Response to Reviewer comments on Hooker et al., "A new skill score for ensemble flood maps: assessing spatial spread-skill with remote sensing observations"

Many thanks to both reviewers for your useful feedback and time reviewing our paper. Our responses to numbered comments in black are detailed below in *blue italics.* We have listed the original line numbers from the reviewers' minor comments. We include a marked-up copy of the revised manuscript where text removed is in strikeout font and additions are underlined.

Response to RC1 comments

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

Section 2.1 and 2.2 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. These Sections have been amended following detailed minor comments from RC1 (see below) so that they now read differently to 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.

Please find our changes and responses to minor comments detailing these issues below.

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. We have added the following to Section 3.2 to clarify: 'GloFAS outputs a gridded (approximately 10 km spatial resolution) ensemble forecast of river streamflow. Each of the GloFAS grid cells are linked to sub-catchments in the Flood Foresight system. The simulation library flood maps are selected when the forecast streamflow exceeds a return period threshold level. Each ensemble member flood map forecast is created by aggregating the individual sub-catchment maps.'

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.

In Section 3.3 we explain that 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 flooded 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. All mapped Figures in the paper now include a permanent water body layer.

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 (for single 'deterministic' forecasts). We have added text to the introduction (L91) to make this clear 'where comparisons were made against conventional binary performance measures'.

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?

Added to the conclusion, paragraph 3. 'We found that one ensemble member outperformed all others in a region close to a confluence zone and nearby observed heavy rainfall. The region correlates to an area of under-spread ensemble members indicating that not enough members were predicting flooding here. Future studies could investigate the physical processes further using the methods presented here. The ensemble flood map spatial spread-skill could be investigated in the context of a particular physical process (such as rainfall intensity/location or an improved aspect of the hydrological model such as antecedent soil moisture) and how these uncertainties translate to the probabilistic flood map forecast'.

To expand on the uses for Figure 9 we add 'The SSS map summarises the whole ensemble, which makes it useful for forecasters attempting to convey uncertainty information to decision makers, highlighting regions where there is high/low confidence in the forecast.'

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.

Added to Section 3.3

'The closest available (cloud free) optical image was a Sentinel-2 image on the 17th August 2017, 5 days after the SAR image acquisition. During this time the flood waters had receded from their peak, which makes this unsuitable for comparison with the SAR-derived flood map. Since no other validation sources are available, for the purposes of this study we assume that the SAR-derived observation of flooding represents the true flooding extent. From October 2021, Sentinel-1 SAR images are processed by CEMS GFM (GFM, 2021) to derive flooding extent and provide an uncertainty estimate of the grid cell classification. This means uncertainty information in the SAR-derived flood map could be accounted for in future evaluation studies.'

Response to minor comments in supplementary attachment

We provide specific responses to many of the minor comments below. We have completed the rest of the minor textual changes suggested by the reviewer as shown in the marked-up manuscript.

1. Title. your proposed metric is agnostic to the 'observed' source. For example, a validated hydrodynamic model could be used. We have not made this change. *In our opinion, forecast systems should be validated against independent observations. In the case of flooding extent, remotely observed observations (open access and freely available) are the only source currently available.*
2. L7. is this flood forecasting? unclear. *The application, ensemble flood forecasting, is mentioned in the same sentence.*
3. L9. is this the 'spatial spread-skill' you're computing? or just a generic term for what you're computing? Unclear. *Changed to 'This determines a skilful scale (agreement scale) of ensemble performance by locally computing a skill metric across a range of length scales.'*
4. L20. This meta paper is not focused on the effects of false alarms or missed warnings. please advise which sections of the paper support your claim. *Reference changed to Arnal et al., 2020 (support in Arnal et al. introduction)*
5. L29. all of these address some uncertainty somewhere in the model chain? I'm sure this is true.. but this sentence isn't useful to the reader. Which publication address which part of the chain? revise to make the sentence more useful or remove. *We include a range of previous work that detail uncertainties in flood forecasting systems (often multiple types included) so that the reader may refer to these for additional details.*
6. L30. how does this sentence support your work? are you just trying to say there is lots of uncertainty in lots of places? revise to include some useful information from Boelee et. al. (e.g., their findings) or remove. *Changed to 'As discussed by Boelee et. al., these uncertainties include those...'*
7. L35. you mean predictiveness (predictive and predictable have different meanings). *This terminology (spatial predictability) has been used in previous literature (e.g., Dey et al., 2016).*
8. L54. define acronym. *ICEYE is the name of the satellite, after further investigation, to the best of our knowledge this is not an acronym.*

9. L56. what is this? *This is one mode (of several) of the SAR satellite, more details can be found on the website referred to.*
10. L65. replace this with a sentence that describes the strengths/weaknesses of the two approaches. *Changed to 'In contrast, here we consider ensemble spatial verification at a single time point'*
11. L68. please explain this better. *Changed to 'A perfect ensemble should encompass forecast uncertainties such that the ensemble spread is correlated to the RMSE of the forecast (Hopson et al., 2014)'.*
12. L70. these claims need a reference. *Added reference Galmiche et al. 2021.*
13. L77. make this a separate sentence and clarify. *Changed to 'When mapping the flood extent prediction, the ensemble mean field alone does not retain the spatial detail of the individual member forecasts.'*
14. L79. Rephrase. *Changed to 'The spatial spread-skill of the ensemble forecast is determined by evaluating the full ensemble against observations of flooding.'*
15. L80. explain and provide citation. *Added reference '(Dey et al. (2014), see Section 2)'.*
16. L83. add context. why are you presenting this work? *Reasoning follows from previous sentence stating that the spatial spread-skill has not received research attention in flood forecasting. Changed to 'However, previous work in numerical weather prediction by…'*
17. L113. this is the new metric you are presenting? revise to make this more clear at the beginning of the paragraph. *Added 'In this Section we present new methods for evaluating and visualising the spatial-spread skill of an ensemble flood map forecast.'*
18. L117. what is this? add some explanation (don't expect the reader to jump ahead to understand something in the intro.) *Added 'such as a combined ensemble or the ensemble median'.*
19. Section 2.1 and 2.2. *Please see main comment (1) response.*
20. L156. Why the change in notation from the previous section? *The notation is changed to D as it is defined differently to FSS and is made clearer now in the revised Section 2.2, L155).*
21. L183. You previously defined 'N' as array dimension. need to be consistent. *Good point. Changed to 'M'.*
22. L212. Difficult on the reader to reference a figure 12pages ahead. restructure so the reader encounters the figure shortly after the first mention. *We prefer to keep the results all together, following the methods Section. Changed to '(an example hexbin plot is presented in Section…'*
23. L220. I suggest removing this section, consolidating into a few sentences, and placing it near the results map. *The SSS map is a crucial aspect of the new method presented in the paper, we prefer to keep this here as part of the methods, rather than with the results.*
24. L233. was there significant levee breaching or avulsion? *Yes, river embankments were damaged in 11 districts (Floodlist, 2017). Added to Section 3.1.*
25. L253. need reference. what does this mean? predicted by who? *Added 'by the South Asian Climate Outlook Forum' and reference (WMO, 2017).*
26. L256. by who? need reference. *Added reference (Central Water Commission, 2023).*

27. L263. Are permanent water bodies filtered from this? *They are, this is detailed on L311.*
28. L297. was this work done by the authors? *Yes.*
29. L297. provide a DOI where the result can be accessed. *Provided in the Code and data availability statement.*
30. L304. what data set is used? From what date? *No additional data set is used. The permanent water bodies are determined from the pre-flood image. Added '...that are detected on the pre-flood image...'.*
31. L313. why? this seems like a major assumption/decision. *See response to RC2 comment (4).*
32. L315. need to give more info on the Flood Foresight setup for your domain. Where are the streamflows calculated? What were they? *See response to main comment (3).*
33. L315. What is the domain of the inundation model? *The domain of the inundation model is global (they are global flood maps). For the application used in this case study the Brahmaputra basin in India and Bangladesh have been mapped to GloFAS grid cells.*

    Added 'Flood Foresight was set-up for the Brahmaputra basin in India and Bangladesh using the simulation library approach to flood mapping described in Section 3.2'.
34. L315. What date is the bathymetry/topography from (this region is super dynamic)? *Added the DSM NEXTMap World DSM (2016) and reference to Section 3.2.*
35. L315. How are levees incorporated? *Levees would be incorporated in the DSM where they are wide enough for detection (DSM is 30 m spatial resolution). Local flood defence data is not included. Added 'Note that the flood maps are undefended'.*
36. Figure 4. how many of these are identical? i.e., pulled from the same JBA model run? *The same model run date/time is used for all ensemble members (the initial conditions vary). The forecast discharge varies across each ensemble member and within each sub-catchment of the domain.*

    Figure 4. It would be nice to label each per JBA model lookup (and provide the counts per lookup). *Unfortunately, this is not possible as there are multiple look-up regions on each flood map.*

    Figure 4. check your 'unflooded' color (white vs. blue). *A closer look shows that the 'white' in the colour bar is light blue as shown on the maps. I think it appears white at a distance because of the surrounding black lines.*

    Figure 4. consider making this a supplement instead. *See response to RC2 minor comment (12).*
37. Figure 5. *Figure 5 has been updated as suggested. Multiple colour scales were tested and we felt this was the best representation of flood probability (a recent visualisation study shows that shades of blue are preferred, e.g. see Boucher et al. 2022 https://doi.org/10.5194/hess-2022-305).*
38. L347. do you mean rather than 'median' and 'all'? If so, I'm not sure I agree. It looks like these aggregate predictions perform better than most member predictions. *Added 'The ens_median and ens_all flood maps outperform the second cluster, however there are individual members with a higher spatial skill score compared to ens_median and ens_all.'*

39. L348. where are these shown/discussed? If I understand the Flood Foresight process, the inundation maps should be nested (larger discharges have more cells flooded)... all cells flooded on a low-discharge map will show as flooded on the high-discharge map. *The ensemble variations can be seen in Figures 4 and 5. Added a reference to Figure 4 in the sentence. '(see Fig. 4 ens_1 compared to ens_median)'.*
40. L362. use a figure call out instead (grid cell location (1100, 250)). *Unchanged. We prefer not to clutter the image with an additional arrow. The coordinates provided tell the reader where to look.*
41. L362. does Flood Foresight consider local rainfall? *The location and intensity of rainfall depends on the ensemble meteorological forecast and will differ between ensemble members.*
42. Figure 6. FSSt? *This is labelled FSS_target so that it is clear to readers of AIC and Figures that this line represents the target without needing to read through the detailed methods to find the meaning of FSS_$_T$.*
43. Figure 7. use consistent terminology. I suggest 'observed inundation'. *We prefer to keep 'correct flooded' here as the most succinct and accurate description of where the forecast accurately predicts the flooding in the observed flood maps. The phrase 'observed inundation' does not capture the same meaning.*
44. L383. I'm confused... Area 1 shows ~80 ensemble/observed agreement (high) and ~0 ensemble/ensemble agreement (low). Either there's an error in your explanation, or the metric is counterintuitive (which is not a good idea). *The metric works in the same way as the general agreement scale, a low/small agreement scale means that they agree within a smaller neighbourhood size.*
45. L383. see comment on caption. need a better way of these. I suggest 'callout 3' or 'arrow 3'. *We prefer to keep these labelled as 'areas' as they refer to a region on the map.*
46. L384. Again, don't expect the reader to remember one sentence from a previous section. Instead, say something like 'ensemble members appear to have high agreement near the confluence mentioned previously'. *Added 'Recall that in this region, most ensemble members did not predict the flooding that occurred with the exception of one ensemble member (ens_1)'.*
47. L386. how? is this just visually determined? if so, this is not obvious to me from fig. 8c. Consider providing some quantitative metric. *Yes, this is a visual impression, the black dashed line is added to the legend. A good idea, we will consider additional quantitative metrics in future work. Added '...visual...'.*
48. Figures 7, 8 and 9. *Permanent water bodies have been added to each of the flood maps along with most other changes suggested. The colours chosen represent the best combinations found for clarity, consistency and colour-blindness considerations. The georeferencing can be inferred from Figures 2 and 5 (added to Figure captions).*
49. L397. did you demonstrate this? what about uncertainty in inundation modelling? *Changed to 'Differences between ensemble members in ensemble forecast flood map systems are mostly driven by initial condition perturbations at the top of the hydro-meteorological forecast chain within the NWP system.' Other uncertainties are mentioned in the following sentence.*

50. L408. how should I interpret this from Figure 8? *Added ‘The hexbin scatter plot summarises the spread-skill relationship so that a trend across the whole domain can be assessed.’*
51. L411. list some here. *Added ‘...such as presenting the ensemble median or other exceedance probability...’.*
52. L412. this seems redundant with the previous paragraph. *Sentences re-ordered.*
53. L420. where did you demonstrate this? *Added ‘The categorical scale maps show...’*
54. L425. This is the third time you use this sentence I believe (which is redundant). But here you do a better job explaining the components. (although 'thresholds' still needs more explanation). *Additions made following comment (3) Section 3.2 explains this in more detail.*
55. code is missing some things useful for publication. *We are unable to publish a full test data set (see data statement in paper).*

JBA has no interest in promoting their Flood Foresight product? *We have checked the NHESS conflict of interest definition and maintain that there is no conflict of interest. JBA provided the data for research purposes alone.*

Response to RC2 comments

1. Details pertaining to the modelling solution used are insufficient and not shown in the workflow figure for some reason.

   Figure 3 has been updated to include additional details in the Flood Foresight workflow process. See also RC1 Comment (3).

2. As the AEP based map library is described, I struggled to understand how the authors arrived at the number 36 for the maps, or the interpolation done to derive these maps

   Thank you for pointing this out. To make this clearer we have added: ‘Between each adjacent pair of modelled return period maps, five additional intermediate flood maps are created by linear interpolation. An additional five flood maps are also created beneath the lowest return period flood map. This gives, in total, a library of 36 flood maps.’

3. The large tracts of quoted text from Hooker et al., 2022 seem rather unusual if not unacceptable, I would summarize and rephrase with a citation, or at least check that the quoted text stays under 10% of the manuscript and reformat correctly. Please consult the APA/MLA guidelines for verbatim quotes >40 words as they should be formatted as block quotes and page numbers from the original article must be provided which currently is not the case. https://research.wou.edu/apa/apa-block-quote

   See RC1 comment (1).

4. The authors choose splines to aggregate the observation data to the model scale, this is strange in my view. Spline interpolation is typically used when we have only a few measured points based on which the underlying random field must be estimated. What the authors try to do here is simply upscaling, thus, the use of splines is not really justified to me. In fact when upscaling one aggregates not interpolates and thus I would expect to see averaging or majority or max or other aggregation techniques or some justification as to why this was chosen.

   This was coded as 0-order spline interpolation, which is in fact bilinear interpolation (equivalent to averaging for aggregation). Changed to 'average aggregation'.

Response to minor comments in supplementary attachment

1. L21. I would be careful with the implication that flooding is ONLY always caused by attached immediate rainfall and thus can never be predicted. *Added 'fluvial flooding caused by intense or prolonged rainfall'.*
2. L68. ref required. *Added reference Hopson, 2014.*
3. L258. Show location on Figure 2? *Figure 2 updated.*
4. L269. could you specify which simplification as there are specific terms for these simplifications in hydraulics, e.g. kinematic/diffusive wave, or inertial. *Added 'diffusion wave approximation'.*
5. L271. what does it mean (with rapid 2D flood spreading). *Added '(created by spreading Normal Depth from upstream to downstream)'*
6. L274. Please see response to main comment 2.
7. L292. What is the double kinematic wave approach? *Added 'which includes bankfull and over bankfull routing'.*
8. L299. I think it is relevant to report the timezone for which this timing is mentioned - note that the acquisition times are typically provided in UTC and would need to be converted to local times, if not done already. *Time zone converted to IST.*
9. L302. I would move the pre-flood image sentence after this one in the interest of clarity. *Sentence moved.*
10. L305. How? Does it use the JRC permanent water layer? worth mentioning IMO. *External permanent water data sources are not used in the HASARD algorithm, 'water' appearing in both the pre-flood and flood image are removed from the flood map by applying thresholding. Added 'by applying a thresholding approach'.*
11. L313. *Please see response to main comment (4).*
12. Figure 4. In the interest of making this figure readable, I suggest to only keep the ens_all, ens_med, and SAR-derived here and move the individual ensemble maps to the appendix. *In operational forecasting practice, it is common to attempt to scan all ensemble member forecasts, usually presented on one page so that quick comparisons can be made. It was a deliberate action to present the flood maps in this way to highlight the difficulty/skill of the forecaster at interpreting an ensemble forecast. Keeping this Figure in this form enables the reader to contrast between Figure 4 and the SSS map (Figure 9), which encompasses the full ensemble information. Added to Section 4.3 'Making comparisons across the different ensemble*

member flood maps in Figure 4 provides a demonstration of these forecasting difficulties.'

13. L332. Do you mean high and low flow? because how would the best and worst performing members be known prior to the model evaluation? *These are known following the FSS evaluation. Added 'determined by the skilful scale calculated in Section 4.1'.*
14. L363. reference to substantiate? *Added reference Flood list, 2017*
15. L397. I don't think this reasoning is necessary in the conclusions. *The authors would prefer to keep this section in the conclusions as it adds context to the findings of the paper.*
16. L407. I saw this implication many times in the paper that this metric is something new but as I understand it was presented in the JoH article from the authors already. I would reword to make it clear that only the application is new. *The spatial-spread skill method for the evaluation of forecast flood maps and the Spatial-Spread Skill (SSS) map are presented here for the first time. Changes made to Section 2 and other minor edits make this distinction clearer.*
17. L426. Are you sure? There are a multitude of papers showing calibration methods using coarse res public satellite data in ungauged regions. I would AT LEAST soften this statement. *Removed 'is currently not possible' added 'are rarely practiced routinely'.*