Responses:

Executive Editor:

In your Code and Data Availability section, with only two exceptions, you link your code and data with websites that do not comply with the standards to be scientific repositories. These include things like Google drives, Google docs, papers that for the availability include links to GitHub sites (something totally unacceptable too, as our policy clearly states), and links to main Copernicus portals, instead to provide the access to the exact data that you use. Therefore, all these issues listed fail to comply with our policy, and instead the current sites that you link or citations to papers, you must provide the exact repositories (link and DOI) from one of the ones accepted according to our policy (please, check it).

Thank you very much for clearly pointing to the "Code and Data Policy" of Geoscientific Model Development. After carefully reading the policy and accordingly checking the manuscript, we have deleted all the website links that do not comply with the standards to be scientific repositories. We have archived all the code and data on the Zenodo repositories:

“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 WeatherBench 2, the code and scripts have been made a push request to its successor, i.e., WeatherBench-X.” (Page 25, Lines 410 to 418)

A big issue here is that manuscripts that do not comply with our Code and Data policy can not be accepted for Discussions and review. However, your manuscript has ended here because of editorial overlook. Because of it, we are granting you a brief time to solve this situation, and reply to this comment with the requested information, which then you should include in any potentially reviewed version of your manuscript, modifying the Code and Data Availability section accordingly.

We are grateful to you for the kind reminding and granting us the time to address this issue. The code and data are now on the Zenodo repositories:

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, 2025a.

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, 2025b.

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, 2025c.

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, 2025d.

Please, reply to this comment addressing this situation and with the information for the repositories as soon as possible. I have to note that failing to comply with this request will make us to have to reject your manuscript for publication in our journal.

Thank you very much for giving us an opportunity to improve the manuscript. By following the insightful comments and the "Code and Data Policy", we have thoroughly revised the code and data availability to address the issue and provide data repositories.

Anonymous Referee #1:

I find the idea of extending WeatherBench2 to climatological extremes very interesting. I believe such evaluations would be really valuable for further comparison between numerical and machine learning based weather models.

We are grateful to you for the positive comments.

However, I personally find the paper more of a summary of possible evaluation metrics demonstrated on WeatherBench2 rather than actual extension of WeatherBench2. The provided codebase and data seem to allow for reproducibility of the results in the paper, but I am not convinced they are easy to reuse for evaluation of new approaches on extreme weather events in the future. Upon some major revisions, I can see this as a valuable contribution to the WeatherBench2 benchmark. I have the following detailed comments I would like to be answered and addressed:

Thank you very much for the constructive comments. This paper aims to exploit more useful comparisons from hydroclimatic forecasts by extending the WeatherBench 2 to binary hydroclimatic events. To this end, this paper illustrates in total seventeen verification metrics on binary forecasts, presents scorecards to showcase the predictive performance on wet and warm extremes and finally examines the sensitivity of different metrics to predefined thresholds of hydroclimatic extremes. Recently, we have made the evaluation scripts a push request to the WeatherBench-X repository.

We agree on that the current code may not be quite reusable. In the revision, the plug-and-play code is provided. In the meantime, the code and scripts are also provided in the form of Jupyter notebooks to facilitate learning and reproducing the entire procedures of forecasts verification.

1) I think what you propose would be a valuable extension of WeatherBench2. However, what you provide is a set of Jupyter Notebooks. I think it would be better to provide a small library with a set of methods that can be imported and run as part of any other codebase. Would that be feasible?

Thank you for the valuable comment. We agree on that it is better to provide a code library:

“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 WeatherBench 2, the code and scripts have been made a push request to its successor, i.e., WeatherBench-X." (Page 25, Lines 410 to 418)

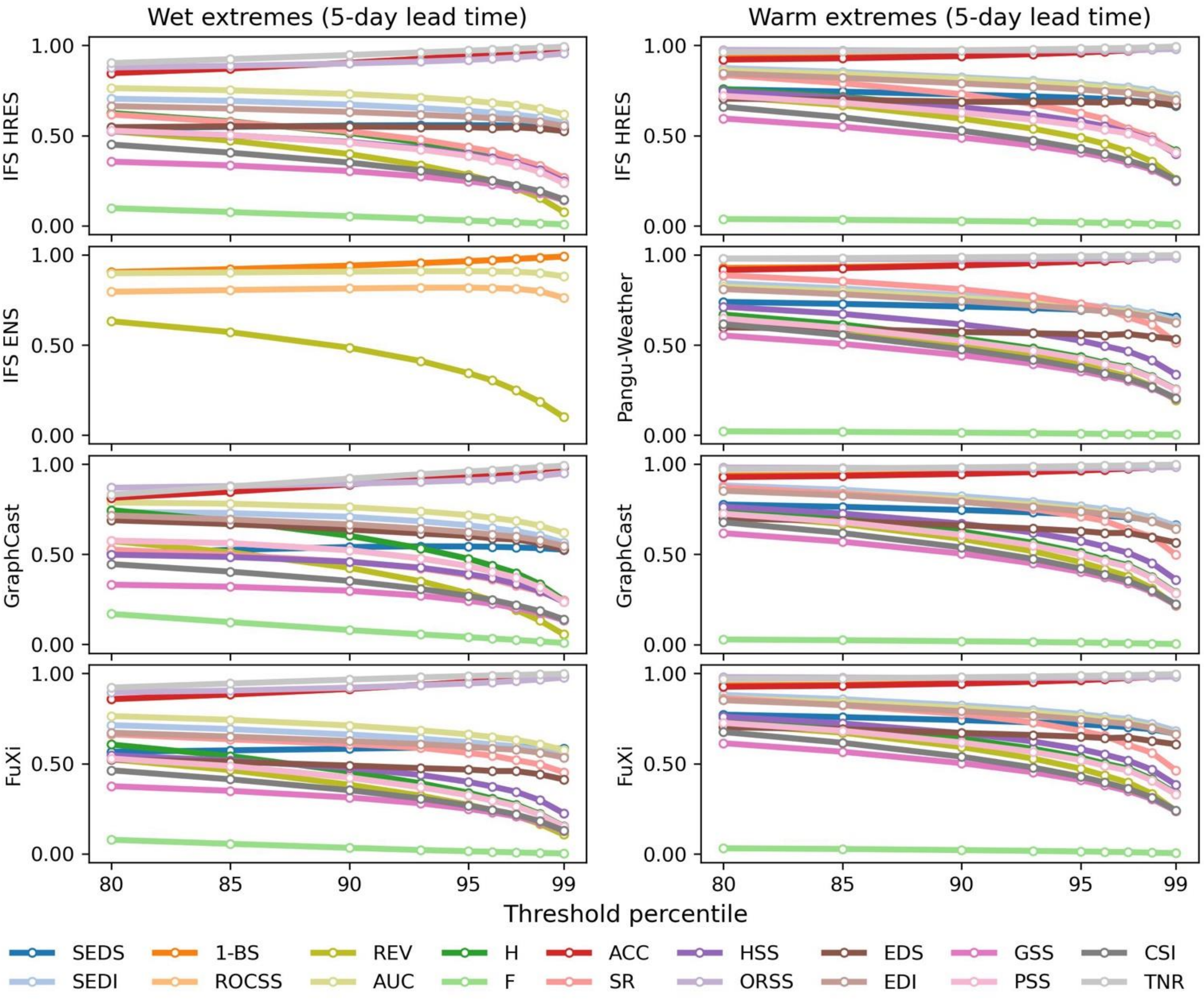
2) Besides, how do you guarantee future compatibility with WeatherBench2, may it undergo any substantial changes in terms of available datasets or codebase? Would it be possible to make your evaluation scripts a push request to WeatherBench2 repository, making it part of the benchmark?

Thank you very much for the instructive comment. We have already made the evaluation scripts a push request (https://github.com/google-research/weatherbenchX/pull/36) to the WeatherBench-X repository. This repository is the successor to the WeatherBench 2 evaluation code and provides the evaluation framework that enables flexible evaluation of various kinds of forecast and ground truth data.

3) You mention GenCast (Price et al., 2025) in your paper, which suggests alternative evaluation metric for weather extremes borrowed from finance, in particular REV curves. I think it would be an interesting score to include in your evaluation. Could you comment on why is it not included and eventually include it?

Thank you for the insightful comment. The relative economic value (REV) is indeed a useful metric that allows us to assess the value of a forecast to a range of users facing different decision problems (Price et al., 2025). Accordingly, we have added the REV to Section 3.2 and updated Figure 8 to illustrate this metric:

"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)." (Pages 7 and 8, Lines 155 to 158)



“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.” (Page 21, Lines 325 to 327)

4) In Section 3.1 you introduce terminology "hits", "false alarm", "misses", "correct rejections". Is this the jargon in natural hazards? I would rather suggest the following terminology: "true positives", "false positives", "false negatives", "true negatives" respectively. This should be changed also anywhere else in the text where the terminology is used.

Thank you for the instructive comment. Following Jolliffe & Stephenson (2012), we used the terminology "hits", "false alarm", "misses" and "correct rejections". For the purpose of easy understanding, we have used "true positives", "false positives", "false negatives" and "true negatives" in the revision:

“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.” (Page 5, Lines 118 to 122)

5) The equations in Table 3 are not understandable without preliminary knowledge. Each equation contains variables, which are not explained anywhere in the text. This, in my opinion, needs to be improved together with description of each score in the text, which is at the moment rather vague (see also my next comment).

We are sorry for the confusion. We have double checked Table 3 for the names of variables across different verification metrics. The variables are explained in the footnotes of Table 2 and Table 3; in the meantime, the description of each score in the text has been improved:

“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 L ópez et al., 2010)
Potential relative economic value (REV)	$REV = \max_{0\leq p\leq 1} \frac{\min\{a+c,r\} - [(a+b)r+c]}{\min\{a+c,r\} - (a+c)r}$	[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.” (Pages 6 and 7, Lines 140 to 144)

“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." (Pages 7 and 8, Lines 146 to 170)

6) On line 153 you introduce HSS, and later on line 161 you introduce ORSS. You describe them as: HSS - accuracy relative to that of the random forecast; ORSS - improvement over the random forecast. It sounds like the two metrics are redundant. Is that the case? If so, why do we need both?

Thank you. In the revision, the introduction of ORSS and HSS are improved to highlight their difference:

"The Heidke skill score (HSS) measures the accuracy relative to that of the random forecasts (Gomis-Cebolla et al., 2023)." (Page 7, Lines 152 and 153)

"The odds ratio skill score (ORSS) examines the improvement over the random forecasts, emphasizing the balance between positive and negative samples (Stephenson, 2000)." (Page 8, Lines 162 to 164)

7) Line 171, I would suggest to reformulate the sentence.

Thank you very much for the instructive comment. We have reformulated this sentence:

“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.” (Page 8, Lines 178 and 179)

8) Line 192 - "As expected, forecasts become less accurate" - why is this expected? You haven't motivated anywhere in the text why this should be the case, neither reference any literature that would explain it. It is mostly the case that forecasts for longer lead times exhibit strong decrease in performance, but I believe your expectation should be somehow grounded.

We are sorry for the incomplete information. The explanation of the expected results has been added:

“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).” (Page 9, Lines 201 to 203)

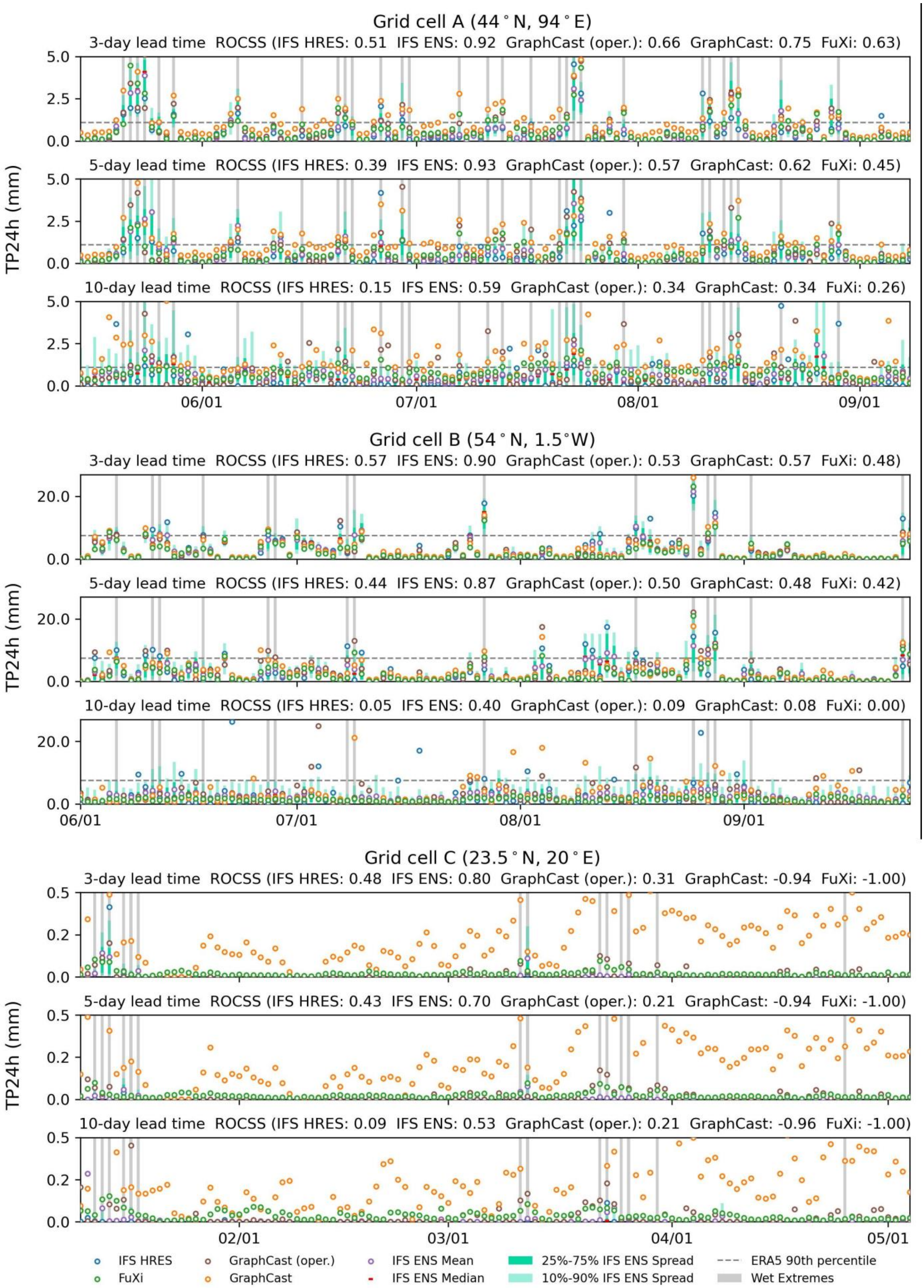
9) Line 197 - "As lead time increases, data-driven forecasts can be less skilful than the IFS HRES". This is an interesting observation, without any follow-up argumentation. It would be great to have more insights into this.

Thank you for the insightful comment. We have added the possible explanation of this phenomenon:

“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).” (Page 9, Lines 207 to 209)

10) Figure 5 legend - gray shading says "Warm Extremes". Is that correct or should it be "Wet extremes" since there is precipitation on the y-axis? And the shaded areas are where precipitation is often high.

Thank you for spotting the typo. We have corrected it accordingly:



“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)." (Page 16, Lines 276 to 278)

11) Depending on the extremes that one wants to forecast, the forecast resolution plays an important role. Many extreme precipitation events cannot be detected with models operating on a coarse global grid as they are subject to very local behaviors. Would it be worth extending WeatherBench2 with new datasets that would allow for finer-scale evaluation of some local extreme events? Do you have any insights on which numerical and machine learning models could be used in order to produce more fine grained forecasts?

Thank you for the constructive comment. We have added a new paragraph to the Discussion section:

"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." (Page 23, Lines 371 to 381)

12) Line 347: "the capability to produce binary forecasts of hydroclimatic extremes warrants further verification" - I would suggest rephrasing

Thank you for the insightful comment. This sentence has been revised as:

"With the availability of more data on hydroclimatic forecasts and baseline ground-truth observations, binary forecasts of hydroclimatic extremes deserve more in-depth verification." (Page 24, Lines 389 and

390)

13) Line 357: "total precipitation of ERA5 data is used as the ground truth" - Do you mean ERA5 forecast or reanalysis? I believe only reanalysis data would make sense as a ground truth.

We are sorry for the confusion. The ground truth data used in this paper is the ERA5 reanalysis rather than ERA5 forecasts. We have revised this sentence to emphasize this point:

"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." (Page 24, Lines 398 to 400)

I do not think it would be objective to compare to ERA5 precipitation forecast directly as we do not want to match ERA5 forecasting capability, but hopefully improve over it, therefore needing ground-truth data corresponding to reality.

Thank you very much. We fully agree on this point:

"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." (Page 24, Lines 382 to 386)

"With the availability of more data on hydroclimatic forecasts and baseline ground-truth observations, binary forecasts of hydroclimatic extremes deserve more in-depth verification." (Page 24, Lines 389 and 390)

References:

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

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. Quarterly Journal of the Royal Meteorological Society, 126(563), 649–667. https://doi.org/10.1002/qj.49712656313

Richardson, David S. (2006). Predictability and economic value. In T. Palmer & R. Hagedorn (Eds.), Predictability of Weather and Climate (1st ed., pp. 628–644). Cambridge University Press. https://doi.org/10.1017/CBO9780511617652.026

Wilks, D. S. (2001). A skill score based on economic value for probability forecasts. Meteorological Applications, 8(2), 209–219. https://doi.org/10.1017/S1350482701002092

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., et al. (2025). Probabilistic weather forecasting with machine learning. Nature, 637(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Bonavita, M. (2024). On Some Limitations of Current Machine Learning Weather Prediction Models. Geophysical Research Letters, 51(12), e2023GL107377. https://doi.org/10.1029/2023GL107377

Anonymous Referee #2:

This manuscript evaluates the performance of leading deterministic deep learning models (GraphCast, PanguWeather, and FuXi) against ECMWF's numerical models, IFS HRES and IFS ENS, in forecasting hydroclimatic extremes. By employing a comprehensive set of skill scores, it provides a detailed assessment of forecast quality. Notably, the study shifts the focus from continuous metrics to binary extremes, expanding upon the work of WeatherBench 2 and offering fresh insights for operational forecasting. In doing so, it also underscores the importance of binary decision-making in high-stakes forecasting scenarios.

We are grateful to you for the positive comments.

I believe this manuscript presents a valuable argument and could make a strong contribution to the evaluation literature on data-driven weather models. However, several key concerns must be addressed before it can be recommended for publication:

Thank you very much for the insightful and detailed comments. Accordingly, we have revised this paper. Below please find the point-to-point responses.

1. A clearer justification and explanation of key methodological choices.

Thank you for the valuable comment. The key methodological choices have been explained more clearly in the revision:

"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)." (Page 5, Lines 107 to 109)

"It is noted that the thresholds at each grid cell are separately calculated." (Page 5, Line 113)

"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)." (Page 8, Lines 182 to 184)

"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)."

(Page 9, Lines 189 to 191)

"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." (Page 15, Lines 265 and 266)

2. Visualizations that provide regional or grid-point-level insights for the main evaluation metrics.

Thank you very much for the instructive comment. We have prepared a new supplement file to provide the grid-point-level results of the Brier score (BS), Heidke skill score (HSS) and the symmetric extremal dependence index (SEDI):

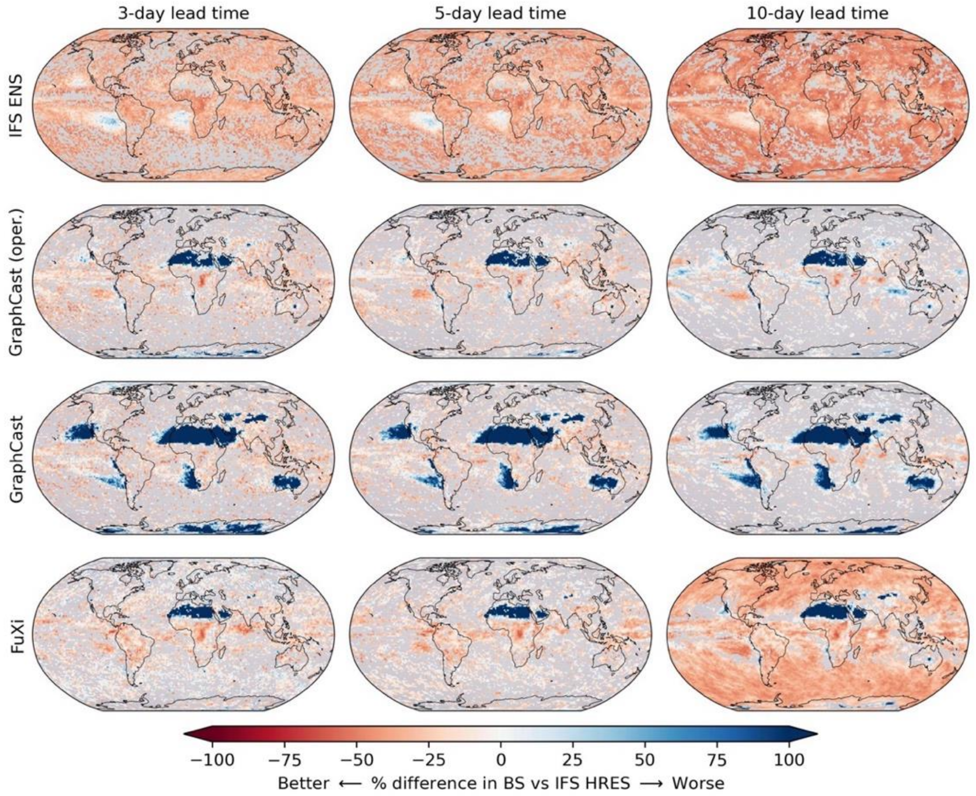


Figure S1. As for Figure 4, but for Brier score (BS).

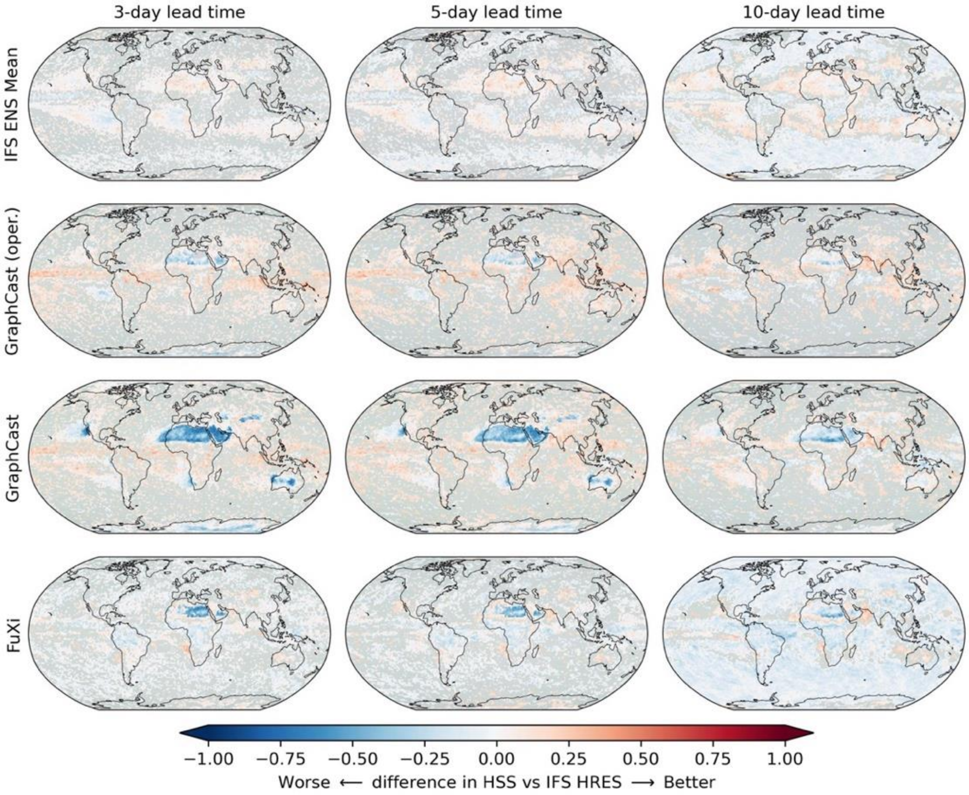


Figure S2. As for Figure 4, but for the Heidke skill score (HSS).

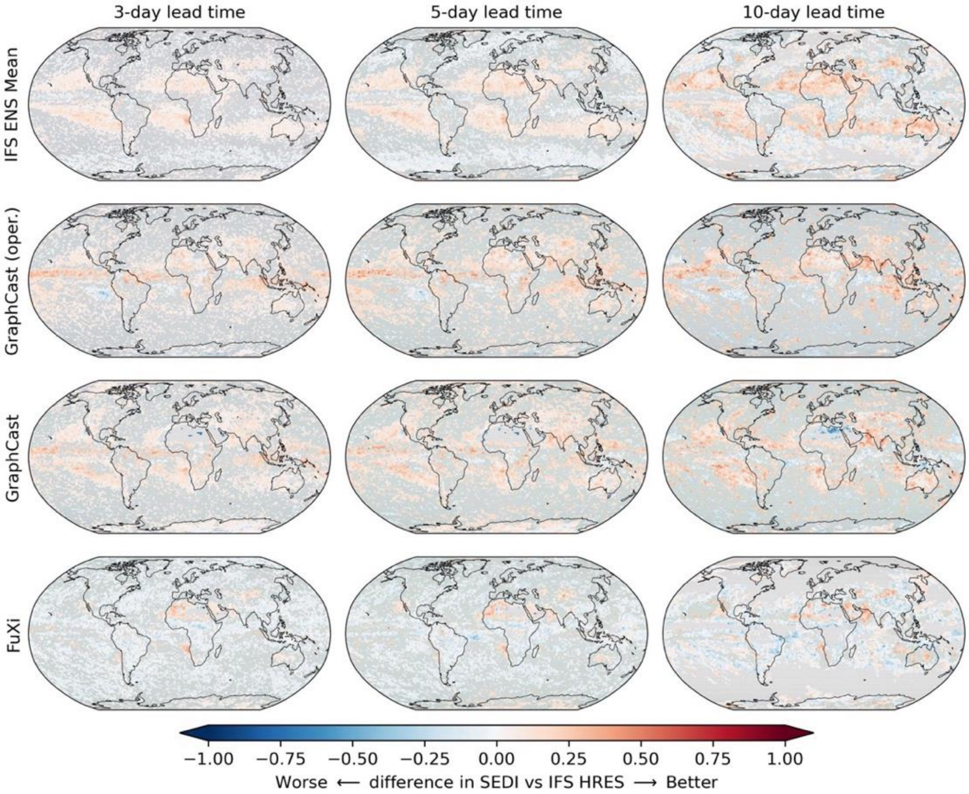


Figure S3. As for Figure 4, but for the symmetric extremal dependence index (SEDI).

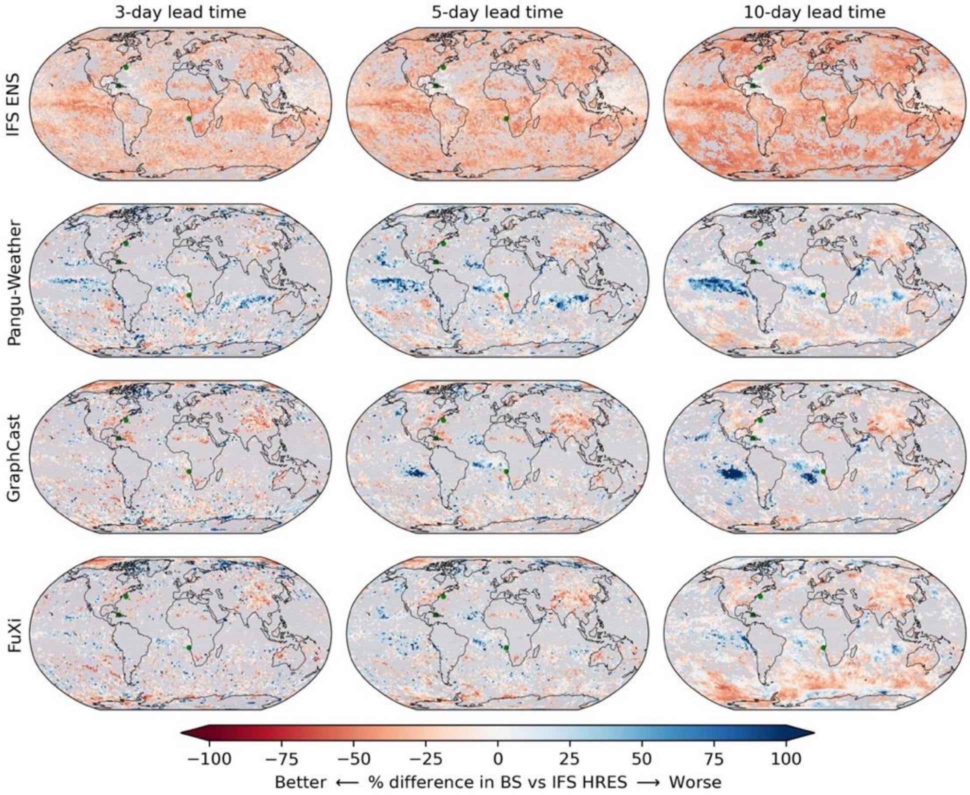


Figure S4. As for Figure 6, but for Brier score (BS).

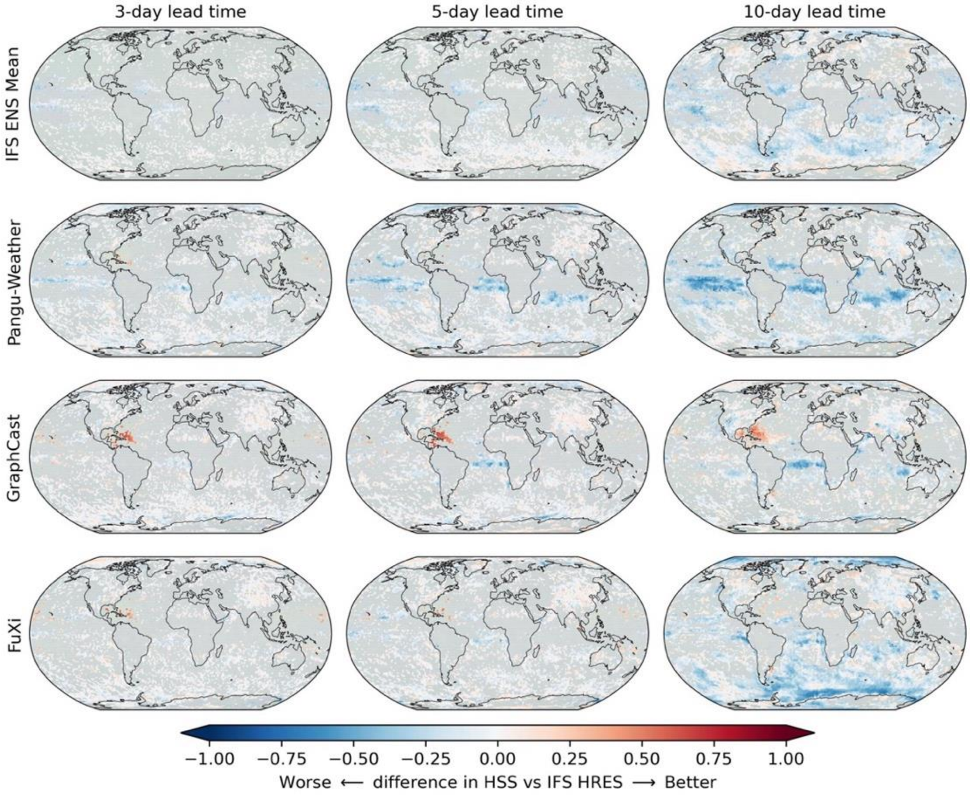


Figure S5. As for Figure 6, but for the Heidke skill score (HSS).

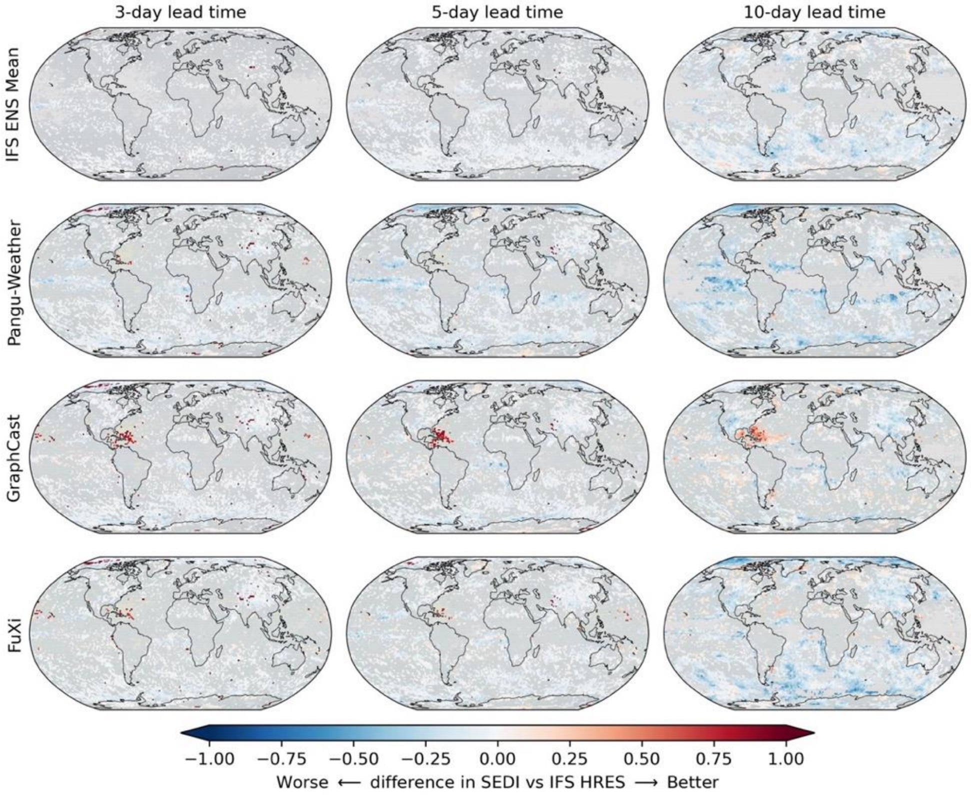


Figure S6. As for Figure 6, but for the symmetric extremal dependence index (SEDI).

” (Pages 1 to 6, Lines 1 to 17 in the supplement)

3. A more comprehensive discussion of the implications of the findings, including potential limitations that may impact the validity of the conclusions.

Thank you for the constructive comment. We have improved the Discussion to elaborate on the implications and limitations:

“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)." (Pages 22 to 24, Lines 343 to 392)

Below, I outline my specific concerns and recommendations:

Thank you very much for the instructive comments. We have improved the paper accordingly and provide the point-by-point responses below.

Main concerns

- I find unclear how global and regional scores are being computed. Are you pooling all data points, defining a single percentile threshold globally, and then taking all data points above that threshold while applying cosine weighting? Or are you defining grid-point-level thresholds, selecting an equal number of extreme data points at each grid point, and then computing an overall score via a cosine-weighted average of individual grid-point scores? Please clarify this in the manuscript. If you are following the

first approach, the global scores may not be particularly meaningful, as they would be dominated by data from the warmest/wettest grid points. If so, it would be especially important to provide additional regional or grid-point-level analyses.

We are sorry for the confusion. Due to that a reasonably rare hydroclimatic event in one location might be common or never observed in another location, it is difficult and not particularly meaningful to define a single threshold at which an event is considered to be extreme (North et al., 2013). Therefore, we calculated the thresholds separately for each grid cell. The following information has been added into the revision to clarify this point:

"It is noted that the thresholds at each grid cell are separately calculated." (Page 5, Line 113)

"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)." (Page 8, Lines 182 to 184)

- Several symbols and abbreviations in Table 3 are undefined, making it difficult to understand the scores without prior knowledge. Please define all terms explicitly and consider adding a short section introducing the main evaluation metrics used in the study.

We are sorry for the incomplete information. Table 3 has been checked and modified carefully. The names of variables here are consistent between different metrics and between Table 2 and Table 3. The variables are explained in the footnotes of Table 2 and Table 3. A new paragraph has been added to introduce the main evaluation metrics. The detailed modifications are as follows:

"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)	$EDS = \dfrac{\ln(a+c)/N - \ln H}{\ln(a+c)/N + \ln H}$	[−1, 1]	1	(Primo and Ghelli, 2009; Stephenson et al., 2008)
Symmetric extreme dependence score (SEDS)	$SEDS = \dfrac{\ln(a+b)/N - \ln H}{\ln(a+c)/N + \ln H}$	[−1, 1]	1	(Orozco López et al., 2010)
Potential relative economic value (REV)	$REV = \max\limits_{0\le p\le 1} \dfrac{\min\{a+c, r\} - [(a+b)r + c]}{\min\{a+c, r\} - (a+c)r}$	[0, 1]	1	(Richardson, 2006, 2000; Wilks, 2001)
Base-rate-independent metrics				
Hit rate (H), sensitivity, recall, probability of detection	$H = \dfrac{a}{a+c}$	[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)	$AUC = \int_0^1 HdF$	[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." (Pages 6 and 7, Lines 140 to 144)

"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." (Page 8, Lines 171 to 175)

- The rationale for selecting specific case studies in Figures 5 and 7 is unclear. Were these grid points chosen because they represent some particular extreme events? Do they highlight specific forecast behavior? If neither, consider moving these figures to the appendix.

We are sorry for the confusion. In the revision, the case studies of grid points are selected respectively due to the better, closer and worse performance of data-driven models in relative to the IFS HRES:

“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.

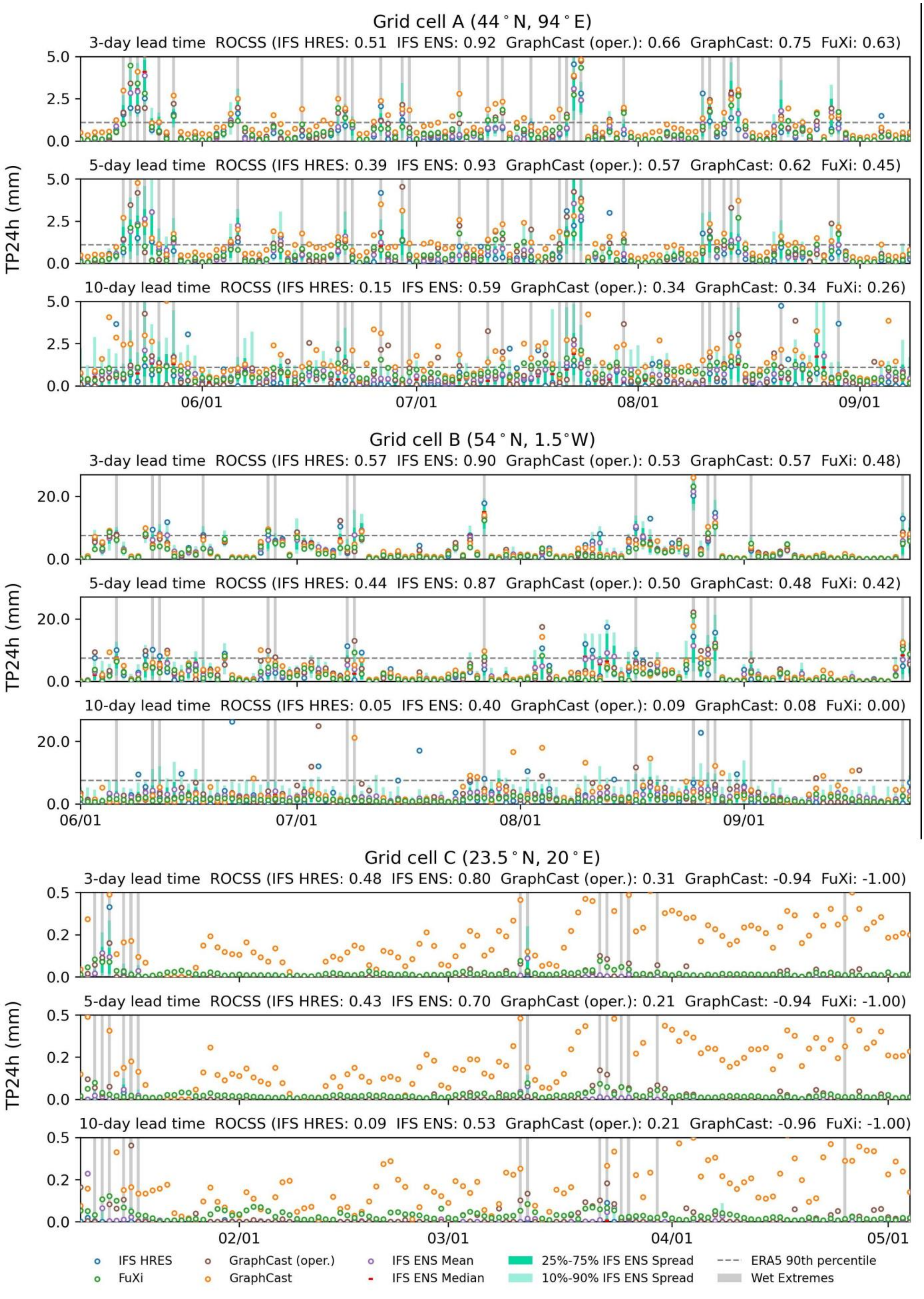


Figure 5. Time series plots of TP24h forecasts initialized at 00 UTC for the IFS HRES, IFS ENS, IFS ENS Mean,

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).” (Pages 14 to 16, Lines 264 to 278)

“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.

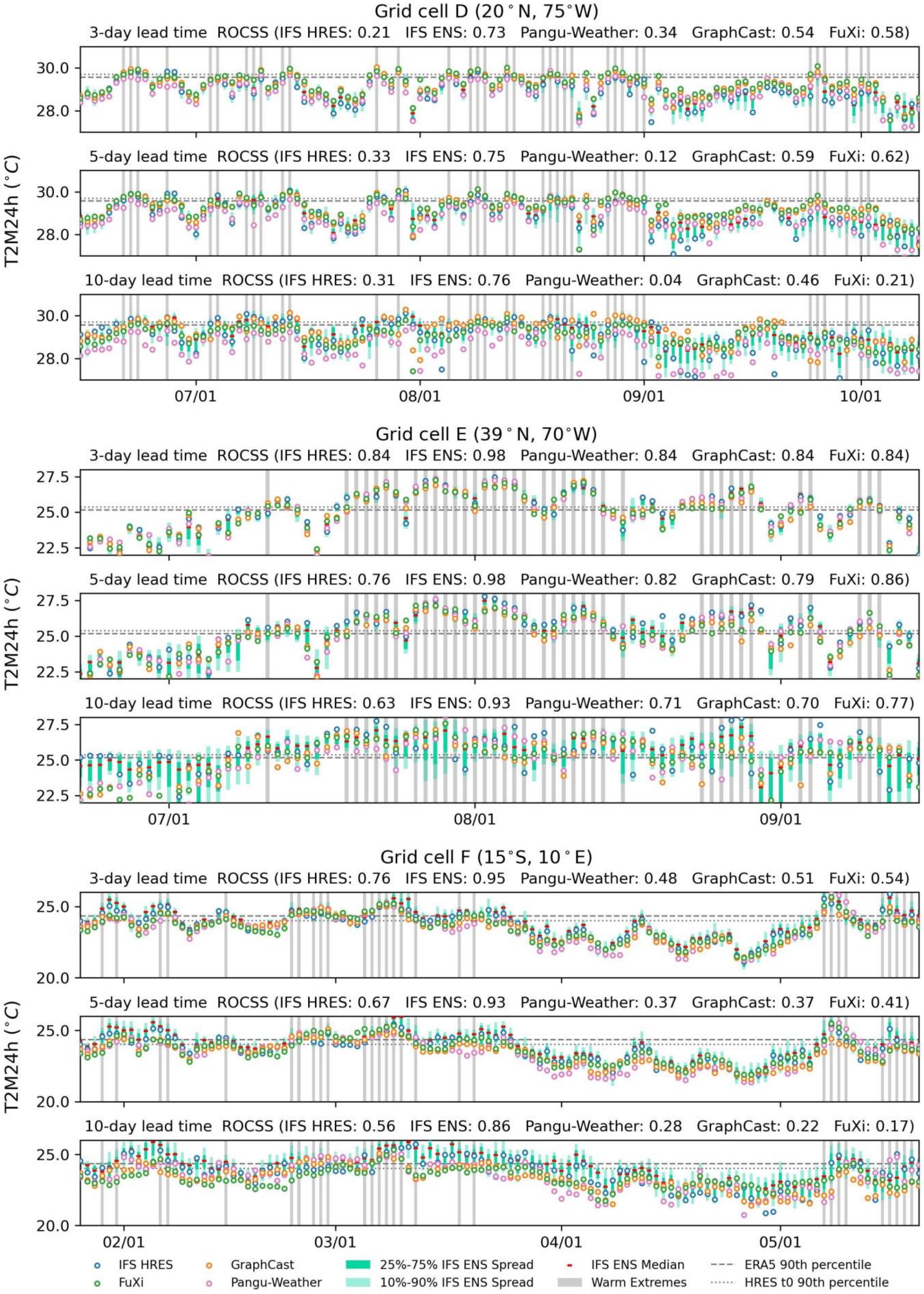


Figure 7. Time series plots of T2M24h forecasts initialized at 00 UTC for the IFS HRES, IFS ENS, Pangu-Weather,

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).” (Pages 18 and 19, Lines 296 to 310)

- Figure 4 is highly informative, as it provides a grid-point-level comparison of different models using a specific skill metric. Could similar visualizations be provided for additional metrics? At present, most metrics are only analyzed at a global scale, which, while interesting, does not offer any insights into regional model performance.

Thank you very much for the instructive comment. We have prepared a supplement file to provide the grid-point-level results of the Brier score (BS), Heidke skill score (HSS) and the symmetric extremal dependence index (SEDI):

“In the supplement, the results across global grid cells in terms of the HSS and SEDI also support this result.” (Page 23, Lines 355 and 356)

“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).” (Page 23, Lines 364 to 367)

- The manuscript presents well-justified points of comparison in the discussion, but it would benefit from a clearer articulation of its novelty relative to prior literature. How do these results improve our understanding of the strengths and limitations of data-driven models compared to numerical models? What new insights does this study provide for operational forecasting? How do these findings extend beyond previous evaluation studies?

Thank you for the valuable comment. The first paragraph of the Discussion has been rewritten:

“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." (Pages 22 and 23, Lines 345 to 356)

- While the manuscript discusses some limitations of individual metrics, a broader reflection on the general limitations of evaluating forecasts based solely on binary performance for hydroclimatic extremes would be valuable. For example, it would be useful to acknowledge that binary metrics alone may not fully capture all the qualities of a good forecast, and could also benefit from integration with standard skill metrics to mitigate the risk for the "forecaster's dilemma" (Lerch, 2017). Additionally, it might be worth discussing why certain models perform particularly well at specific lead times, potentially due to trade-offs between accuracy and forecast activity (Ben Bouallègue and the AIFS team, 2024).

Thank you very much for the constructive comment. We have added a new paragraph into the Discussion to account for these points:

"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).” (Page 23, Lines 357 to 368)

Minor considerations and typos

- The introduction could further emphasise the necessity of this work. Expanding on why binary forecasts are operationally important and how they complement deterministic or probabilistic forecasts would benefit readers unfamiliar with operational forecasting.

Thank you very much for the instructive comment. We have added the information below into the last paragraph of the Introduction:

“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).” (Pages 2 and 3, Lines 54 to 58)

- The manuscript focuses on temperature and wet extremes but does not explain why other extreme variables (e.g., wind) were excluded. A brief justification would be helpful, particularly given that ERA5 may be a more reliable ground truth for other variables than precipitation.

Thank you for the valuable comment. The reason for focusing on the temperature and wet extremes has been added into the revision:

“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).” (Page 5, Lines 107 to 109)

- The choice of approach to statistical testing could do with some more thorough justification. The manuscript currently suggests that the approach follows prior literature, but is a paired t-test valid in this case? For instance, have you verified that the assumption of normality for ROCSS score differences holds? A brief explanation of why this method is appropriate would strengthen the argument for this approach.

We are sorry for the incomplete information. The reason why this method is appropriate has been added into the revision:

"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)." (Page 9, Lines 189 to 191)

- Section 2.1: For clarity and conciseness, consider removing descriptions of models not used in the analysis.

Thank you. We have removed the descriptions of the ERA5 forecasts, Keisler's graph neural network, Esteves's spherical convolutional neural network and NeuralGCM that are not used in the analysis.

- Line 108: Typo ("Weatherbence 2" should be "WeatherBench 2").

Thank you for spotting the typo. We have corrected it.

References

Lerch, Sebastian, et al. "Forecaster's Dilemma: Extreme Events and Forecast Evaluation." Statistical Science, vol. 32, no. 1, 2017, pp. 106–27. JSTOR, http://www.jstor.org/stable/26408123. Accessed 24 Mar. 2025.

Ben Bouallègue and the AIFS team, "Accuracy versus Activity", 2024, doi: 10.21957/8b50609a0f.

References:

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.

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.

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.

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.

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.

Jolliffe, I. T. and Stephenson, D. B.: Forecast verification: a practitioner’s guide in atmospheric science, 2nd ed., John Wiley & Sons, 2012.

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.

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.

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.: 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, 2024.

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.

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.

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.

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, 2024.

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.

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.