Response to reviewer 1

We would like to thank the reviewer for the time they took to provide valuable feedback that led to an improved manuscript. The reviewers' points are bold.

This article contributes to the discussion on the performance of AI models for weather forecasting, with a particular focus on their ability to predict extreme precipitation events. The methodology incorporates several novel ideas in verification, including spatial verification using a neighbourhood pseudo-ensemble, a threshold-weighted CRPS, and a decomposition of the CRPS using post-processing.

The paper reads very well overall. The data and the verification methodology are generally well described, and the figures are clear and easy to read. However, I found myself going back and forth between Figures 3, 4, and 7 to compare the results. Perhaps the authors could find a way to keep the results from Figure 3 visible in Figures 4 and 7 (and those from Figure 4 in Figure 7). This would make it easier to follow the presentation of the results.

While I find the study very interesting, I would encourage the authors to add a couple of discussion points

Thank you. We explored if we could keep the results from earlier figures displaying on latter figures but found that the figures became too cluttered and that the y-axis range became so wide that it was hard to interpret the lines as they overlapped more.

I would encourage the authors to add a couple of discussion points:

1. **In Section 3.1, it would be important, to my opinion, to discuss representativeness and to what extent a grid-box average can be directly compared to a point observation. In Section 6 (Model climatology), the Q-Q plot is quite compelling. How much of the off-diagonal behaviour is due to the smoothness in the forecast versus representativeness issue. A discussion on this could be interesting.**
   To test this, we added the neighbourhood mean of the HRRR 7x9 model to the Q-Q plot. We also added the following text when discussing the Q-Q plots

   “To explore the relationship between grid resolution and representativeness, we also take the neighbourhood mean of HRRR 7X9. The mean of the HRRR 7X9 neighbourhood shifts the distribution away from the diagonal line, but not nearly to the same extent as GraphCast-GFS 1X1, suggesting that differences in the

climatology of extremes may be driven more by the smoothness of GraphCast-GFS than by differences in grid resolution”

2. **About the climatology used to define the thresholds, using ERA5 instead of long time series of observations touches upon the representativeness issue too. Also, a question is whether the thresholds are season dependent or constant throughout the year. If the latter is correct, what are the implications for the interpretation of the results?**

   Yes, the ERA5 thresholds will likely be less than a station climatology based on the representativeness issue, however, the exact way that we define the extreme threshold is somewhat arbitrary.

   The thresholds are constant throughout the year. Either option could be chosen. We chose to keep them constant as at least at the BoM, warnings for heavy rainfall are based on a constant threshold and do not vary with season. This is because infrastructure is engineered to withstand impacts up to a certain annual exceedance probability which generally do not vary with seasons (noting that there will be some variation in practice. I.e., compound events if it rained heavily the day before).

   We have updated the paragraph on this topic in the observations section to read as:

   “To define thresholds for extreme precipitation events, several options can be considered, each addressing slightly different questions. These include using fixed thresholds across all stations, annual climatological thresholds for each station, or seasonally varying climatological thresholds at each station. In this study, we adopt annual climatological thresholds for each station, which aligns with approaches used by some meteorological agencies in operational warning services.

   Specifically, we define extreme precipitation events using the 99th and 99.9th percentile thresholds of six-hour precipitation accumulations at each station. As long observational time series are not available for all stations in the ASOS dataset, these thresholds are derived from ERA5 reanalysis data (1990–2020) at the grid point corresponding to each station location. These thresholds may differ from those derived directly from station observations; however, in the absence of a universally defined threshold for extreme precipitation, the exact choice of threshold is inherently somewhat subjective.”

3. **In Section 5, you mention CRPS being the integral of BS scores with the BS for small thresholds contributing most to the overall score. Could you comment on how twCRPS works with that respect? Would the main contribution to this score be the BS for the 99% percentile in your example?**

   The relationship between twCRPS and the integral of the Brier score depends on the chosen weighting function. The twCRPS is defined as:

$$\text{twCRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbf{1}\{y \leq z\})^2 \, w(z) \, dz$$

   where $(F(z) - \mathbf{1}\{y \leq z\})^2$ is the Brier score at threshold z. The function $w(z)$weights the Brier score at each threshold before the integral is taken.

   If the weighting function were $w(z) = \mathbf{1}\{z > 50\}$ the twCRPS would equal the area under the curve in Figure 5 to the right of 50 mm.

   However, since our weighting function uses thresholds that vary between stations (based on climatology), Figure 5 cannot be interpreted in this way, as it reflects scores aggregated across all stations. For this reason, we chose not to modify the paper to discuss this.

4. **You show the discrimination ability based on the twCRPS. It would be interesting to compute the discrimination based on the "full" CRPS too. That would help the interpretation of the results. Also, one could simply show the results of the post-processed forecasts (instead of showing the discrimination), that would ease the comparison with the other results and help assessing the impact of post-processing more directly.**
   Thank you for both these suggestions. We have taken on both these suggestions. We have rewritten the discrimination section to accommodate these changes.
   Note that we have set the y-axis range to match between the CRPS fig and the new potential CRPS fig, as well as the twCRPS fig and the potential twCRPS fig. We don't match the y-axis range on the confidence interval plots between those figures as it makes them harder to interpret as the confidence intervals become too small.

We also would like to highlight that in the previous version we used isotonic regression that targeted the mean functional. We have updated it to use a version that targets the median functional so that it consistent with MAE for the single ensemble member calculation of CRPS. We also made a small adjustment to the final sentence of the abstract.

Minor comments:

1. **Section 2.1. There are a couple of studies based on precipitation observations that have been published. See for example Jin et al (2025) and Ben Bouallegue et al (2026).**
   We already had cited Jin et al (2025). We have added the Ben Bouallegue et al (2026) paper. It wasn't possible to cite that paper as we submitted this paper to GMD before the SEEPS4ALL paper appeared as a preprint.
2. **Section 3.1. When mentioning pseudo-ensemble, I would cite the original paper describing this idea: Theis et al 2005.**
   Thank you. Added in the revised introduction.
3. **Figures 3,4, 7, 8. It is not explained in the text how the confidence intervals are computed (Diebold Mariano is only mentioned in the Code Availability Section).**
   Thank you for highlighting that we missed this. We have added the following sentence:
   "Confidence intervals for the difference in mean scores between the two models are calculated as follows: spatial means are first taken across stations to account for spatial correlation, and the Hering–Genton modification of the Diebold–Mariano test (Diebold and Mariano, 1995; Hering and Genton, 2011) is then applied to account for temporal correlation."
4. **Section 7. The CORP-like decomposition approach: it is not explained how it works. Equation 9 reminds me of Equation 15 in Siegert 2017. Is it the same idea?**
   Correct. It is the same idea as the RES component in equation 15. We rewrote the discrimination section to just show the potential (tw)CRPS like you suggested so we now no longer discuss the CORP-like decomposition approach and this section has now been removed.