Responses:

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 information 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 previous code may not be quite accessible. In the revision, the code and scripts are provided in the form of Jupyter notebooks to facilitate learning and reproducing the entire procedures of forecasts verification instead. In the revised version, the plug-and-play code is also provided.

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 small code library. In the revision, we rewrite and reorganize the scripts to be plug-and-play:

“Code and data availability

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, 2025a). All the analysis results are archived on the Zenodo under https://doi.org/10.5281/zenodo.15067178 (Li and Zhao, 2025b).

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, 2025d) and under https://doi.org/10.5281/zenodo.15066898 (Li and Zhao, 2025c).

To ensure the compatibility with the WeatherBench 2, the code and scripts have been made a push request to its successor, i.e., WeatherBench-X.” (Page 24, Lines 393 to 401).

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 comments. 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 relative economic value (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 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).” (Page 8, Lines 162 to 164)

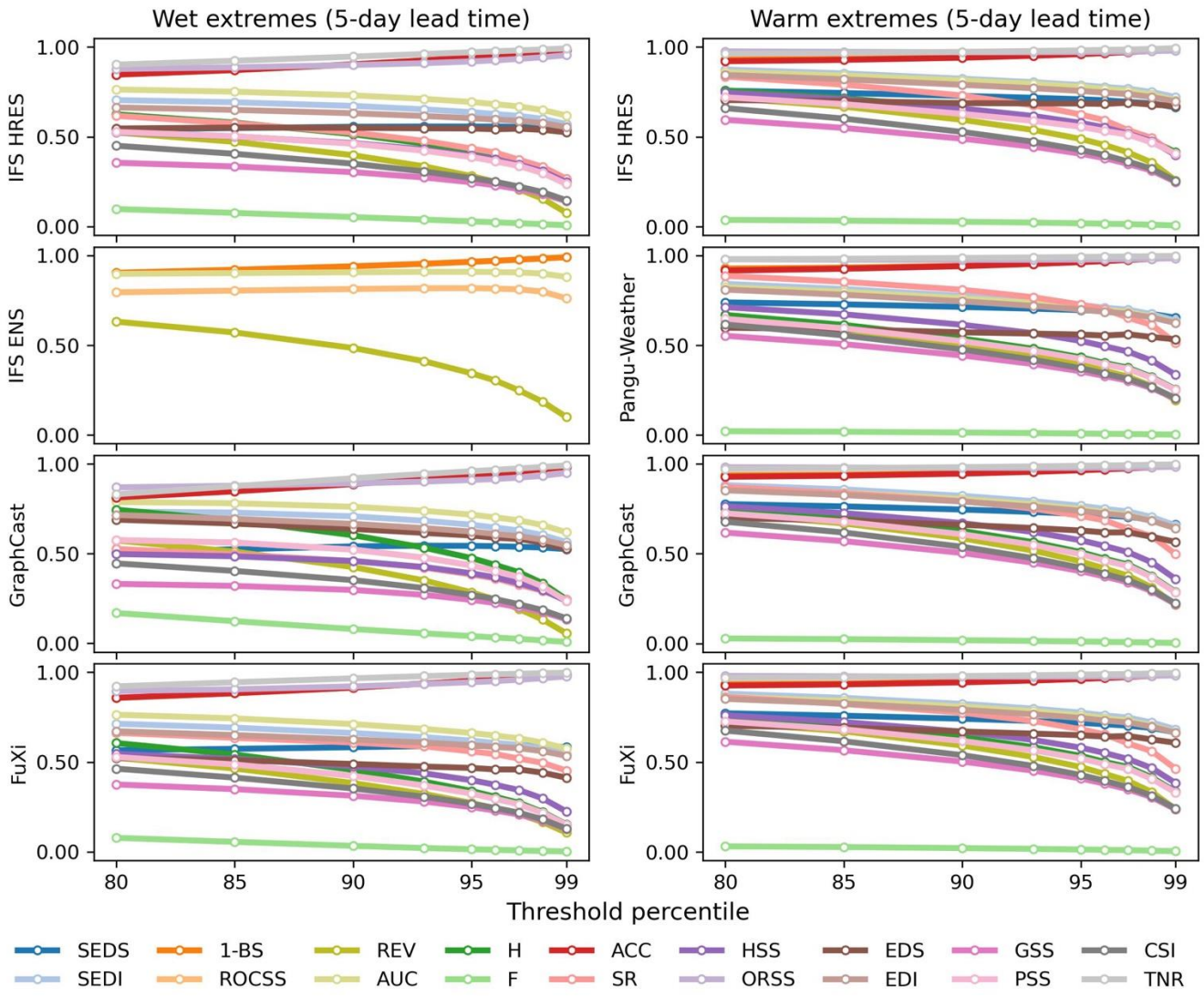

Figure 8. Globally area-weighted performance in forecasting wet extremes and warm extremes with different threshold percentiles at 5-day lead time. The REV is calculated with a fixed cost-loss-ratio of 0.2 only for purposes of illustration.

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 use 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 relation to the corresponding observations, binary forecasts can be divided into 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). The true positives represent events that are successfully forecasted; the false positives are non-events that are incorrectly forecasted as events; the false negatives denote events that are incorrectly forecasted as non-events; and the true negatives represent non-events that are correctly forecasted as non-events." (Page 8, Lines 162 to 164)

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:

"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 147 to 151)

In the meantime, the description of each score in the text has been improved:

"The 8 base-rate-dependent metrics in Table 3 are influenced by the underlying distribution of observed events and non-events (Jolliffe and Stephenson, 2012). The accuracy is calculated as the ratio between the number of true positives and the total number of events and non-events (Finley, 1884). The success ratio (SR) measures the number of true positives divided by the number of forecasted events (Lagadec et al., 2016). The critical success index (CSI) is the number of true positives divided by the total number of forecasted and observed events (Chakraborty et al., 2023; Gilbert, 1884; Donaldson et al., 1975). The Heidke skill score (HSS) measures the accuracy relative to that of the random forecasts (Gomis-Cebolla et al., 2023). The Gillert skill score (GSS) evaluates the fraction of true positives over the observed and forecasted events after adjusting for the random true positives (Chen et al., 2018; Coelho et al., 2022). 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 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 events (Ferro and Stephenson, 2011). The hit rate and false alarm rate respectively quantify the proportion of true positives in observed events and the proportion of false positives in observed non-events (Swets, 1986). The specificity measures the percentage of true negatives to observed non-events (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 (Swets, 1986). The ROC skill score (ROCSS) compares the discriminative ability over random forecasts and a higher ROCSS indicates better predictive skill. For probabilistic forecasts, the ROCSS can be calculated by considering the hit rate and false alarm rate for all possible thresholds of probability (Huang and Zhao, 2022)." (Pages 7 and 8, Lines 153 to 177)

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?

We are sorry for the confusion. 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, Line 158)

"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 168 to 170)

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

Thank you. 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 180 and 181)

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 202 to 204)

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 given that the problem of over-smoothing is more prominent among data-driven models than physical models (Bonavita, 2024; Lam et al., 2023).” (Page 9, Lines 208 to 210)

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:

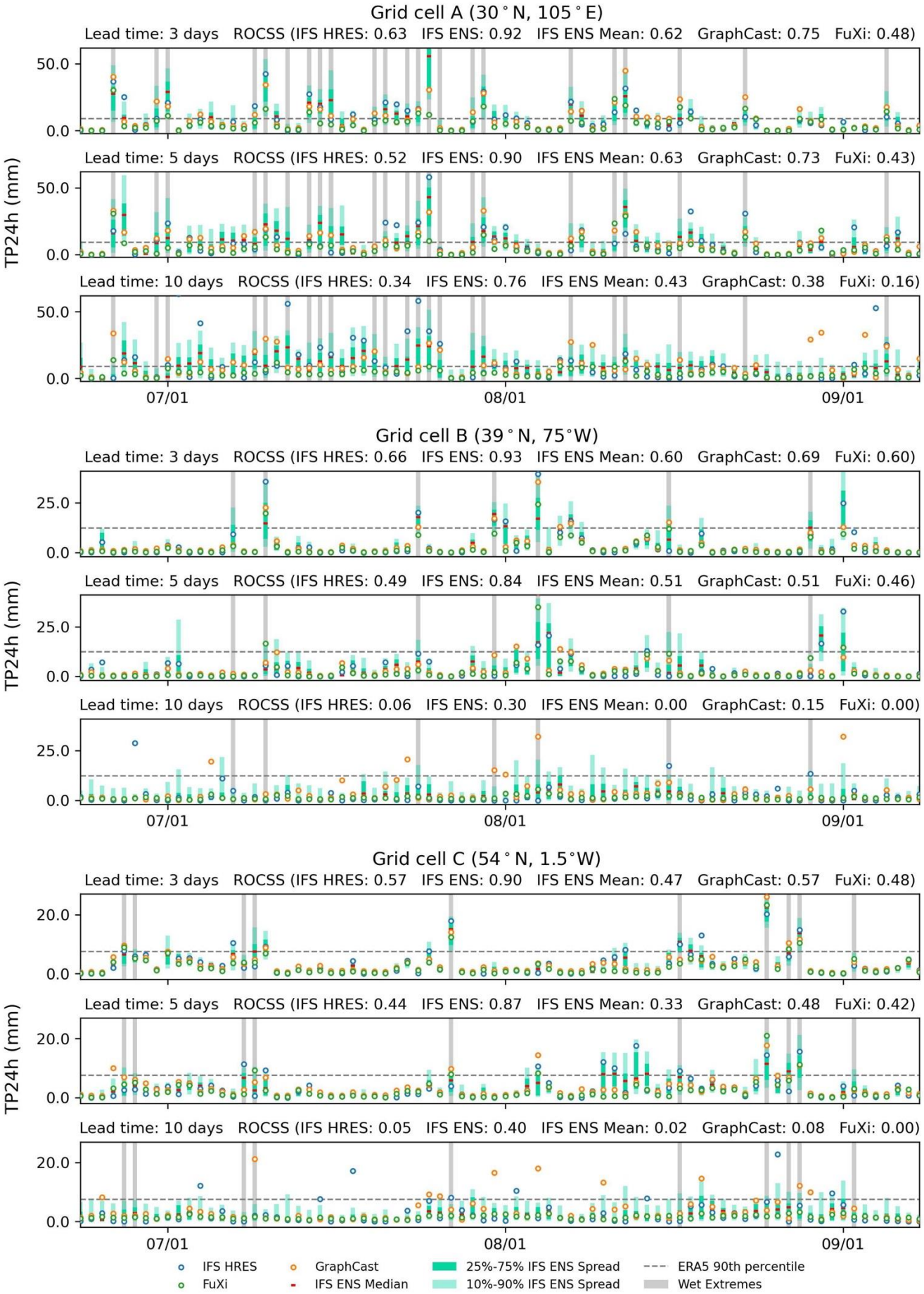

Figure 5. Time series plots of TP24h forecasts initialized at 00 UTC for the IFS HRES, IFS ENS, IFS

ENS Mean, GraphCast and FuXi over three selected grid cells, i.e., A (30°N, 105°E), B (39°N, 75°W) and C (54°N, 1.5°W). (Pages 16 and 17, 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 comments. 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). Given the complexity of regional geographic and climatic conditions, 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). In the meantime, hydroclimatic forecasts of coarse temporal resolution might miss extreme values and the underlying evolution processes due to the mismatch between forecast time step and event timescale (Pasche et al., 2025). 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). Therefore, there exists a demand to enhance the spatial and temporal resolution of data-driven models (Xu et al., 2024b; Zhong et al., 2024). 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 353 to 364)

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:

"With the availability of more data on hydroclimatic forecasts and baseline ground-truth observations, binary forecasts of hydroclimatic extremes deserve more in-depth verifications." (Page 23, Lines 371 to 373)

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 ERA5 reanalysis rather than ERA5 forecasts. We have revised this sentence to emphasize this point:

"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 382 and 383)

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.

We fully agree on this point:

"High skill of data-driven models in forecasting wet extremes can stem from the unfair setting of ground truth data (Rasp et al., 2024; Lam et al., 2023). As for the WeatherBench 2, it has been noted 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 23, Lines 365 to 369)

"With the availability of more data on hydroclimatic forecasts and baseline ground-truth observations, binary forecasts of hydroclimatic extremes deserve more in-depth verifications." (Page 23, Lines 371 to 373)

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