Response to RC1 comments on "A new skill score for ensemble flood maps: assessing spatial spread-skill with remote sensing observations"

Thank you to Seth Bryant for your time reviewing our paper. Our responses to numbered comments in black are detailed below in *blue italics.*

1. Two pages are copied verbatim from Hooker et al., (2022). Instead, these should be summarized, and the reader directed to this other publication.

These pages relate to methods explaining the scale-selective approach to spatial scale validation and are an integral part of evaluating the spatial spread-skill of an ensemble flood map forecast. We feel that the full method and equations should be included here to enable understanding of the ensemble application. If we are invited to revise the manuscript, we will make some small edits (for example, in response to comments in the supplement) and it will no longer be exactly the same as in our previous paper.

2. There are numerous grammatical issues, redundant sentences/phrases, imprecise/inaccurate vocabulary, and a confusing overall sequence/structure which make the manuscript difficult to follow. The authors should consider the perspective of the reader, striving to be as concise and logical as possible.

These issues, detailed in the reviewer's supplement, will be addressed in full should we be invited to revise our paper.

3. While I'm unfamiliar with the details of flood forecasts, I can imagine and appreciate the motivation for such a metric. However, I'm skeptical the method proposed is appropriate for application against a simulation-library like Flood Foresight. For example, if each inundation raster within the library is monotonically nested (i.e., cells become progressively more flooded), a neighborhood approach seems unnecessary. More information on the Flood Foresight simulation implemented in this study is needed to evaluate this properly.

The Flood Foresight system is not set up in the way that the reviewer suggests. Instead, the catchment is divided into impact zones. Each impact zone is linked to a GLoFAS grid cell (providing discharge data for the event) and a different return-period-threshold-discharge and flood map selection. We will include more detail on this aspect of the Flood Foresight system in the revised manuscript. Please note that this validation procedure can also be used to improve individual flood maps held within the simulation library.

4. Similarly, additional details of the application of the permanent water body layer (in both the SAR-derived layer and the Flood Foresight layer) are necessary to evaluate the utility of the proposed method to the case study. For example, if the same source layer pre-filter is implemented in both the

‘observed’ and the ‘simulation’ data, rewarding the simulation for accuracy in these cells seems inappropriate.

The SAR-derived flood maps have permanent surface water bodies removed as part of the flood mapping process as a pre-flood image is used to compare against the flood image. To enable evaluation of the flood prediction accuracy alone, the pre-flood occurrence of surface water using the JRC Global Surface Water database has been removed from the forecast flood maps to allow a fair comparison to be made. We will add a comment to the revised manuscript to explain this.

5. To demonstrate the utility of the metric, the authors should consider comparing against some alternative. When is the proposed two-phased sophisticated method more appropriate than existing simple methods?

The novelty of this ensemble validation approach means that there are no other existing methods to compare to. In our previous paper (Hooker et al., 2022) we have compared the scale-selective evaluation to multiple existing commonly used binary performance measures. We will add this to the introduction section.

6. Additional synthesis of the results would be helpful to demonstrate the utility of the proposed method. For example, the authors suggest the metric can provide some ‘link to physical processes’, but no discussion of this is provided for the case study. How can the metric help us understand the role of dynamic morphology and levee performance in ensemble accuracy? How should we use Figure 9? For emergency response?

The physical processes mentioned in the discussion and conclusion include multiple types throughout the forecast-chain, from atmospheric processes which determine the location and intensity of precipitation though to hydraulic processes. Each of these will impact the spatial-spread-skill of the ensemble flood maps. For example, including observations of antecedent soil moisture could lead to reduced uncertainty in the forecast discharge and an improvement in the spatial predictability of the flood maps. Inaccuracies linked to the hydrodynamic modelling used to produce the flood maps in the simulation library will be evident where the observed and forecast discharge are similar, but the flood map skill score is low. Often these inaccuracies relate to the DTM and local infrastructure such as roads, embankments, bridges or dams and their impact on the hydraulic modelling. We will expand the discussion (Section 4.2) of how this validation can be related to physical processes.

Figure 9 could be used by model developers or researchers aiming to improve the flood forecasting system. It could also be used operationally by flood forecasters to summarise the full spatial detail in the ensemble forecast and to communicate the uncertainty in the flood extent forecast to decision makers. We will add these details to the discussion section.

7. The accuracy of the derived SAR layer should be evaluated carefully, and its quality demonstrated to the reader. If this ‘observed’ layer is poor, the case study is not useful.

The SAR-derived flood map accuracy is an assumption made in this study. We will add this statement to the data section. The closest available (cloud free) optical image was a Sentinel-2 image on the 17 August 2017, 5 days after the SAR image. During this time the flood waters had receded from their peak which makes this unsuitable for comparison with the Sentinel-1 image.

Since October 2021, Sentinel-1 SAR images are processed by CEMS GFM and this product provides an uncertainty estimate with the derived flood extent. This will allow the observation uncertainty information to be used in conjunction with the new scores for proper interpretation in future.