# Response to Reviewer 2

August 2026

### Reviewer #2

> *I thank the authors for partly incorporating the remarks from my first review (Reviewer #2). Nevertheless, I don't feel that they have sufficiently addressed my remarks regarding the potential need for regridding before calculating the twCRPS.*
>
> *I can only accept the manuscript for publication once my concerns have been properly addressed.*

### Response

Thank you for reiterating this concern. We agree that our previous response did not sufficiently address it. Our previous revision focused on the distinction between empirical and fair scores, whereas the reviewer's concern was primarily about differences in spatial scale. We have now addressed these two issues separately and revised the manuscript accordingly.

We agree with the reviewer's point that the CRPS and twCRPS are not invariant to spatial aggregation. Coarse-graining a high-resolution precipitation field generally changes its variance, upper tail and predictive CDF, and can therefore change its CRPS or twCRPS. Consequently, our original statement that no re-gridding is required when comparing models with different grid spacings was too broad. We have now removed that general claim. The revised manuscript now states that re-gridding is required for a common-scale comparison, whereas this study evaluates the native-grid neighbourhood empirical CDF.

The revised manuscript now distinguishes between two different scientific questions:

1. Which forecast system performs better when both are evaluated at a common spatial scale, so that differences in grid spacing are not retained in the comparison?
2. How well does each native-grid neighbourhood empirical CDF predict precipitation at the observation site when the neighbourhood covers a specified physical area?

The reviewer's proposed coarse-graining addresses the first question. This paper addresses the second question.

The reviewer's proposed coarse-graining is appropriate for the first question. We now state explicitly that this comparison requires the forecasts to be re-gridded to a common grid and evaluated against observations represented at a comparable spatial scale. Such an analysis would be appropriate if the objective were to isolate differences between the forecast systems after removing the effects of grid spacing and spatial scale.

The updated paper addresses the second question. This is because we wanted to evaluate the forecasts in a user-focused framework. We are evaluating the models in a specific way in which deterministic model output can be interpreted by an operational meteorologist or used within a simple neighbourhood post-processing system. For each model, the $M$ native-grid values within a specified neighbourhood $\mathcal{N}$ define the equally weighted empirical CDF

$$\widehat{F}_{\mathcal{N}}(z) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{x_m \leq z\}. \tag{1}$$

In this framework, the empirical CDF is itself the forecast being evaluated. The neighbourhood values are not treated as a finite sample used to estimate a hypothetical underlying distribution. The CRPS and twCRPS are then calculated directly for this finite empirical forecast.

The native-grid comparison deliberately retains effects associated with grid spacing, effective resolution and spatial sampling. These are attributes of the forecasts evaluated in the second question. Re-gridding would change the empirical CDFs being evaluated and would instead answer the first question. We therefore interpret the rankings only within the native-grid empirical-CDF framework.

We have also clarified the distinct roles of the empirical and fair scores. The fair CRPS and fair twCRPS are appropriate under the alternative interpretation in which the neighbourhood values are treated as a finite sample from an underlying predictive distribution, provided that they can be regarded as suitable draws from that distribution. They do not remove differences in grid spacing or spatial scale. It is worth noting that neighbouring grid values are spatially dependent, which affects whether they can be regarded as suitable draws from an underlying predictive distribution. Since our analysis evaluates the finite neighbourhood empirical CDF itself, we use the empirical CRPS and twCRPS. This choice does not make the comparison independent of spatial scale, and the revised manuscript now states this explicitly.

To address the reviewer's feedback, we have made the following changes in this updated manuscript:

- We removed the general claim that forecast systems with different grid spacings can be compared without re-gridding.
- We added a new subsection, "Constructing and interpreting neighbourhood pseudo-ensembles", that distinguishes the two questions and explains the appropriate evaluation approach for each.
- We restored the fair CRPS and fair twCRPS equations to explain this distinction.
- We added the explicit statement that, if the objective is a comparison at a common spatial scale, the forecasts should first be re-gridded to a common grid and evaluated against observations at a comparable spatial scale.
- We revised the Introduction and Conclusions to state that our comparison includes, rather than removes, effects associated with grid spacing, effective resolution and spatial sampling.
- We revised the Abstract to describe the comparison in terms of the empirical CDFs produced from the two forecast systems rather than presenting the result as a resolution-independent ranking of the underlying models.

Additionally in our methods section we now state:

> *Evaluating native-grid neighbourhoods compares the resulting forecast products, including effects associated with grid spacing, effective resolution, and spatial sampling. If the aim is instead to compare forecast systems at a common spatial scale, the forecasts should first be re-gridded to a common grid and evaluated against observations at a comparable spatial scale. In this study, we use the native-grid empirical-CDF interpretation.*

We hope this resolves the concern more directly than our previous response. We accept the reviewer's argument that re-gridding is required for the common-scale comparison described in the first question. The present study instead addresses the second question, for which re-gridding would change the finite neighbourhood empirical CDF being evaluated. We have therefore retained the native-grid analysis while

revising the manuscript to define that forecast precisely and to restrict the interpretation of the results accordingly.