Response to RC4 comments on "Assessing the spatial spread-skill of ensemble flood maps with remote sensing observations".

Many thanks to reviewer 4 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 new line numbers (clean copy) for additions made. We include a marked-up copy of the revised manuscript where text removed is in strikeout font and additions are underlined.

Overall comments

• Firstly, although it is clear from the abstract that this is a new application of existing methods (apart from the new diagram in Figure 9), this is not clear through the manuscript. For example the start of Sections 2.1 and 2.2 should refer to original work (e.g. Roberts and Lean,2008, Dey et al. 2014, 2016a), Further details of already published applications of the applications of the agreement scale method for statistical spread-skill evaluation should also be referenced in Section 2.4, e.g.
Dey, S. R., Plant, R. S., Roberts, N. M., & Migliorini, S. (2016b). Assessing spatial precipitation uncertainties in a convective-scale ensemble. Quarterly Journal of the Royal Meteorological Society, 142(701), 2935-2948.

An excellent point, references added as suggested.

• There needs to also be more clarity in the differences between the FSS useful/skilful scale (Eq. 4) and agreement scales (Eq. 6). (The former linking directly with the spatial differences between objects e.g. Skok, and Roberts 2018, the latter reflecting a pre-defined "acceptable" bias at different scales).

L188 Added 'Note that the skilful scale determined by the FSS (Section 2.1) differs from the agreement scale defined here. The former links directly with the spatial differences between objects e.g. Skok, and Roberts (2018), whereas the latter reflects a pre-defined "acceptable" bias at different scales.'

• As far as I understand it, the method uses a library of different return period flood maps for sub catchments, with the appropriate map selected for each member of a streamflow ensemble based on the predicted stream flow values. Thus, each sub catchment corresponds to one particular return period threshold. This is not discussed with respect to the spatial results, which seem to relate directly to the sub catchments (e.g. Fig 9). It would be a useful justification and advert for this new method if it could discern these aspects of the forecasting system.

Another very good point.
L444 added 'The areas identified (1, 2 and 3) lie within different sub-catchments, which are linked to different GloFAS grid cells, driving the ensemble flood map selection for each sub-catchment. Inferences can be made about the spread-skill of the driving discharge data at sub-catchment level across the domain.'

• A discussion should be added about the effect of bias on both the FSS and the spatial scales. I agree with a previous reviewers comment on this regarding (then) Figure 4, which was (as far as I understand not addressed by the authors). In particular I don't agree with the assertion that "There is no evidence in the literature to suggest that the FSS score is biased towards overprediction.". In fact the FSS reflects a bias between the fields being compared, which is why many studies use percentile thresholds for FSS calculation. E.g.

Roberts and Lean 2008
..."Figure 3 shows the way the FSS typically varies with neighborhood length n, given a sufficiently large sample. It has a range from 0 to 1. A forecast with perfect skill has a score of 1; a score of 0 means zero skill. Skill is lowest at the grid scale, that is, when the neighborhood is only one grid point and the fractions are binary ones or zeros. As the size of the neighborhood is increased, skill increases until it reaches an asymptote at n = 2N − 1. If there is no bias (an equal number of observed and forecast pixels exceeding the threshold) the asymptotic fractions skill score (AFSS) (FSS at n = 2N − 1) has a value of 1, indicating perfect skill over the whole domain. If there is a bias, then the observed frequency fo (fraction of observed points exceeding the threshold over the domain) is not equal to the model-forecast frequency fM, and from Eqs. (5), (6), and (7) it can be shown that
Equation 8
This descriptor of the bias is useful because it relates the bias to the spatial accuracy of a forecast and is linked to the conventional frequency bias (fo/fM), with the advantage of being less sensitive to biases from small frequencies (AFSS = 0.8 is a factor of 2, AFSS = 0.5 is a factor of 4, and AFSS = 0.2 is a factor of 10 frequency bias)."

Mittermaier, M., Roberts, N., & Thompson, S. A. (2013). A long-term assessment of precipitation forecast skill using the fractions skill score. Meteorological Applications, 20(2), 176-186.
"the use of frequency (percentile) thresholds is recommended because of the implicit bias removal this approach provides, as any rain in a forecast period is treated as 'the event of interest"

Additionally, in Dey et al 2016b (reference suggested above) there is a direct discussion linking Spatial scales to fractional coverage.

Our understanding is that the previous reviewers' comment referred to the FSS score being biased if a particular grid cell was over-predicted in comparison to under-predicted (as discussed in Stephens et al. 2014, who investigated bias in binary performance measures such as the CSI), i.e. a tendency to score higher over the domain if the forecast was over-predicted rather than referring to the background bias of the flooded area (difference in total flood extent between the forecast and observed fields over the domain of interest). In this study percentage thresholds were not used to threshold the data due to the binary nature of the flood extent data.

- This paper would be considerably strengthened by including another case study, or at least another snapshot in time.

We agree that additional applications of the spatial-spread skill methods to other flood events would be an ideal next step to this first presentation of the methods in the flood forecasting field. Unfortunately, additional ensemble forecast flood maps at different lead times were not available for this flood event. However, the principals of the application of the spatial spread-skill methods can be adequately demonstrated using this example.

Specific comments

Introduction - It would strengthen the argument to mention in the introduction other publications applying the FSS method to data other than precipitation, e.g.
Harvey, N. J. and Dacre, H. F.: Spatial evaluation of volcanic ash forecasts using satellite observations, Atmos. Chem. Phys., 16, 861–872, https://doi.org/10.5194/acp-16-861-2016, 2016.
Simecek-Beatty, D., & Lehr, W. J. (2021). Oil spill forecast assessment using Fractions Skill Score. Marine Pollution Bulletin, 164, 112041.
Skok, G., & Hladnik, V. (2018). Verification of gridded wind forecasts in complex alpine terrain: A new wind verification methodology based on the neighborhood approach. Monthly Weather Review, 146(1), 63-75.

These publications refer to the deterministic application of FSS, which is different to ensemble spatial-spread skill applications that would be relevant to discuss in this research article.

Section 2.2. I think there needs to be more justification for the choices of alpha and Slim. Is it coincidental that Slim=80 is chosen the same as Dey et al. 2016? How was this chosen/physically justified? How does this relate to the catchment/sub-catchment size and other physical scales of the catchment?

A good spot. It was a coincidence! In previous work for smaller flood events in the UK, we used a smaller value of Slim.
L174 Added '...The parameter value α indicates an acceptable bias at grid level such that 0 ≤ α ≤ 1. Additional historical forecast data of flood events is not available for the region in this study, so we assume there is no background bias between the forecast and the observations and set α = 0. A fixed maximum scale Slim is predetermined using human judgement considering the physical characteristics of the flood event. The value chosen for Slim depends on the magnitude of the flood extent relative to the size of the sub-catchment. For the case study presented here, we set Slim = 80 (2400 m), which is approximately 1/4 to 1/2 of the sub-catchment widths in the domain '

L 239 "Our new Spatial Spread-Skill (SSS) map" Please rephrase. This map is not new, being published in
Dey SRA. 2016. 'A spatial approach to the analysis of convective-scale ensemble systems', PhD thesis. Department of Meteorology, University of Reading: UK.
http://centaur.reading.ac.uk/65945/

Rephrased as suggested.

I recommend removing Fig 1. As it is published elsewhere and available, and best described in context.

We prefer to keep Figure 1 as following comments from previous reviewers we feel it aids understanding the spatial spread-skill methods.

Fig 5 is fascinating. However, a lot of the detail is too small and largely not discussed. Could only the key members be shown? Or those mentioned specifically in the text? Perhaps the full figure could go into supplementary information?

We are glad that you find Figure 5 fascinating! We prefer to keep it in this form as a demonstration of the difficulty of interpreting ensemble flood map forecasts (akin to postage stamp plots used to show ensemble synoptic charts).

I understand that it has already been published (Hooker et al 2022a), but I find Fig 8 very confusing and somewhat misleading. The colour bar itself suggests (at first appearance) negative Agreement Scales which doesn't make sense. It might be more intuitive to either colour the agreement scales one colour then use e.g. hashing to show the negative areas form the contingency table analysis, or to split into two separate colour bars, 1 blue with the title "over prediction", one red with title "under prediction", both going from 0 to 80?

A valid point. Figure 8 has been updated as suggested.

Could some sub catchments be added to Fig 9 and 10 to aid the interpretation? If not then maybe something like Figure 4 could be included as a subplot of Fig 9 and 10 (same size, area scale) to aid interpretation?

An excellent idea, added a sub-plot as suggested to Figures 9 and 10.

It would help interpretation if a "key" for the spatial scales (or at least Slim) could be added to Figures 8,9 and 10. E.g. a black square of the relevant scale with annotation for grid points and size in km.

A key for Slim has been added as suggested to Figures 8, 9 and 10.

Response to RC3 comments

Note that the code and observation data (with the exception of the forecast flood maps) are shared as stated in the Code and data availability section.