

I thank the authors for their thorough revision, and I believe the manuscript has improved substantially. In particular, the paper is now much clearer in terms of its structure, the discussion of the literature in the Introduction, and the inclusion of additional analyses. Overall, I think the study is now close to being publishable.

I only have a few minor comments:

L187a: “requires”?

L187b: I am not sure I agree with the statement that “reliability assessment, unlike many other skill metrics, requires aggregation of many grid points”. My understanding is that the need for spatial aggregation mainly arises from the limited hindcast sample available in this study, which would otherwise provide too few forecast-observation pairs at individual grid points to estimate robust reliability diagrams.

L308: Since the CRPS analysis is now presented in the Results and contributes to the interpretation of forecast performance, I think it should also be briefly introduced in the Methods section. In addition, my understanding is that the agreement metric represents the percentage of cases for which the two verification datasets assign the same binary-event classification after pooling all grid points and years within each region. I suggest explicitly defining this metric as well.