

Reply to the Report #2

We thank the anonymous referee for providing additional comments. Please find our replies below.

L187a: “requires”?

A: The typo has been corrected.

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.

A: The sentence has been rephrased to “Reliability assessment of seasonal forecasts, unlike many other skill metrics, requires aggregation of many grid points to obtain large samples due to the limited number of predictions available at a single grid point (equal to the number of years in this study).”.

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.

A: The CRPS was introduced as an additional point to support the conclusions. It is not, however, the key diagnostics used throughout the paper. We believe that the references and explanation provided when introducing the CRPS analysis are sufficient for the reader to interpret the results. The same stands for the agreement metric that is well described in the text. Thus, we believe that no manuscript changes are required in this regard.