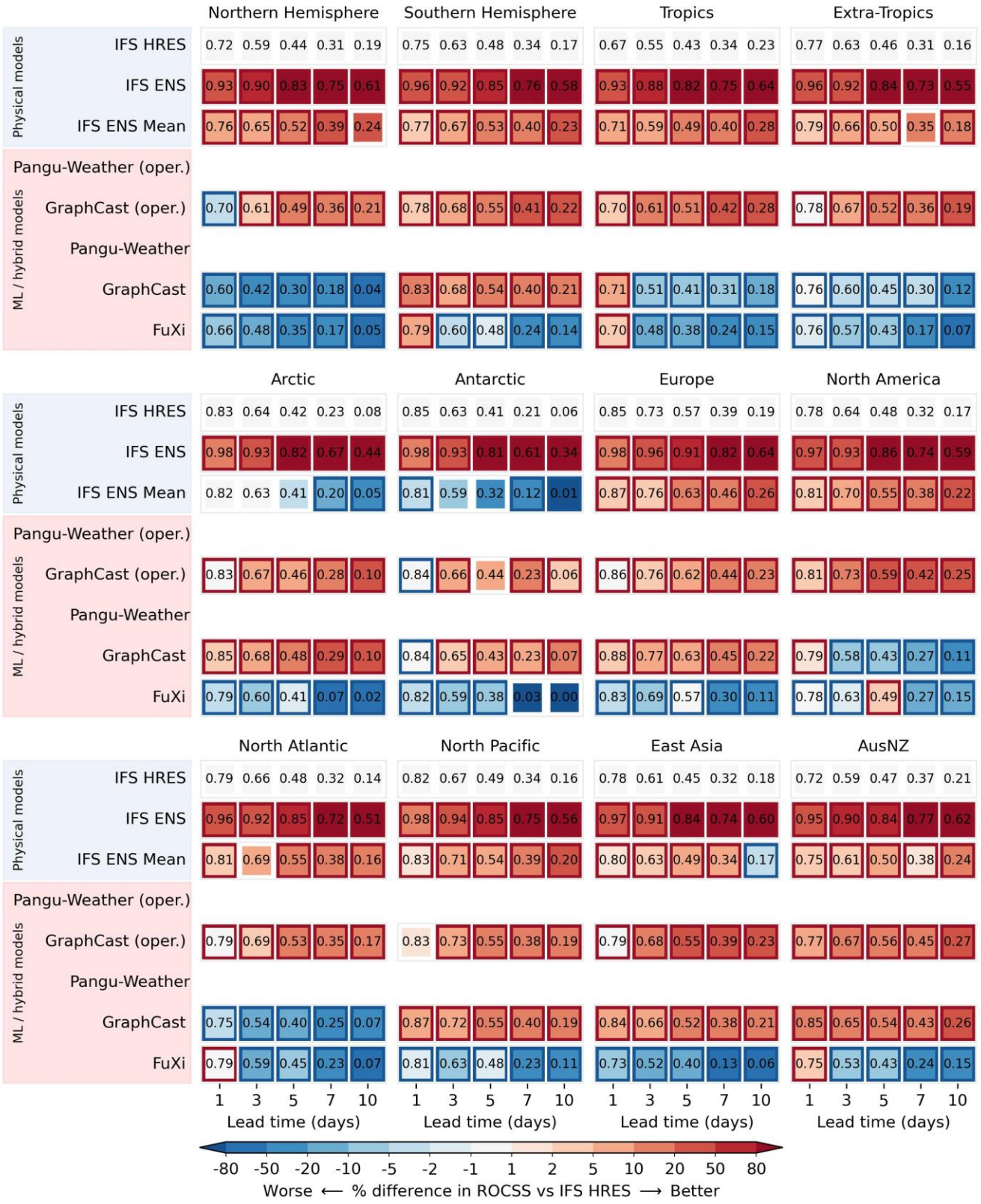

Figure 2. Regionally area-weighted ROCSS of different forecasts for wet extreme. The red and blue borders indicate significantly different performance compared to the IFS HRES at the significance level of 0.05.

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 the magnitude for warm extremes (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 times 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 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).

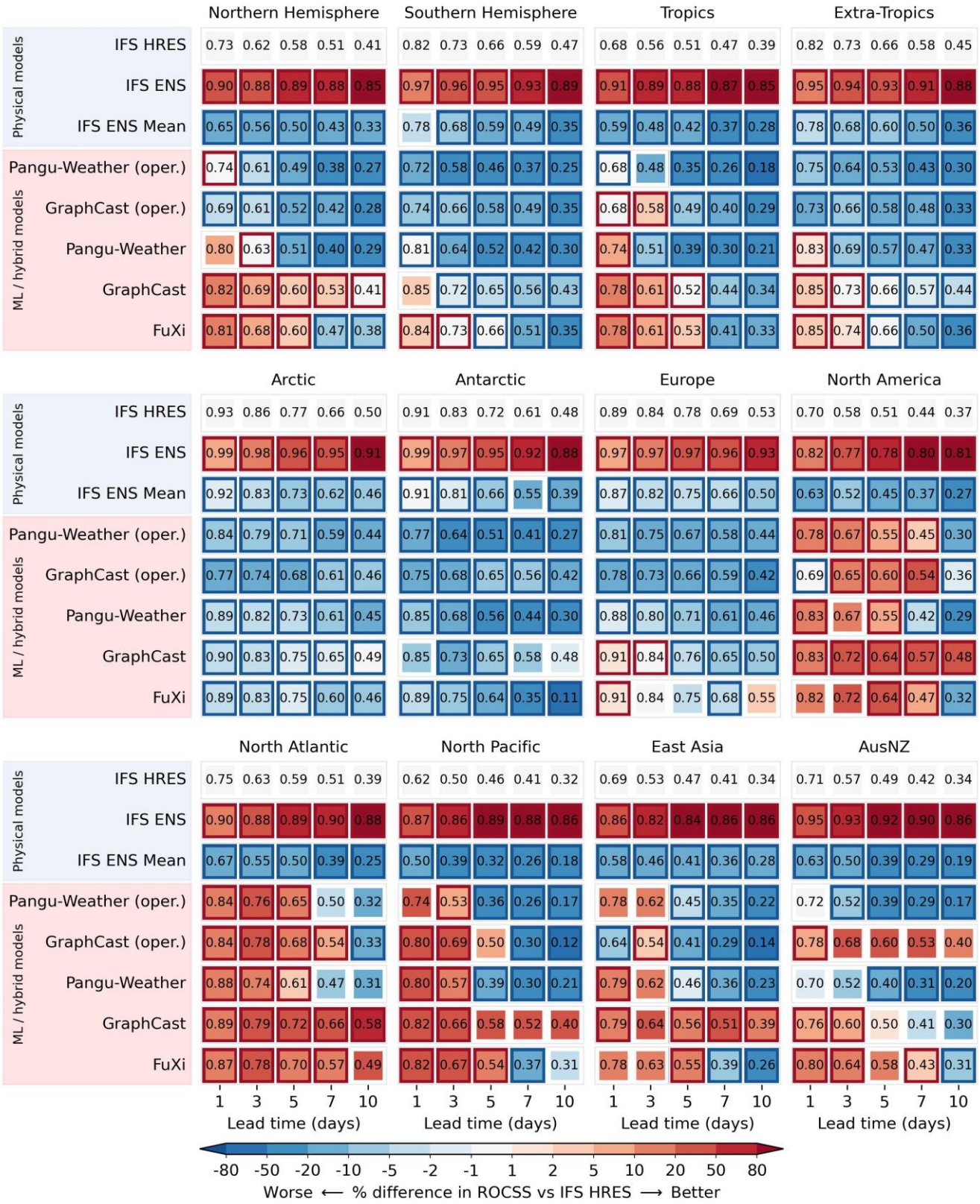

Figure 3. As for Figure 2, but for warm extremes.

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 GraphCast is not as skilful as the IFS HRES in more grid cells, such as the Northern Africa, Central Australia and Central Asia. The FuXi tends to be less skilful than the IFS HRES in most grid cells, such as the Northern Africa, 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 and so many false positives for many of or even almost all the grid cells in this region, the ROCSS is nearly -1 so that their forecasts tend to be worse than the IFS HRES in the Northern Hemisphere (extra-tropics) and Tropics.

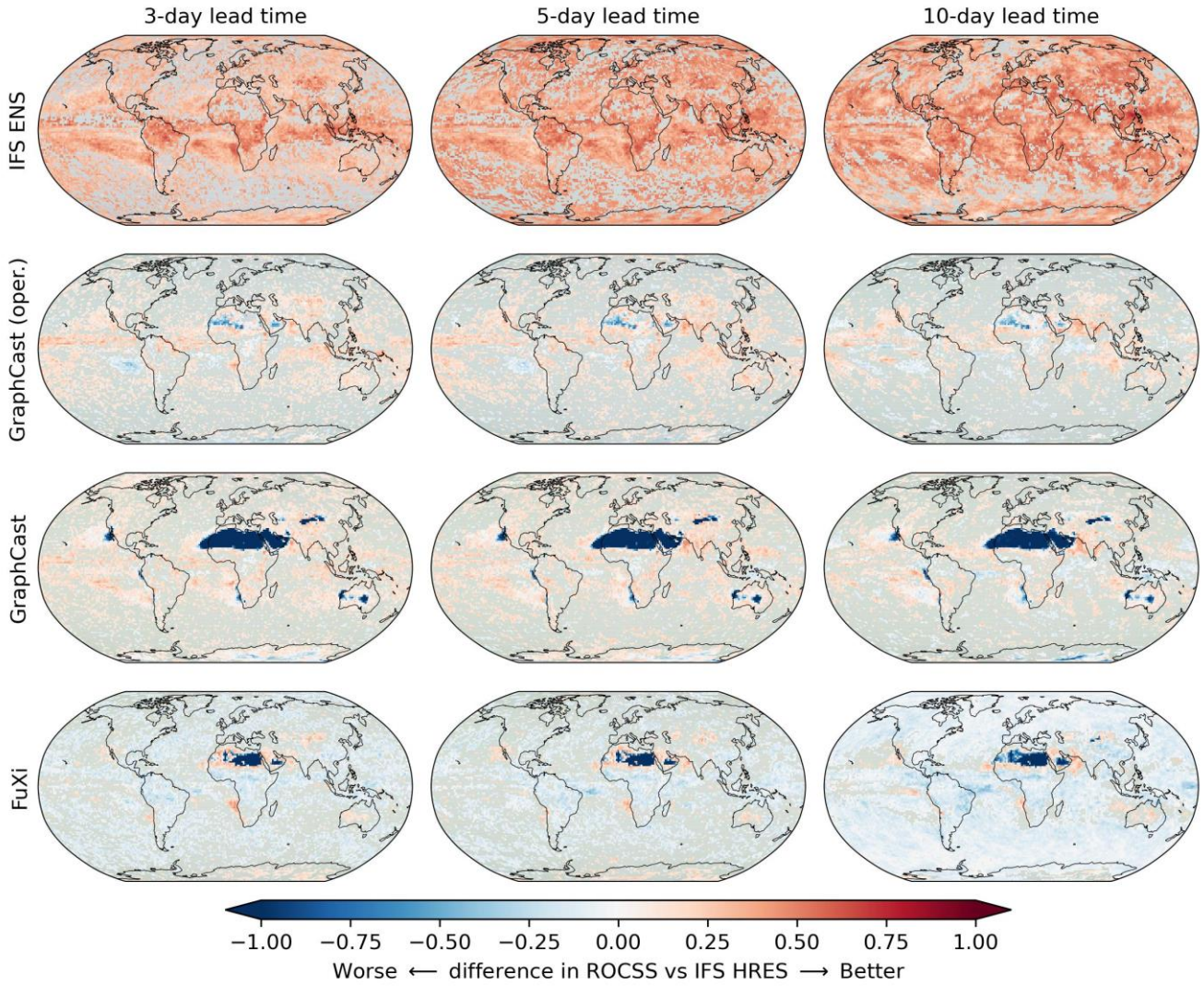

Figure 4. Differences of IFS ENS, GraphCast (operational), GraphCast and FuXi in ROCSS to the IFS HRES for wet extremes at each grid cell. The grey colour indicates grid with no statistically significant differences at the significance level of 0.1.

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

for three grid cells in Figure 5. The grid cells A, B and C are selected respectively due to the better, close and worse performance of data-driven models in relative to the IFS HRES. 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). For grid cells A and B, the five sets of forecasts have nearly equal number of true negatives; the IFS HRES show more true positives but more false negatives; the GraphCast is more capable of capturing the wet extremes but tends to produce more false positives; the IFS ENS Mean and FuXi tend to underestimate the wet extremes, resulting in more false negatives but fewer false positives. For grid cell C that is located in the Northern Africa, the GraphCast and FuXi tend to overestimate the low precipitation and underestimate the high precipitation, leading to zero numbers of true negatives for the FuXi and zero numbers of false negatives for both. At the lead times of 3 and 10 days, the ROCSS is respectively 0.48 and 0.09 for the IFS HRES, 0.80 and 0.53 for the IFS ENS, 0.31 and 0.21 for the operational GraphCast, -0.94 and -0.96 for the GraphCast and -1.00 and -1.00 for the FuXi.

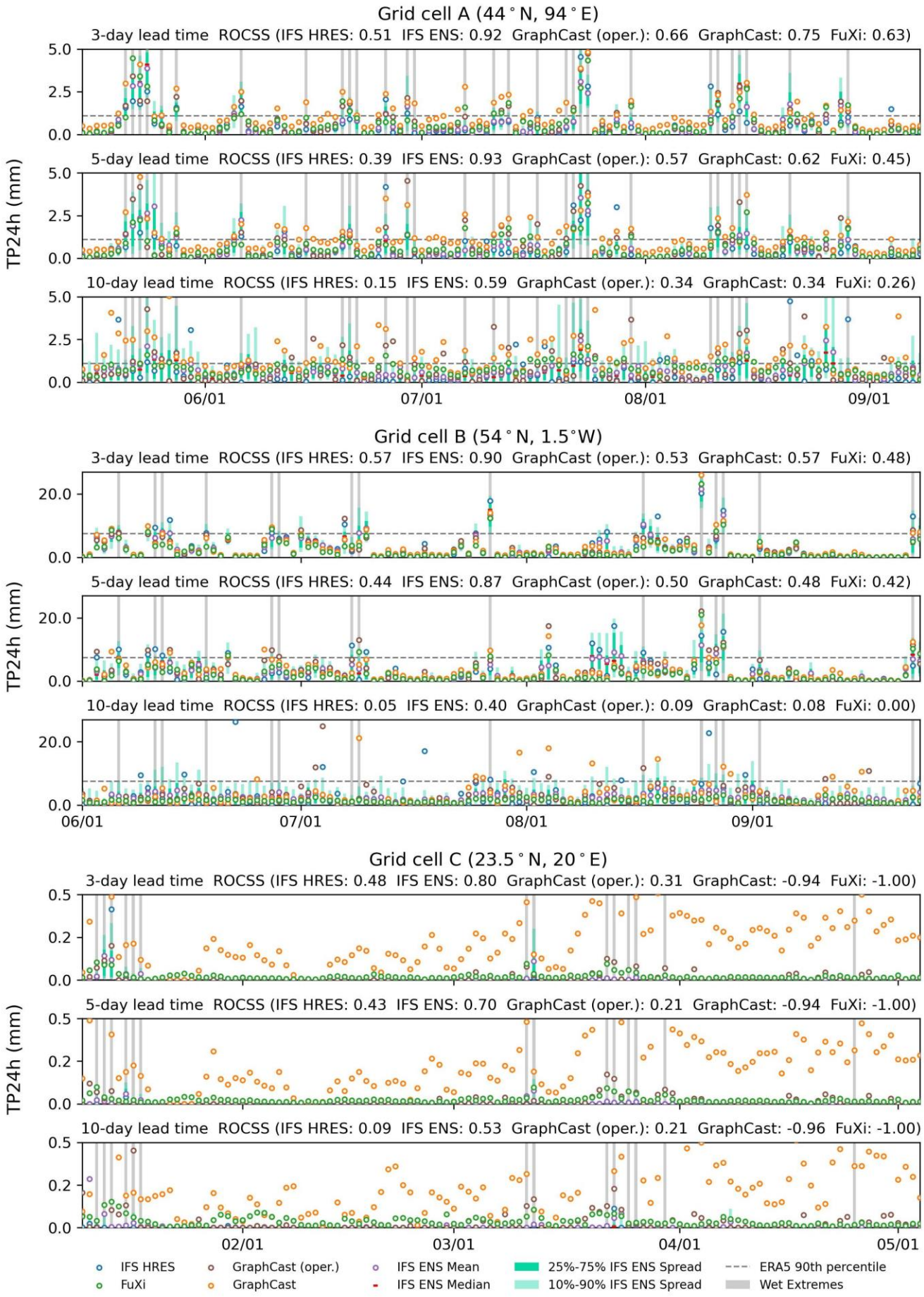

Figure 5. Time series plots of TP24h forecasts initialized at 00 UTC for the IFS HRES, IFS ENS, IFS ENS Mean, GraphCast and FuXi over three selected grid cells, i.e., A (44 °N, 94 °E), B (54 °N, 1.5 °W) and C (23.5 °N, 20 °E).

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 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 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).

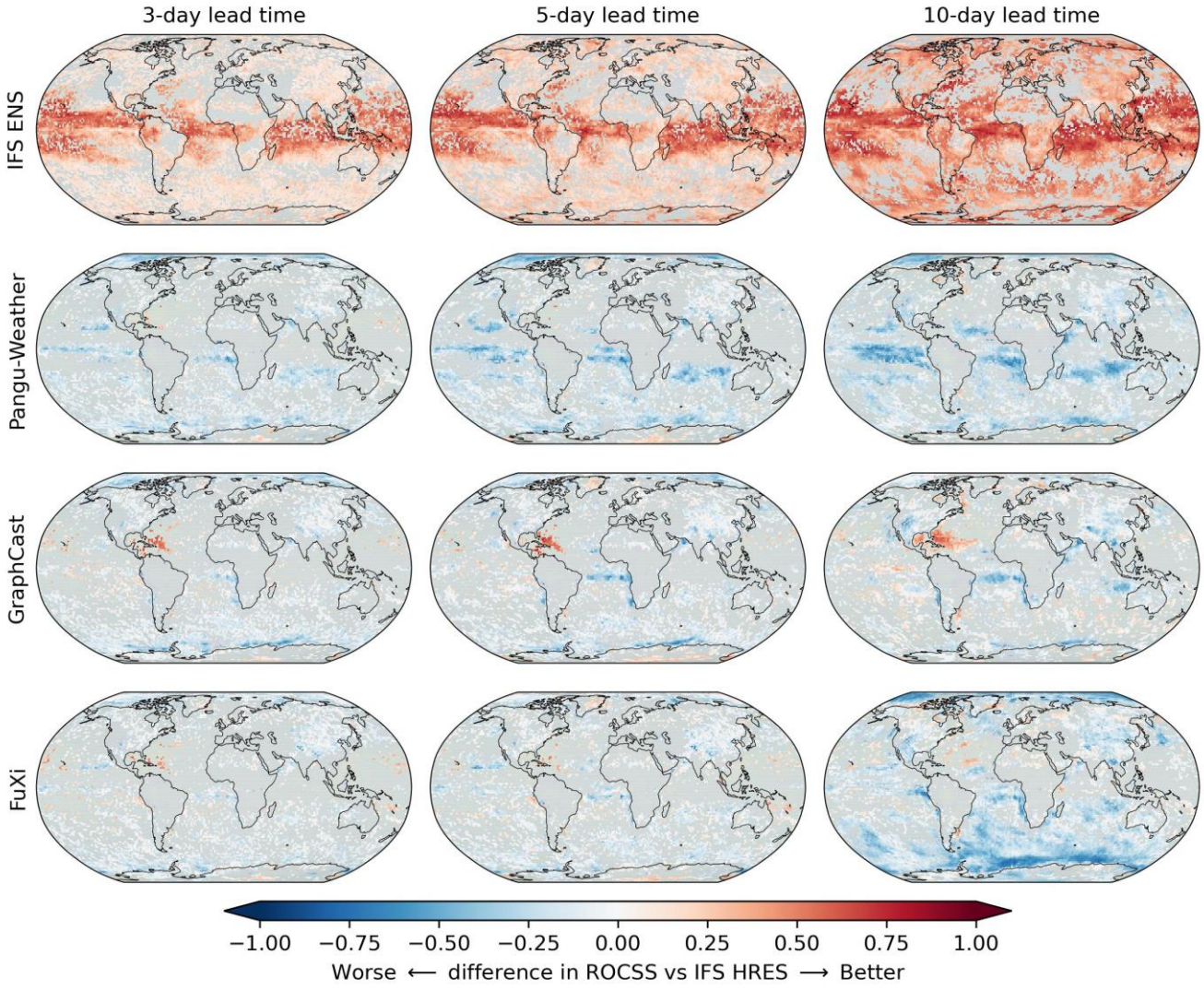

Figure 6. Differences of IFS ENS, Pangu-Weather, GraphCast and FuXi in ROCSS to the IFS HRES for warm extremes at each grid cell. The grey colour indicates grid with no statistically significant differences at the significance level of 0.1.

The time series for 24-hour maximum of 2m temperature from different forecasts initialized at 00 UTC are shown for three grid cells in Figure 7. The grid cells D, E and F are also selected respectively due to the better, close and worse performance of data-driven models in relative to the IFS HRES. Overall, the Pangu-Weather, GraphCast and FuXi exhibit similar temperature dynamics over time to those of the IFS HRES. For grid cell D, the Pangu-Weather, GraphCast and FuXi tend to outperform the IFS HRES. The Pangu-Weather tends to underestimate the temperature, leading to less true positives and more false negatives. The GraphCast and FuXi show more true positives. For grid cell E, these models show a nearly equal number of true positives and true negatives, resulting in similar ROCSS. For grid cell F, the data-driven models tend to be less accurate than the IFS HRES. The Pangu-Weather, GraphCast and FuXi tend to underestimate the temperature, leading to more false negatives and less true positives. As the lead time increases from 3 to 10 days, the ROCSS reduces from 0.48 to 0.28 for the Pangu-Weather, from 0.51 to 0.22 for the GraphCast and from 0.54 to 0.17 for the FuXi. By contrast, the IFS HRES and IFS ENS change less. The ROCSS decreases from 0.76 to 0.56 for the IFS HRES and from 0.95 to 0.86 for the IFS ENS.

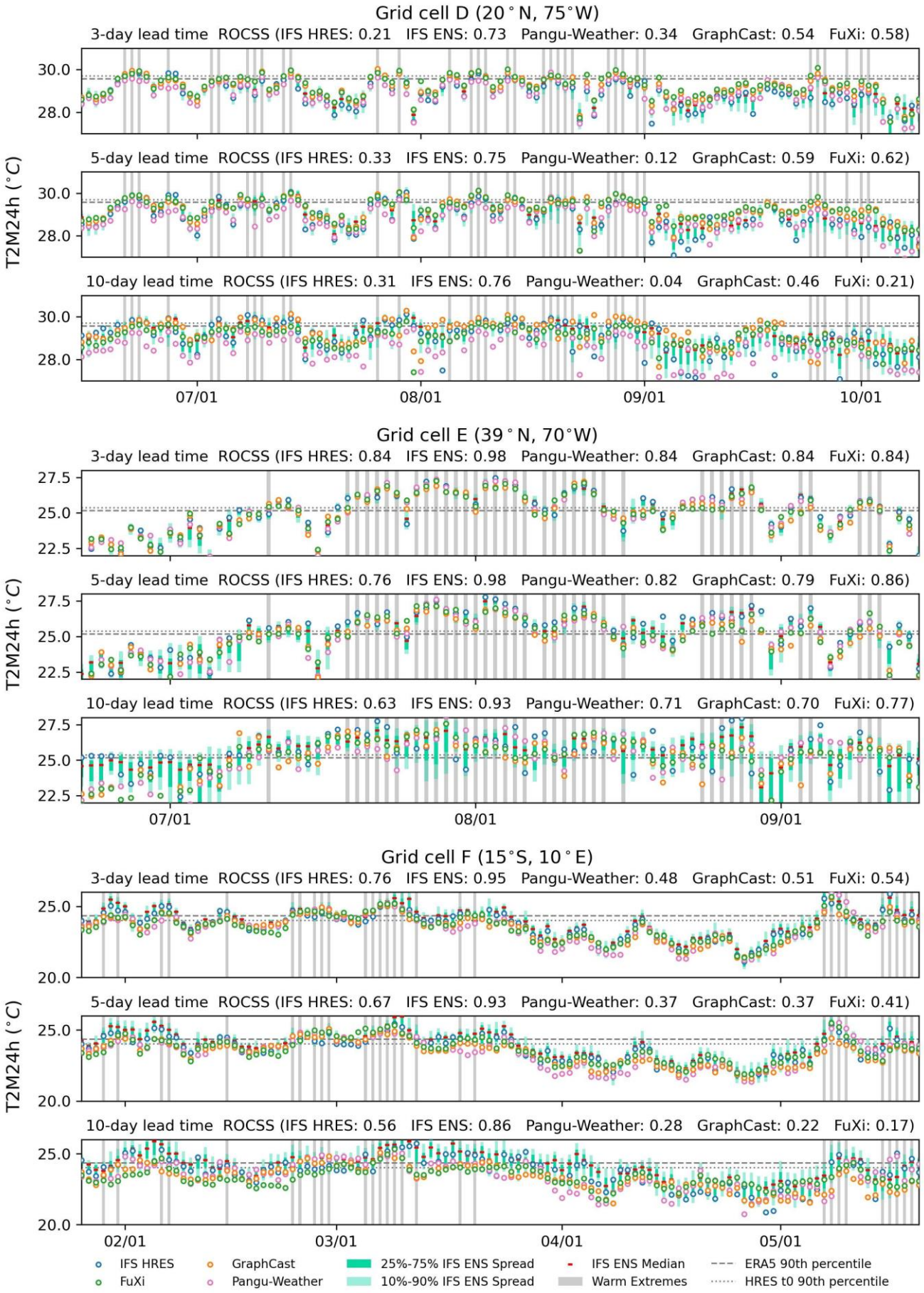

Figure 7. Time series plots of T2M24h forecasts initialized at 00 UTC for the IFS HRES, IFS ENS, Pangu-Weather, GraphCast and FuXi over three selected grid cells, i.e., D (20 °N, 75 °W), E (39 °N, 70 °W) and F (15 °S, 10 °E).

4.4 Sensitivity to predefined thresholds

The globally area-weighted performance under different predefined thresholds is illustrated for 5-day lead time in Figure 8. The ROCSS is base-rate-independent and suitable simultaneously for deterministic and probabilistic forecasts of binary events. It is noted that the REV needs predefined cost-loss ratios to calculate the potential values of forecasts, while the cost-loss ratios may be different for hydroclimatic extremes with different threshold percentiles. In the meantime, the SEDI is also applicable to extreme events because of its base-rate independence and nondegenerate limit (North et al., 2013; Jolliffe and Stephenson, 2012; Brodie et al., 2024). The base-rate-independent 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).

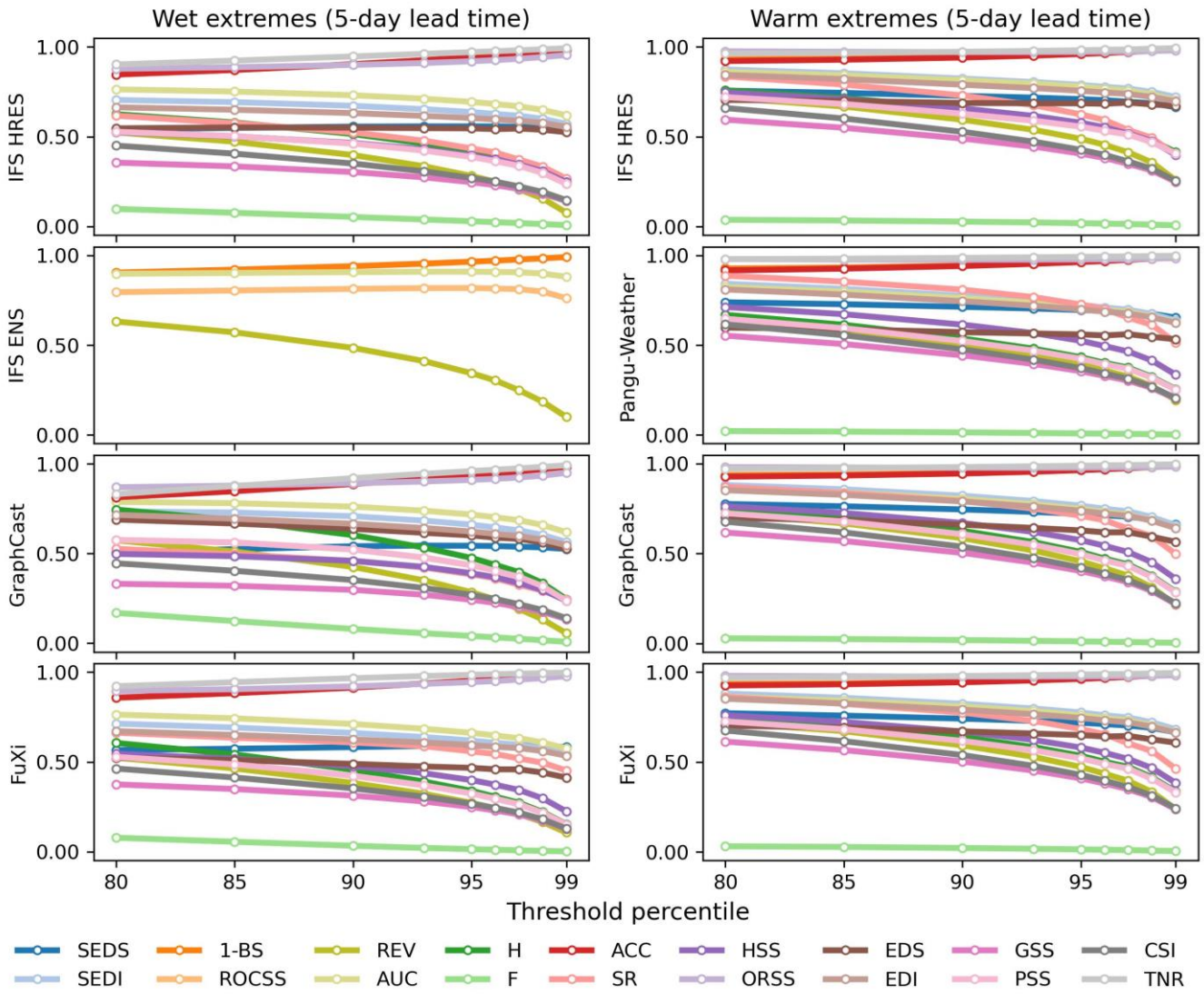

Figure 8. Globally area-weighted performance in forecasting wet extremes and warm extremes with different threshold percentiles at 5-day lead time. The REV is calculated with a fixed cost-loss-ratio of 0.2 only for purposes of illustration.

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 at lead times less than 5 days and the GraphCast tends to be better at lead time more than 5 days. Specifically, as to forecast wet 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 than the Pangu-Weather and Pangu-Weather (operational) in predicting warm extremes (Olivetti and Messori, 2024b).

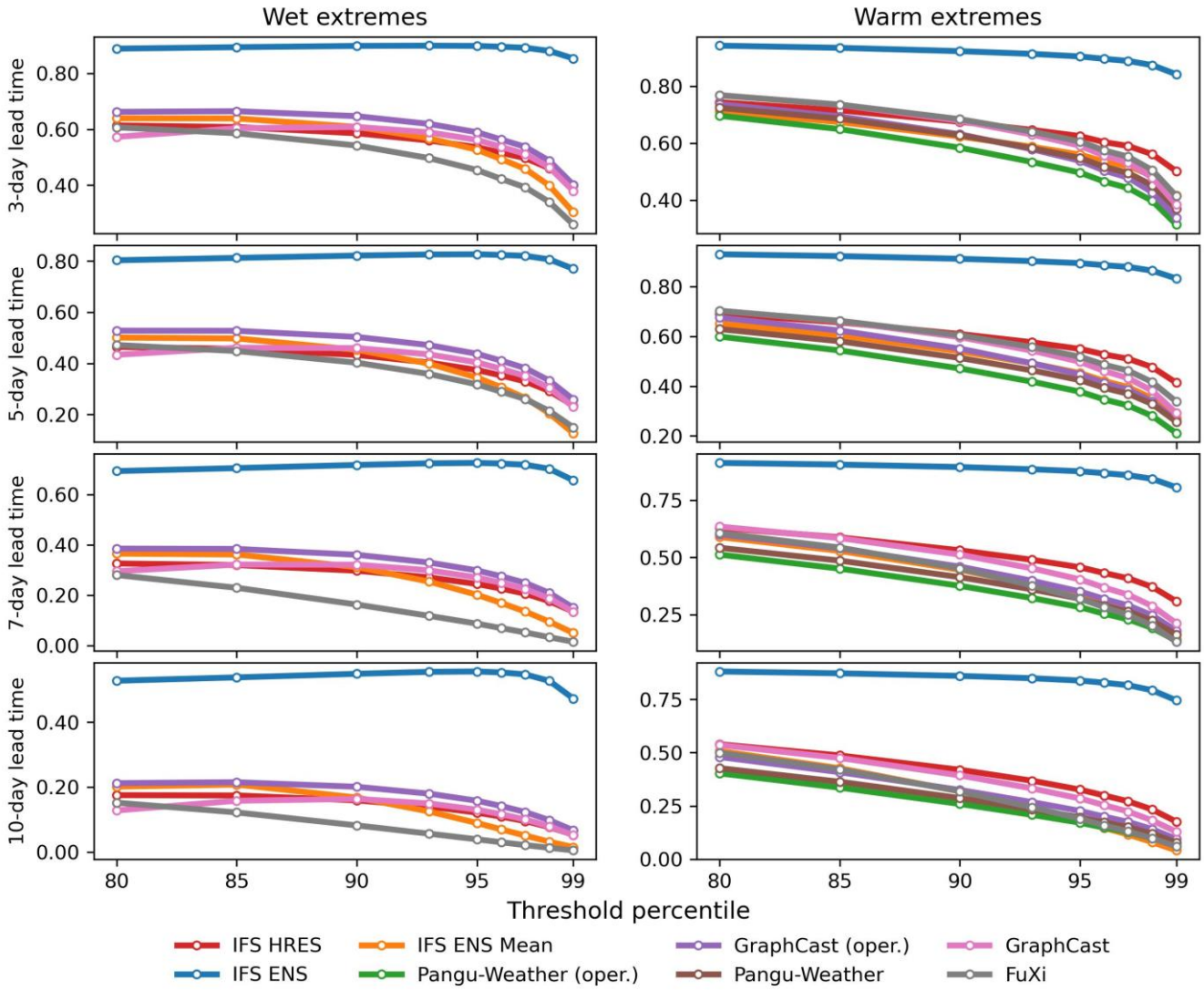

Figure 9. Globally area-weighted ROCSS for wet extremes and warm extremes with different threshold percentiles.

5 Discussion

5.1 Implications on forecaster's dilemma

Binary hydroclimatic forecasts provide useful information for disaster prevention and risk mitigation (Ben Bouallègue et al., 2024; Merz et al., 2020). Verification metrics of deterministic and ensemble forecasts, such as the RMSE and CRPSS, in general focus on the overall predictive performance across a range of events (Huang and Zhao, 2022; Rasp et al., 2024). They tend to reward models that minimize average errors and unrealistically smooth forecasts, leading to limited guidance to forecast hydroclimatic extremes (Ferro and Stephenson, 2011; Rasp et al., 2020). By contrast, verification metrics of binary forecasts provide valuable additional information by emphasizing the ability to discriminate certain hydroclimatic extremes that do not directly relate to average errors (Larraondo et al., 2020). In this paper, the results show that for warm extremes, the Pangu-

Weather, GraphCast and FuXi tend to be more skilful than the IFS HRES within 3-day lead time but become less skilful as lead time increases. The verification of binary hydroclimatic forecasts seems to be more stringent for data-driven models since the observed lead time in which there exists outperformance of data-driven models tends to be shorter than that under continuous forecasts (Lam et al., 2023; Bi et al., 2023; Chen et al., 2023). In the supplement, the results across global grid cells in terms of the HSS and SEDI also support this result.

The climate system is high-dimensional and complex so that there won't be a single verification metric to showcase all essential characteristics of a good forecast (Rasp et al., 2024; Jolliffe and Stephenson, 2012). While verifications metrics of binary forecasts emphasize the discrimination, they are unable to reflect other attributes to quantify the forecast quality, such as reliability, resolution and uncertainty (Murphy, 1993). Although the GraphCast is more capable of capturing the wet extremes, it tends to produce more false positives. This result implies the “forecaster’s dilemma”, i.e., conditioning on outcomes is incompatible with the theoretical assumptions of established forecast evaluation methods (Lerch et al., 2017). From this perspective, a combination of multiple verification metrics and diagnostic plots is in demand (Larraondo et al., 2020; Huang and Zhao, 2022). As shown in Fig. S1 and Fig. S4 in the supplement, the values of BS for the FuXi are better than that for the HRES at the lead time of 10 days, which is different to the results for ROCSS in Fig.4. Considering that the BS tends to reflect the average performance and is influenced by the unbalanced number of occurrences and non-occurrences, better values of a single metric do not mean a more useful forecast (Rasp et al., 2024). Overall, the process of forecast verification needs to be guided by the demand of operational applications (Ben Bouallègue and the AIFS team, 2024; Rasp et al., 2024).

5.2 Use of ground truth data

High-resolution forecasts are essential for accurately capturing multi-scale processes of hydroclimatic extremes (Liu et al., 2024a; Charlton-Perez et al., 2024; Xu et al., 2025). It is noted that hydroclimatic forecasts of coarse spatial resolution tend to miss the required small-scale variability, such as the intensity and structure of typhoon (Ben Bouallègue et al., 2024; Selz and Craig, 2023). Also, they may miss extreme values and the underlying evolution processes due to the mismatch between forecast time step and event time (Pasche et al., 2025). Therefore, there exists a demand to enhance the spatial and temporal resolution of data-driven models (Xu et al., 2024b; Zhong et al., 2024). It is noted that diffusion models have recently been shown to be effective for km-scale atmospheric downscaling (Mardani et al., 2025). In addition, hybrid models that utilize global forecasts from data-driven models to drive high-resolution regional models, such as the weather research and forecasting (WRF) model, can improve the forecast accuracy and resolution for extreme precipitation and tropical cyclones (Liu et al., 2024b; Xu et al., 2024b, 2025). Given that the metrics listed in Table 3 are suitable to different spatial and temporal scales, the WeatherBench 2 is capable of evaluating for high-resolution forecast data.

Part of forecast skill of data-driven models on wet and warm extremes can stem from the unfair setting of ground truth data (Rasp et al., 2024; Lam et al., 2023). As for the WeatherBench 2, it is worthwhile to note that the verification of precipitation using ERA5 reanalysis 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 occurring in 2020 due to current data availability. The short verification period can provide limited information about the model performance and sensitive results to the climate variability (Olivetti and Messori, 2024b). With the availability of more data on hydroclimatic forecasts and baseline ground-truth observations, binary forecasts of hydroclimatic extremes deserve more in-depth verification. In the meantime, the different roles that the operational IFS analysis and ERA5 reanalysis data play in the initial conditions to generate forecasts also deserve further verification (Ben Bouallègue et al., 2024; Liu et al., 2024a; Xu et al., 2024b).

6 Conclusions

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

Code and data availability

The raw data, i.e., forecasts and ground truth data, used in this paper are downloaded from the WeatherBench 2 and are archived on the Zenodo under https://doi.org/10.5281/zenodo.15066828 (Li and Zhao, 2025a) and under https://doi.org/10.5281/zenodo.15066898 (Li and Zhao, 2025b).

The code and scripts performing all the analysis and plots are archived on the Zenodo under https://doi.org/10.5281/zenodo.15067282 (Li and Zhao, 2025c). All the analysis results are archived on the Zenodo under https://doi.org/10.5281/zenodo.15067178 (Li and Zhao, 2025d).

To guarantee future compatibility with the WeatherBench 2, the code and scripts have been made a push request to its successor, i.e., WeatherBench-X.

Author contributions

TZ: Writing – original draft, Visualization, Software, Methodology, Conceptualization. QL: Validation, Resources, Data 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.

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).

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.

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-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 Forecasts, 2020.

Ben Bouallègue, Z. and the AIFS team: Accuracy versus activity, ECMWF, 2024.

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-D-23-0162.1, 2024.

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.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, NATURE, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Bonavita, M.: On Some Limitations of Current Machine Learning Weather Prediction Models, GEOPHYS RES LETT, 51, e2023GL107377, https://doi.org/10.1029/2023GL107377, 2024.

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

Charlton-Perez, A. J., Dacre, H. F., Driscoll, S., Gray, S. L., Harvey, B., Harvey, N. J., Hunt, K. M. R., Lee, R. W., Swaminathan, R., Vandaele, R., and Volonté, A.: Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán, npj Clim Atmos Sci, 7, 1–11, https://doi.org/10.1038/s41612-024-00638-w, 2024.

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, 11,693-11,701, 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.

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.

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,

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.

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.

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, 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: 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, 2nd ed., John Wiley & Sons, 2012.

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

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, 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.

Lang, Y., Ye, A., Gong, W., Miao, C., Di, Z., Xu, J., Liu, Y., Luo, L., and Duan, Q.: Evaluating Skill of Seasonal Precipitation and Temperature Predictions of NCEP CFSv2 Forecasts over 17 Hydroclimatic Regions in China, J HYDROMETEOROL, 15, 1546–1559, https://doi.org/10.1175/JHM-D-13-0208.1, 2014.

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.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T.: Forecaster's Dilemma: Extreme Events and Forecast Evaluation, STAT SCI, 32, 106–127, https://doi.org/10.1214/16-STS588, 2017.

Li, Q. and Zhao, T.: Code for the extension of the WeatherBench 2 to binary hydroclimatic forecasts (v0.3.0), https://doi.org/10.5281/zenodo.15067282, 2025a.

Li, Q. and Zhao, T.: Data for the extension of the WeatherBench 2 to binary hydroclimatic forecasts (v0.2.0),

https://doi.org/10.5281/zenodo.15067178, 2025b.

Li, Q. and Zhao, T.: Data for the extension of the WeatherBench 2 to binary hydroclimatic forecasts: ensemble forecasts for 24h maximum temperature (v0.1.0), https://doi.org/10.5281/zenodo.15066898, 2025c.

Li, Q. and Zhao, T.: Data for the extension of the WeatherBench 2 to binary hydroclimatic forecasts: ensemble forecasts for 24h precipitation (v0.1.0), https://doi.org/10.5281/zenodo.15066828, 2025d.

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.

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, 2024a.

Liu, H., Tan, Z., Wang, Y., Tang, J., Satoh, M., Lei, L., Gu, J., Zhang, Y., Nie, G., and Chen, Q.: A Hybrid Machine Learning/Physics-Based Modeling Framework for 2-Week Extended Prediction of Tropical Cyclones, Journal of Geophysical Research: Machine Learning and Computation, 1, e2024JH000207, https://doi.org/10.1029/2024JH000207, 2024b.

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.

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Nabian, M. A., Ge, T., Subramaniam, A., Kashinath, K., Kautz, J., and Pritchard, M.: Residual corrective diffusion modeling for km-scale atmospheric downscaling, Commun Earth Environ, 6, 1–10, https://doi.org/10.1038/s43247-025-02042-5, 2025.

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, Rev. Geophys., 58, e2020RG000704, https://doi.org/10.1029/2020RG000704, 2020.

Murphy, A. H.: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, WEATHER FORECAST, 8, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, 1993.

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, 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-17-7915-2024, 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.

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., and Engelke, S.: Validating Deep Learning Weather Forecast Models on 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 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, 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, 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.: 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.

Richardson, D. S.: Skill and relative economic value of the ECMWF ensemble prediction system, Q J ROY METEOR SOC, 126, 649–667, https://doi.org/10.1002/qj.49712656313, 2000.

Richardson, D. S.: Predictability and economic value, in: Predictability of Weather and Climate, edited by: Palmer, T. and Hagedorn, R., Cambridge University Press, 628–644, 2006.

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.

Selz, T. and Craig, G. C.: Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?, GEOPHYS RES LETT, 50, e2023GL105747, https://doi.org/10.1029/2023GL105747, 2023.

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

Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R. L., Wood, A., and Zappa, M.: Hybrid forecasting: blending climate predictions with AI models, Hydrol Earth Syst Sc, 27, 1865–1889, https://doi.org/10.5194/hess-27-1865-2023, 2023.

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, 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, https://doi.org/10.1029/2020MS002109, 2020.

Wilks, D. S.: A skill score based on economic value for probability forecasts, METEOROL APPL, 8, 209–219, https://doi.org/10.1017/S1350482701002092, 2001.

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.

Xiong, S., Zhao, T., Guo, C., Tian, Y., Yang, F., Chen, W., and Chen, X.: Evaluation and attribution of trends in compound dry-hot events for major river basins in China, Sci. China Earth Sci., 67, 79–91, https://doi.org/10.1007/s11430-022-1174-7, 2024.

Xu, H., Duan, Y., and Xu, X.: Evaluating AI's capability to reflect physical mechanisms: a case study of tropical cyclone 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.

Xu, H., Zhao, Y., Dajun, Z., Duan, Y., Xu, X., Xu, H., Zhao, Y., Dajun, Z., Duan, Y., and Xu, X.: Exploring the typhoon intensity forecasting through integrating AI weather forecasting with regional numerical weather model, npj Clim Atmos Sci, 8, 1–10, https://doi.org/10.1038/s41612-025-00926-z, 2025.

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

Zhao, T., Xiong, S., Tian, Y., Wu, Y., Li, B., and Chen, X.: Compound dry and hot events over major river basins of the world from 1921 to 2020, WEATHER CLIM EXTREME, 44, 100679, https://doi.org/10.1016/j.wace.2024.100679, 2024.

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.