Response to reviewer 2

We are grateful to the reviewer for their thorough and insightful review of our manuscript, which prompted us to improve the manuscript. The reviewers' points are bold.

In their study, the authors perform an evaluation of the forecast of extreme precipitation for two models: a traditional numerical weather prediction model and a data-driven AI weather prediction model. To avoid the "double penalty" issue of models with higher spatial resolution when performing point-to-point comparison, they apply the HiRA (High-Resolution Assessment) approach, where they also account for neighboring grid-cells, and compare at equivalent neighborhood sizes for the respective models. They combine this spatial sampling with the threshold-weighted Continuous Ranked Probability Score (twCRPS) to analyse the performance of both models for extreme precipitation.

This study introduces a novel approach to compare models at different horizontal resolutions, specifically focusing on extreme events. My major criticism is regarding the claim that no re-gridding is necessary when comparing models at different horizontal resolutions. The manuscript is well-written and fits well into the scope of GMD and merits publication once my concerns are addressed.

Thank you very much. We address the specific criticisms below.

Specific comments:

- **The authors state that when comparing models with different spatial resolutions, no re-gridding is required when using (tw)CRPS with HiRA. I'm not sure whether this is an assumption or whether the authors have verified this claim. I understand that different ensemble sizes are accounted for using the fair version of (tw)CRPS. I had a quick look at previous studies (e.g., Crocker et al., 2020) that used a similar approach and could not find an evaluation of the claim. In the following, I will argue why this claim might be overstated, and different horizontal resolutions will introduce systematic biases in the CRPS.**

  As CRPS is a strictly proper scoring rule, there exists a minimum for CRPS if the distribution of the forecasted quantity matches that of the observed quantity. Let's assume we have such a situation and that both follow Gaussian distributions (which might hold for a quantity like temperature)

with a variance of σ^2. Let's now assume we coarse-grain our forecast by forming n x n sub-grids within the fine grid and average over those sub-grids. This operation will keep the area of interest the same for coarse and fine grids. For the new coarse grid, the mean will be the same, but the variance will be reduced to σ^2/n^2. A similar behavior can be expected when one uses an atmospheric model with a coarser resolution, where one also would expect the variance to shrink. As the variance on the coarser grid is now different compared to the finer grid, the CRPS of the coarser grid will be larger. This follows from the CRPS being a proper scoring rule.
For rain, which doesn't follow a Gaussian but rather a log-normal distribution L(μ, σ^2), the situation is a little more complex but comparable. To not lose the lognormal distribution when averaging, let's use the geometric mean. This would cause the coarse-grained distribution to be L(μ, $\sigma\%^! = \sigma^!/n^2$). The arithmetic mean of a lognormal distribution is exp(mu + $\sigma\%^!/2$), so the arithmetic mean in the coarse-grained will shrink as the variance is reduced. Even if the model is recalibrated to preserve the mean—analogous to tuning a low-resolution model—the resulting distribution becomes less right-tailed. Performing the coarse-graining has again changed the distribution, being less right-tailed. With a similar argumentation as before, the CRPS of the coarser grid will be larger.

The examples above indicate that CRPS is not scale-invariant, and some form of regridding is required to make models at different horizontal resolutions comparable. I understand that from a user's perspective, a higher resolution might be better as more details might be captured, but in the context of a fair statistical validation, this higher fidelity will introduce systematic differences in the statistical quantities. These examples are worst-case assumptions. In the case of spatial correlation in the respective fields, the increase in CRPS would be reduced. Furthermore, one should not expect the distribution of the observations to be identical with the forecasted distribution of the fine model, even though the Q-Q plot in Fig. 6 shows that they agree well.

In summary, I would strongly recommend coarse-graining of the high-resolution model before making a statistical comparison between the two models.

This was a valuable perspective that prompted us to reconsider our approach and recalculate the scores accordingly. It also helped us clarify the specific objective of the verification. Our aim is to assess model performance from the

perspective of an operational meteorologist in a particular use case: treating the forecast neighbourhood as an empirical cumulative distribution function (CDF). From this perspective, we are interested in the overall practical value of the model for decision-making, rather than in estimating the CRPS of the underlying predictive distribution (from which the neighbourhood values are merely a sample).

As a result, we have made several changes:

1. We recalculated all the scores with their empirical versions rather than their fair versions. This was to assess the case where an operational meteorologist treats the neighbourhood as a predictive distribution. Note that the scores are extremely similar and it does not change the result of the paper.
2. We have updated the paper to be more explicit in the question that we are trying to answer. In the introduction, before the list of positives about this approach we added: “One practical use case for weather models is when meteorologists visually assess a deterministic precipitation forecast around a point to qualitatively infer the likelihood of different precipitation amounts. A quantitative forecast can be constructed by generating an empirical cumulative distribution function (CDF) from a neighbourhood pseudo-ensemble around the observation. Similarly, post-processing techniques often employ neighbourhood approaches on NWP output to generate probabilistic forecasts \citep{Theis_2005, Schwartz_2017}. In this paper we demonstrate a verification approach that merges two existing methods and assesses the model in one particular framework that aligns with a specific way that operational meteorologists may use a model. That is, we evaluate how well extremes are predicted when the model's spatial neighbourhood around a point is treated as an empirical CDF.
3. We removed the fair (tw)CRPS equations and corresponding text from section 3.2

- **I don’t fully understand the rationale behind Fig. 6, comparing GraphCast-GFS and HRRR at the grid point level in a Q-Q plot. As outlined in my previous point, a model with a coarser horizontal resolution will be less right-tailed compared to a model with a finer resolution. Even the authors acknowledge this fact. Therefore, not much can be learned from this comparison. If the authors want a scale-aware comparison of precipitation between models at different horizontal resolutions, they could perform coarse-graining as outlined in my previous point (geometrical mean and readjusting the arithmetic mean) to allow for a scale-aware comparison. The comparison between HRRR and GraphCast-GFS is therefore biased toward the higher-resolution system regardless of actual forecast skill.**

We have updated this figure to also include a Q-Q plot of the mean neighbourhood value of the HRRR 7X9 model. Rather than a geometric mean, we chose to take the arithmetic mean as it handles zero values (and has other minor benefits, e.g., conserves water volume).

- **I do struggle a bit with the structuring of the manuscript, in particular with the placement of Section 3.3. While being within Section 3, which mainly focuses on HiRA and (tw)CRPS, it reads more like an introduction and/or discussion, and only slightly touches on HiRA and (tw)CRPS.**
  Thank you for this suggestion. We have adopted it.
  We moved most of section 3.3 to the introduction and just after the paragraph on evaluating extremes. We also re-worded parts that we moved across to better fit in the introduction. The latter parts of section 3.3 were reworded and placed just before the list in the introduction. We also cut the second half of the last sentence under the heading "Using HiRA and twCRPS"

Minor Remarks:

- **Fig. 3/4/7: Why not use consistent ranges in the y-axis for the respective difference plots (b, c, d) in each figure, as actual values are not too different? One could even consider putting it in the same subfigure if it doesn't become too busy.**
  Thank you. We have updated them to be consistent. We found that it was too busy combining them all on the one subfigure.
- **I assume all figures referring to (tw)CRPS use the fair version. In this case, please mention it in the manuscript and/or in the figures that (tw)CRPS is synonymous with the fair version. On that note, if it is the fair CRPS, you are also looking at the fair Brier score (Ferro, 2014) in Section 5, are you?**
  Yes, in the original manuscript and yes, we originally used the fair Brier score in section 5. However, as discussed above, we have moved to use the empirical (tw)CRPS/Brier scores.

Other changes

- Updated arXiv citations to the published papers.
- Since this paper was submitted 6 months ago, there has been several new papers that have focused on evaluating how well AIWP models predict extremes. We have included these and simplified the paragraphs in the introduction that focus on the evaluation of extreme forecasts.
- Minor grammar/spelling fixes.

- Updated code and data availability section