Responses:

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 the paper. Below please find the point-to-point responses.

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

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

"It is noted that the thresholds are calculated separately for each grid cell to obtain an equal number of extreme samples per grid cell." (Page 5, Lines 111 and 112)

"For comparison at the grid scale, the 17 metrics are computed separately for each grid cell. To facilitate comparisons at regional to global scales, the 17 metrics are calculated using the area-weighting method based on the scores that are separately calculated for each grid cell (Rasp et al., 2024)." (Page 8, Lines 179 to 181)

"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 264 and 265)

"As the critical variables to understand and forecast the hydroclimatic processes such as floods and heatwaves, the precipitation and temperature are of concern in operational forecasts (Huang and Zhao, 2022; Lang et al., 2014; Zhao et al., 2022; Slater et al., 2023)." (Page 5, Lines 105 to 108)

"Given the skewness and censoring characteristics of hydroclimatic variables, the standard paired t test is not applicable in this case (Huang et al., 2023, 2022). In the meantime, the selected hydroclimatic observations are subject to the problems of heteroscedasticity and autocorrelation due to the spatial and temporal clustering of hydroclimatic extremes (Olivetti and Messori, 2024). To address these issues, the cluster-robust standard errors are used to correct the paired t test (Liang and Zeger, 1986; Shen et al., 1987)." (Page 9, Lines 186 to 190)

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

Thank you very much for the instructive comments. 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):

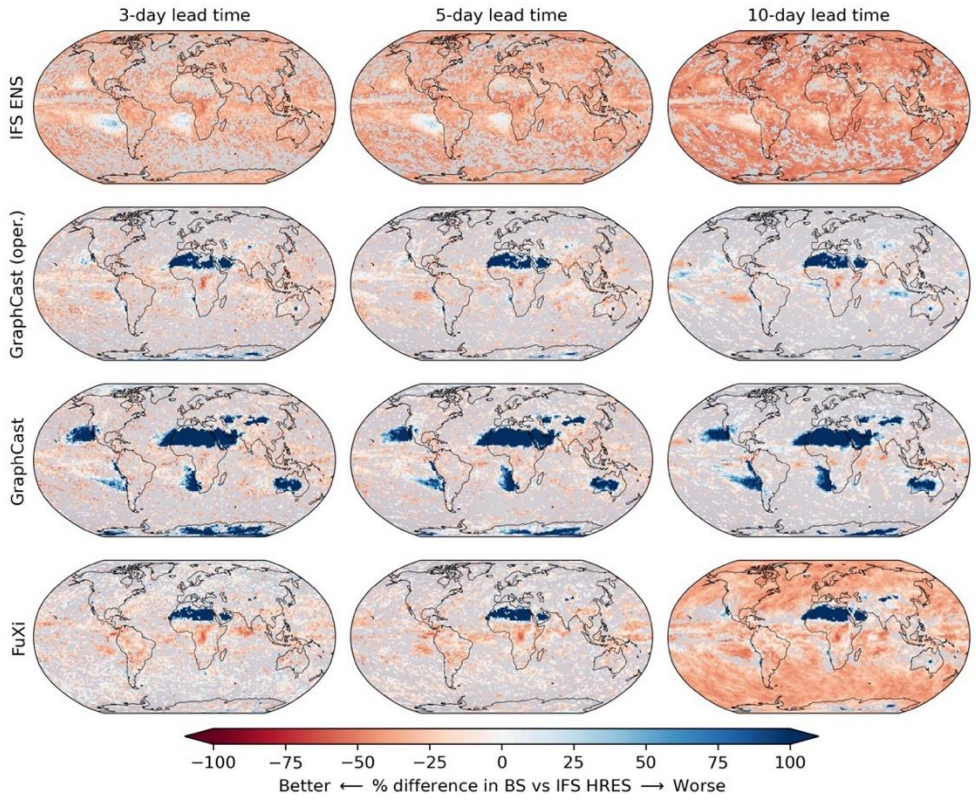

Figure S1. As for Figure 4, but for Brier score (BS).

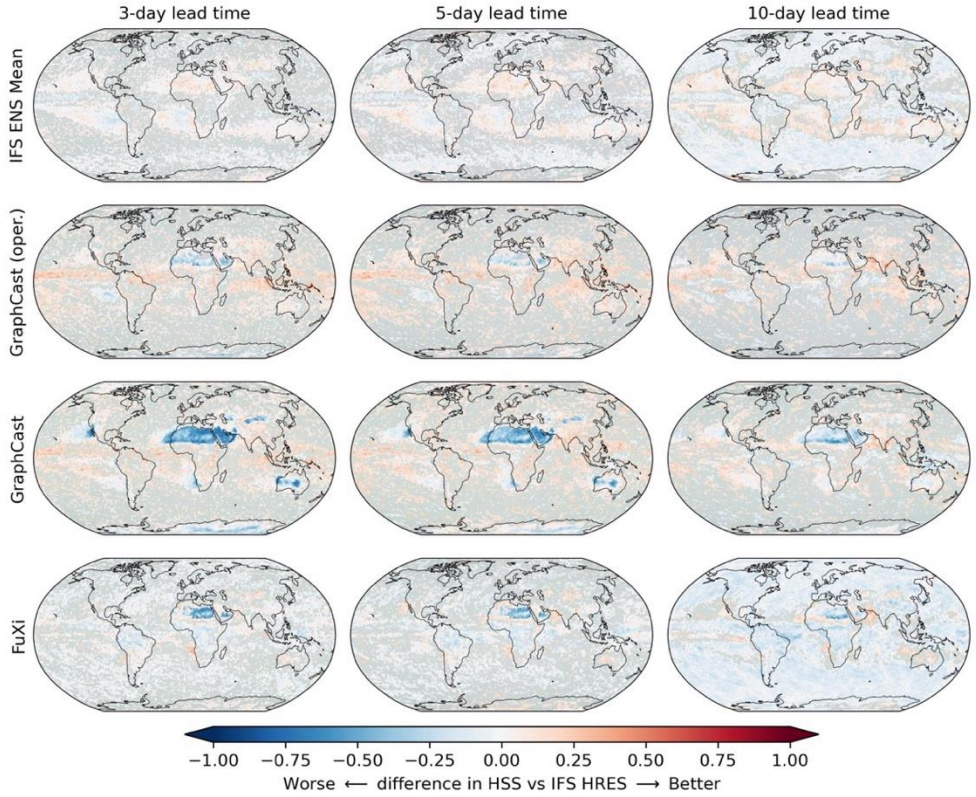

Figure S2. As for Figure 4, but for the Heidke skill score (HSS).

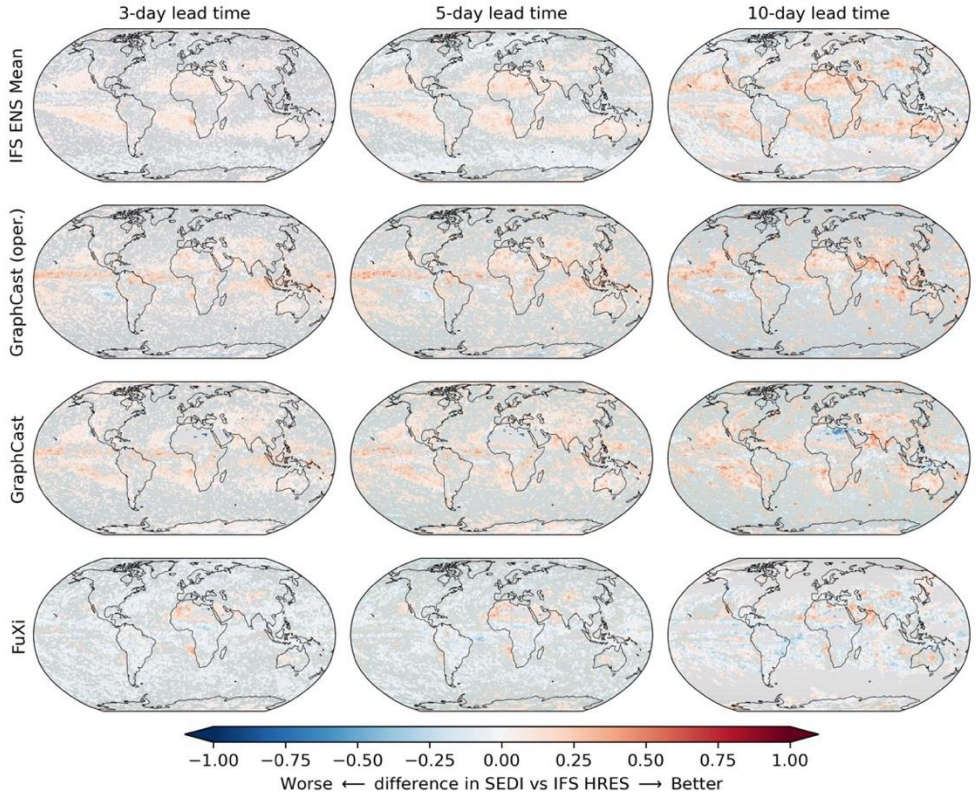

Figure S3. As for Figure 4, but for the symmetric extremal dependence index (SEDI).

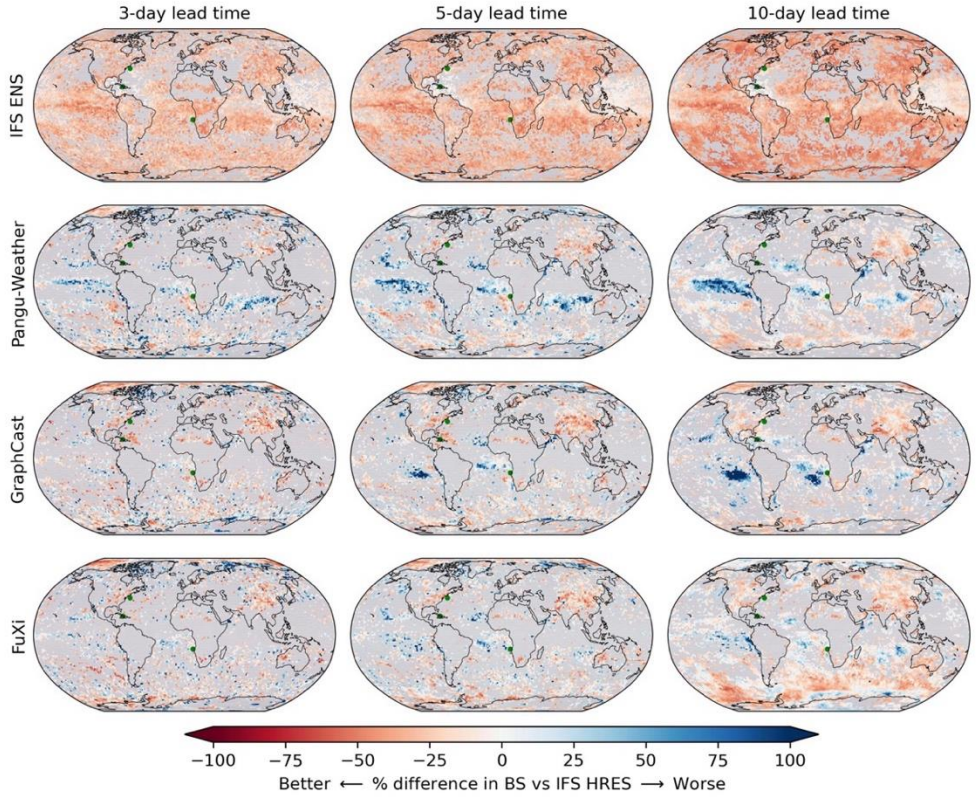

Figure S4. As for Figure 6, but for Brier score (BS).

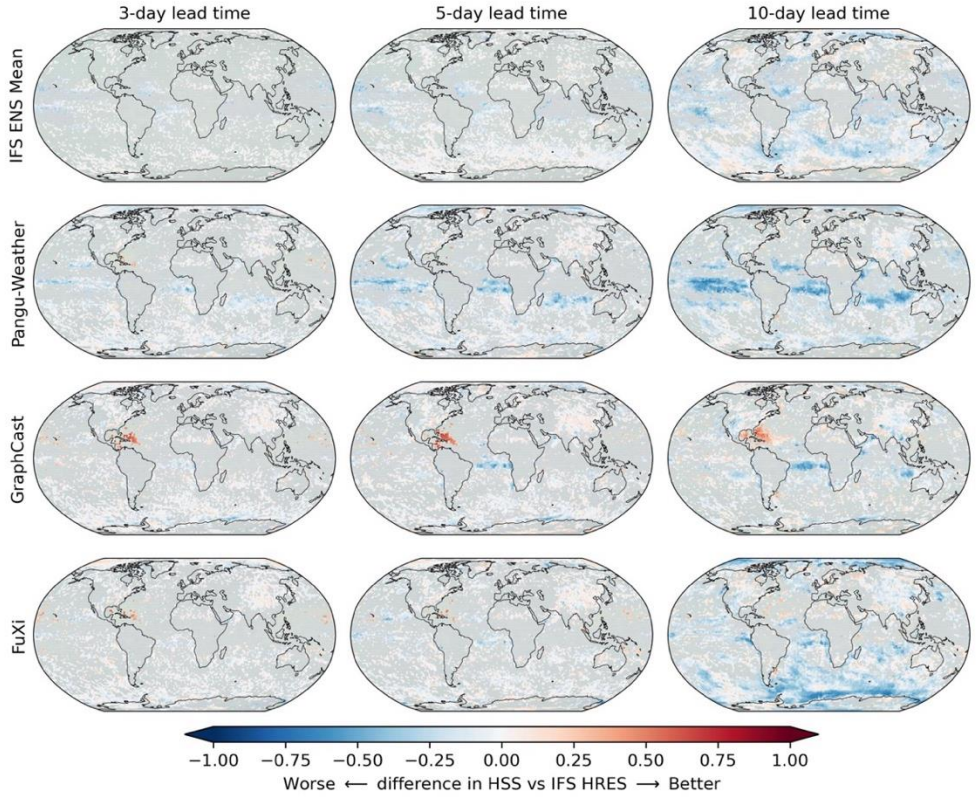

Figure S5. As for Figure 6, but for the Heidke skill score (HSS).

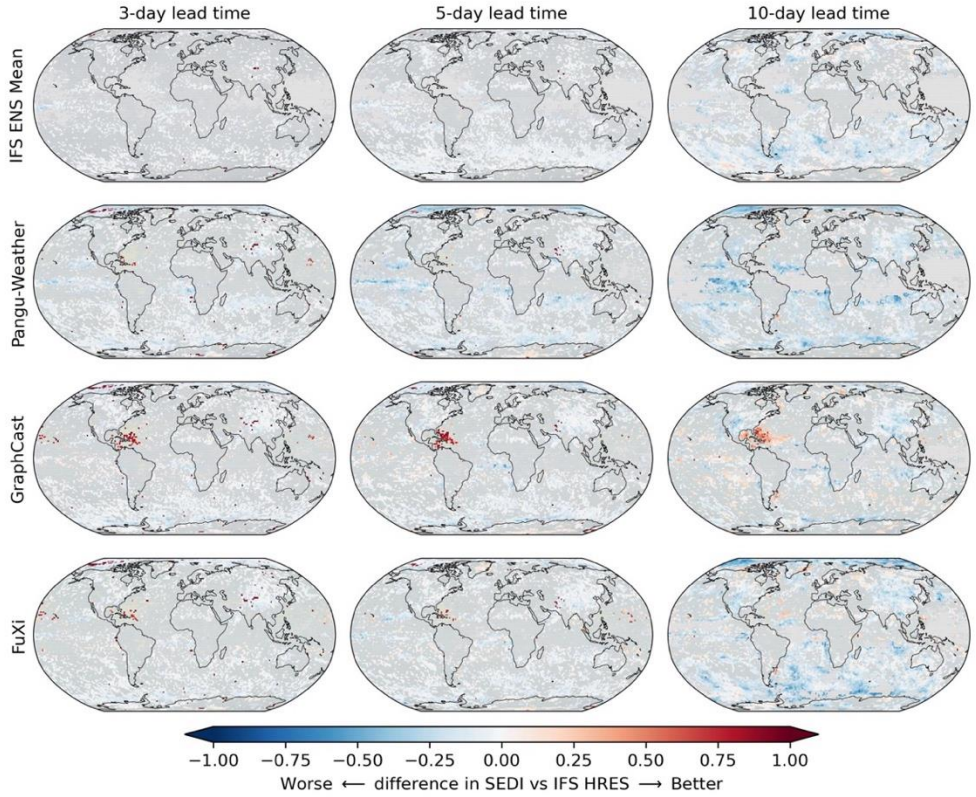

Figure S6. As for Figure 6, but for the symmetric extremal dependence index (SEDI).

" (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:

"**5.1 Binary forecasts and forecaster's dilemma**

Binary hydroclimatic forecasts can provide useful information for disaster prevention and risk mitigation (Ben Bouallègue et al., 2024; Merz et al., 2020). The evaluation of deterministic forecasts and ensemble forecasts is usually based on average metrics that focus on the overall predictive performance, such as the RMSE and CRPSS (Huang and Zhao, 2022; Rasp et al., 2024). Models that have minimized average errors and unrealistically smooth forecasts can also be rewarded by these metrics, leading to their limited guidance to forecast hydroclimatic extremes (Ferro and Stephenson, 2011; Rasp et al., 2020). By contrast, metrics for binary forecasts can emphasize the ability to discriminate certain hydroclimatic extremes that contribute little to average metrics (Larraondo et al., 2020). Binary forecasts are thus more suitable than continuous forecasts in these cases. 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 valid period of lead time when data-driven models are more skilful is shorter than that of the previous studies for 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 outcome, which indicates the unique insights of binary forecasts for hydroclimatic extremes.

The hydroclimatic system is high-dimensional and complex so that there won't be a single verification metric to determine all essential characteristics of a good forecast (Rasp et al., 2024; Jolliffe and Stephenson, 2012). While the metrics for binary forecasts can emphasize the discrimination, they are unable to reflect other attributes to quantify the forecast quality, such as reliability, resolution, uncertainty and etc. (Murphy, 1993). In the meantime, as shown in Fig. 5, when the GraphCast is more capable of capturing the wet extremes, it tends to produce more false positives, leading to the "forecaster's dilemma" (Lerch et al., 2017). It has been shown that using a combination of different types of verification metrics and diagnostic plots is effective (Larraondo et al., 2020; Huang and Zhao, 2022). As shown in Fig. A1 and Fig. B1 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 events and non-events, better values of a single metric do not mean more useful forecasts (Rasp et al., 2024). Therefore, the forecast verification needs to be guided by the operational applications and to account for the trade-offs between accuracy and forecast activity (Ben Bouallègue and the AIFS team, 2024; Rasp et al., 2024). " (Pages 22 and 23, Lines 343 to 367)

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 physical 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 are calculated separately for each grid cell to obtain an equal number of extreme samples per grid cell." (Page 5, Lines 111 and 112)

"For comparison at the grid scale, the 17 metrics are computed separately for each grid cell. To facilitate comparisons at regional to global scales, the 17 metrics are calculated using the area-weighting method based on the scores that are separately calculated for each grid cell (Rasp et al., 2024)." (Page 8, Lines 179 to 181)

- 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)
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)
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)
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 numbers 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 138 to 142)

"Among the seventeen metrics, the ROCSS is base-rate-independent and suitable simultaneously for deterministic and probabilistic forecasts of binary events. By contrast, other metrics need the predefined probability threshold to convert the probabilistic forecasts to deterministic forecasts. Therefore, the ROCSS is selected as the main 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). A higher ROCSS indicates better predictive skill." (Page 8, Lines 168 to 172)

- 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., 2024). 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 and 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.

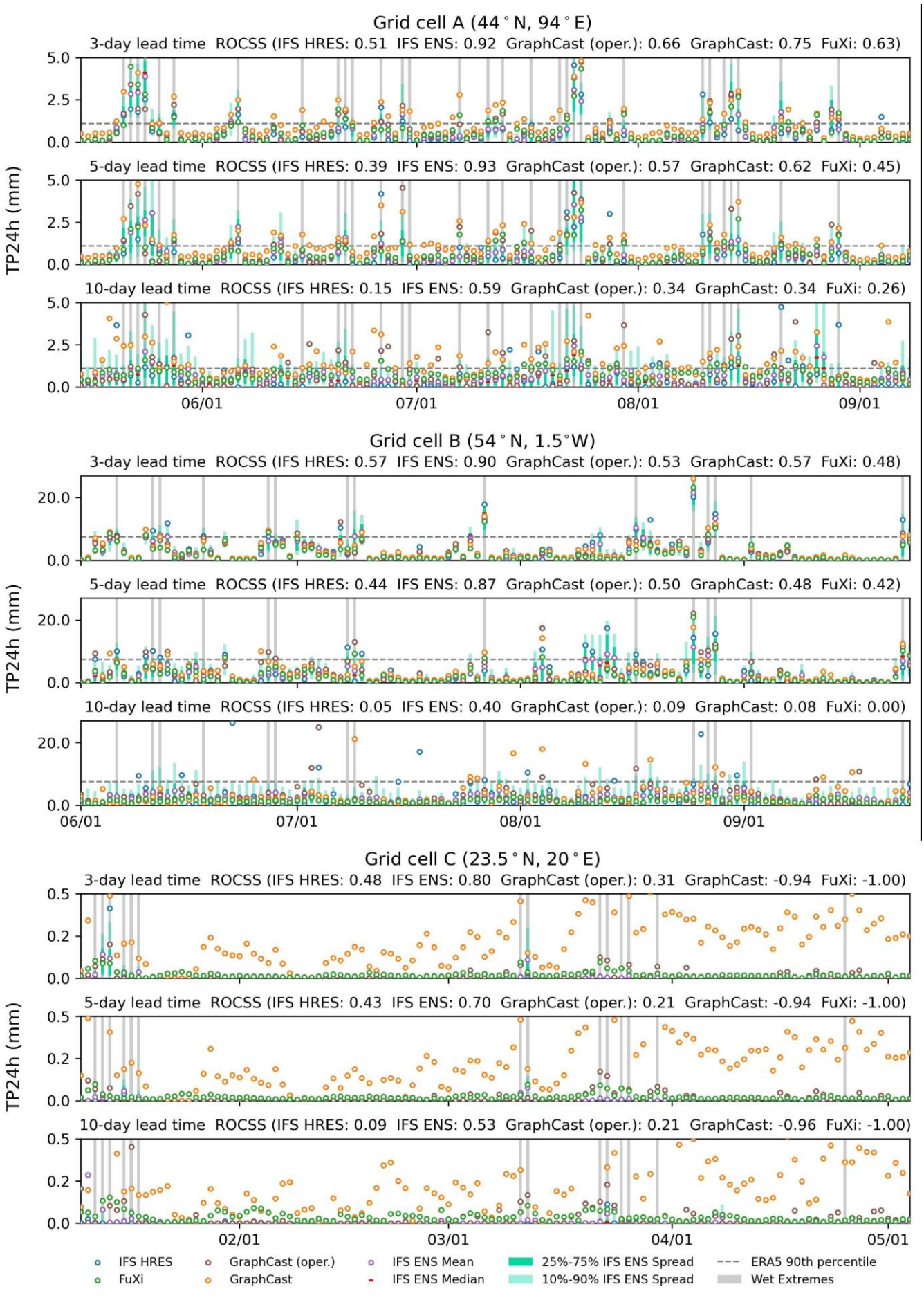

Figure 5. Time series plots of TP24h forecasts initialized at 00 UTC for the IFS HRES, IFS ENS, IFS ENS Mean,

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 17, Lines 263 to 277)

“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.

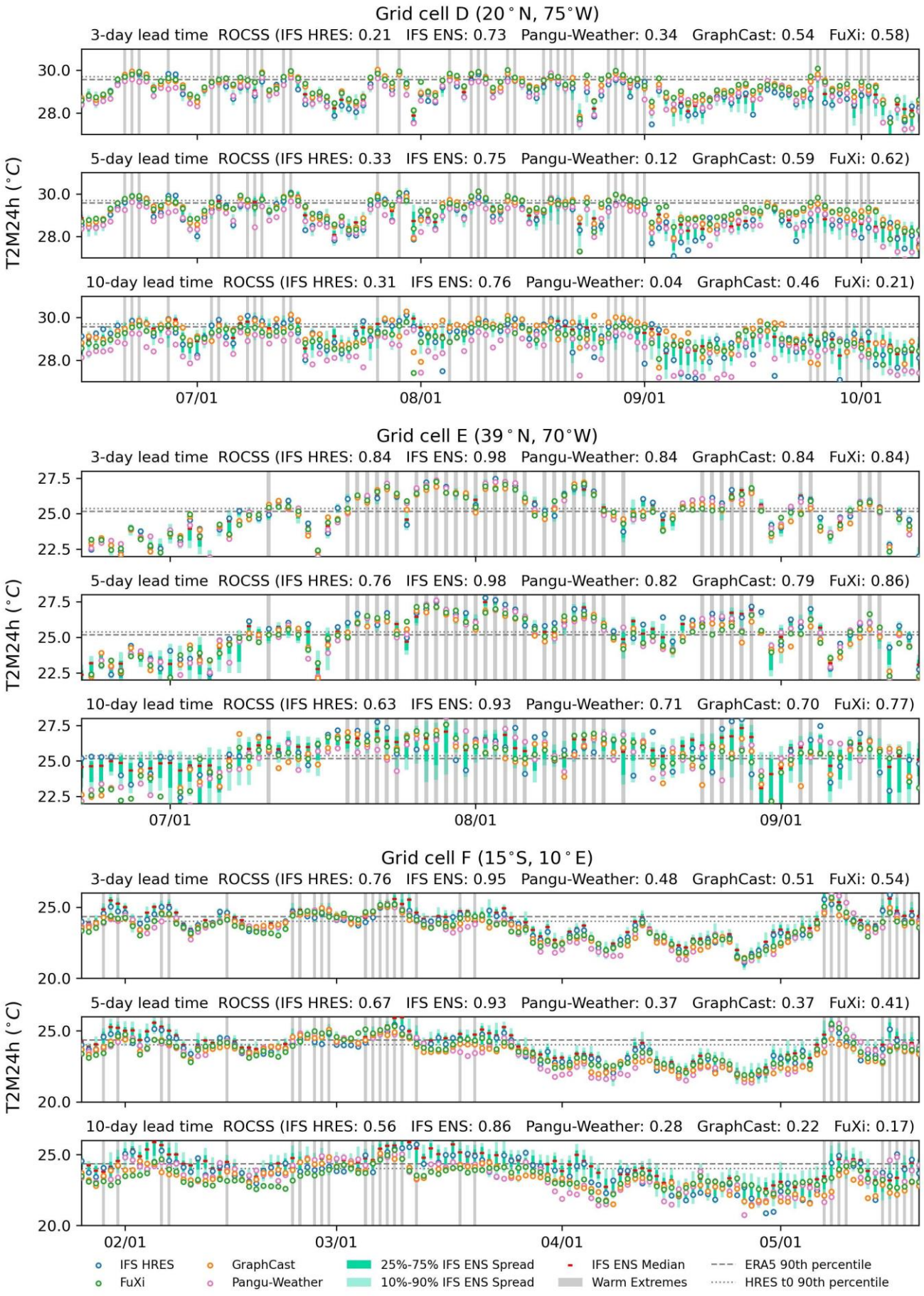

Figure 7. Time series plots of T2M24h forecasts initialized at 00 UTC for the IFS HRES, IFS ENS, Pangu-Weather,

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 295 to 309)

- *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 outcome, which indicates the unique insights of binary forecasts for hydroclimatic extremes.” (Page 23, Lines 362 to 365)

“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 events and non-events, better values of a single metric do not mean a more useful forecast (Rasp et al., 2024).” (Page 23, Lines 354 and 355)

- *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 can provide useful information for disaster prevention and risk mitigation (Ben Bouallègue et al., 2024; Merz et al., 2020). The evaluation of deterministic forecasts and ensemble forecasts is usually based on average metrics that focus on the overall predictive performance, such as the RMSE and CRPSS (Huang and Zhao, 2022; Rasp et al., 2024). Models that have minimized average errors and unrealistically smooth forecasts can also be rewarded by these metrics, leading to their limited guidance to forecast hydroclimatic extremes (Ferro and Stephenson, 2011; Rasp et al., 2020). By contrast,

metrics for binary forecasts can emphasize the ability to discriminate certain hydroclimatic extremes that contribute little to average metrics (Larraondo et al., 2020). Binary forecasts are thus more suitable than continuous forecasts in these cases. 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 valid period of lead time when data-driven models are more skilful is shorter than that of the previous studies for 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 outcome, which indicates the unique insights of binary forecasts for hydroclimatic extremes." (Pages 22 and 23, Lines 344 to 355)

- *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 hydroclimatic system is high-dimensional and complex so that there won't be a single verification metric to determine all essential characteristics of a good forecast (Rasp et al., 2024; Jolliffe and Stephenson, 2012). While the metrics for binary forecasts can emphasize the discrimination, they are unable to reflect other attributes to quantify the forecast quality, such as reliability, resolution, uncertainty and etc. (Murphy, 1993). In the meantime, as shown in Fig. 5, when the GraphCast is more capable of capturing the wet extremes, it tends to produce more false positives, leading to the "forecaster's dilemma" (Lerch et al., 2017). It has been shown that using a combination of different types of verification metrics and diagnostic plots is effective (Larraondo et al., 2020; Huang and Zhao, 2022). As shown in Fig. A1 and Fig. B1 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 events and non-events, better values of a single metric do not mean more useful forecasts (Rasp et al., 2024). Therefore, the

forecast verification needs to be guided by the operational applications and to account for the trade-offs between accuracy and forecast activity (Ben Bouallègue and the AIFS team, 2024; Rasp et al., 2024). ” (Page 23, Lines 356 to 367)

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:

“For operational applications such as disaster warning, the emphasis is usually on the occurrence versus non-occurrence of certain hydroclimatic extremes instead of their precise magnitude (Larraondo et al., 2020; Rasp et al., 2020). Binary forecasts match this need and can be directly used for decision-making (Jolliffe and Stephenson, 2012; Larraondo et al., 2020). In the meantime, binary forecasts can emphasize 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 59)

- 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:

“As the critical variables to understand and forecast the hydroclimatic processes such as floods and heatwaves, the precipitation and temperature are of concern in operational forecasts (Huang and Zhao, 2022; Lang et al., 2014; Zhao et al., 2022; Slater et al., 2023).” (Page 5, Lines 105 to 108)

- 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:

"Given the skewness and censoring characteristics of hydroclimatic variables, the standard paired t test is not applicable in this case (Huang et al., 2023, 2022). In the meantime, the selected hydroclimatic observations are subject to the problems of heteroscedasticity and autocorrelation due to the spatial and temporal clustering of hydroclimatic extremes (Olivetti and Messori, 2024). To address these issues, the cluster-robust standard errors are used to correct the paired t test (Liang and Zeger, 1986; Shen et al., 1987)." (Page 9, Lines 186 to 190)

- 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:

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.

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.

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.

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.

Huang, Z., Zhao, T., Tian, Y., Chen, X., Duan, Q., and Wang, H.: Reliability of ensemble climatological forecasts, Water Resour. Res., 59, e2023WR034942, https://doi.org/10.1029/2023WR034942, 2023.

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

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.

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.

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.

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.

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.

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., 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, https://doi.org/10.1017/CBO9780511617652.026, 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.

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.

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.

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.

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.