Responses:

Editor:

Topic editor decision: Publish subject to minor revisions (review by editor)

Please reply to the reviewer's comments in round 2, then we will make a decision for your manuscript.

We are grateful to you for the kind decision and to the reviewers for the insightful and constructive comments. Accordingly, we have revised this paper:

1. More details on the relationship between binary forecasts and deterministic and ensemble forecasts of continuous variables have been added;

2. The equations in Table 2 have been improved.

In the meantime, we have checked the whole manuscript carefully and corrected the typos in Lines 14, 62, 82, 89, 90, 95, 175 and 410.

<u>Report #2 from Anonymous referee #3:</u>

This manuscript presents a valuable extension of the Weatherbench2 framework by introducing a set of metrics tailored to the evaluation of binary forecasts of hydroclimatic extremes. Given the growing interest in evaluating extreme events for risk-based decision-making, this work has clear relevance for both research and operational forecasting communities. The core contribution is strong and well-executed.

We are grateful to you for the positive comments.

I believe the manuscript would benefit from clearer conceptual framing, particularly around the definition and role of binary forecasts:

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. Throughout the manuscript, the term "binary forecasts" is sometimes introduced as if it were a separate category of forecast alongside deterministic and ensemble forecasts (e.g., Page 2, Line 53 'Besides deterministic and ensembles, there is a demand of binary forecasts in disaster …'). Some sentences conflate forecast generation types (e.g., deterministic vs. ensemble) with forecast formats (e.g., continuous vs. binary), which could confuse readers unfamiliar with forecast verification practices. Binary forecasts are derived from deterministic or probabilistic (ensemble) outputs via thresholding. I would recommend explicitly defining binary forecasts in the introduction and consistently describing them as categorical expressions of the underlying continuous forecasts.

Thank you very much for the constructive comment. We have added more details:

"While the recent WeatherBench 2 provides a versatile framework for verifying deterministic and ensemble forecasts of continuous variables, this paper presents an extension to binary forecasts on the occurrence versus non-occurrence of hydroclimatic extremes." (Page 1, Lines 6 to 9)

"By following established practices in the World Meteorological Organisation (WMO), the WeatherBench 2 pays attention to both deterministic and ensemble forecasts of continuous variables generated by physical and data-driven NWP models (Jin et al., 2024)." (Page 2, Lines 50 to 53)

"Binary forecasts on the occurrence versus non-occurrence of target events can be generated from deterministic and ensemble forecasts of continuous variables by using predefined thresholds of hydroclimatic events (Ben Bouallègue et al., 2024)." (Page 5, Lines 116 to 118)

"Verification metrics for deterministic and ensemble forecasts of continuous variables, such as the RMSE and the CRPS, in general focus on the overall predictive performance across a range of events (Huang and Zhao, 2022; Rasp et al., 2024)." (Page 22, Lines 379 to 381)

2. Table 1 includes metrics (e.g., RMSE, CRPS) used for continuous-valued forecasts, while the paper later introduces binary metrics (e.g., ROCSS, CSI). This distinction can be made clearer in the text, especially in Sections 2 and 3. For instance, the manuscript might benefit from explicitly contrasting continuous forecast verification with binary event verification.

Thank you for the insightful comment. The verification of binary forecasts of hydroclimatic events is based on the contingency table by comparing the forecasts and the corresponding observations respectively with the predefined thresholds. By contrast, the verification of deterministic and ensemble forecasts of continuous variables compares the forecasts with the observations directly. The tables and the text in Sections 2 and 3 has been improved to make the distinction clearer:

"Table 1. Verification metrics for deterministic and ensemble forecasts of continuous variables in the WeatherBench 2." (Page 4, Line 107)

"Table 3. Verification metrics for binary forecasts." (Page 6, Line 164)

"The contingency table plays a key part in the verification of binary forecasts of hydroclimatic events (Larraondo et al., 2020). As shown in Table 2, there are four parts of the contingency table, i.e., true positives (a), false positives (b), false negatives (c) and true negatives (d)." (Page 5, Lines 129 to 131)

"Table 2. Contingency table for binary forecasts.

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

Where $M = 1$ *and* $M \geq 2$ *respectively represent deterministic and ensemble forecasts;* N *is the number of pairs of observations and forecasts for verification;* o_n *represents the* n*-th observation;* p *denotes the probability thresholds above which the occurrences are forecasted to occur for ensemble forecasts;* $I()$ *denotes the indicator function*" (Pages 5 and 6, Lines 138 to 153)