Dear Jim,

Please find below our responses to the three reviewers. We highlight where we have made changes to the resubmitted manuscript considering their comments.

Kind regards,
Shaun Harrigan on behalf of all co-authors

Response to R#1

Anonymous Referee #1 (R#1)'s original text in black *with our response in* blue.

My apologies to the editor and the authors for the lateness of this review. Thanks very much to the authors for comprehensively addressing the first of my objections relating to CRPSS calculated with respect to persistence. It's clear from their explanation and additional figure that this was not an error, and I appreciate the detailed explanation and figure. (We can disagree on whether this result is 'intuitive' or not - in the responses the authors stated that the decline in accuracy of persistence is faster than GloFAS, which is what I (intuitively) assumed would happen, but it's clear this is not the case as skill falls with lead time. But this is basically irrelevant.) On re-reading the description of persistence skill (L234-235) there was one thing that was a little ambiguous to me - if a forecast is issued on Jan 2, where the first forecast value is for Jan 3, is the persistence forecast taken from Jan 1 or Jan 2? (I had assumed Jan 2, but I wasn't sure from their description.) This may be worth clarifying.

We are glad our previous explanation and figure have cleared up your previous concern on the CRPSS.

To avoid any ambiguity in the description of the persistence benchmark forecast we have added how the persistence forecast is calculated for an example forecast issued on 3 January. The following text was added to L235-238 in the revised manuscript:

"For example, for a forecast issued on 3 January at 00UTC, the persistence benchmark forecast is the average river discharge over the 24 h time step from 2 January 00UTC to 3 January 00UTC, and the same value is used as benchmark for all 30 lead times (i.e., 4 January to 2 February)."

Added to this is that the many strengths of the paper remain: GloFAS is a system of international significance and this update is in my view ultimately absolutely worthy of publication. The paper is very clear and exceptionally well presented, and remains a join to read.

We thank the reviewer for their positive words and recommendation.

However, the authors chose not to address my second major comment about reliability (and I note they have the support of the editor in this, so I am probably howling into the wind here). Essentially I disagree with both the authors and the editor that omitting an assessment of reliability is acceptable. To me there are two fundamental reasons to present ensemble predictions (and I'm not alone here, going back at least as far as Krzystofowicz 2001): 1) The predictions are, on average, more accurate than deterministic predictions and 2) the forecasts are more 'honest' because they give a representation of predictive uncertainty.

The authors have clearly and admirably addressed (1) with their analysis of CRPS. The second they have ignored. I note also that Emerton et al. (2018) presented a basic assessment of reliability in their assessment of GloFAS-seasonal - I don't understand why such an analysis could not be performed here. (Though as I stated previously - a summary statistic of PIT values would be more appropriate, though attributes diagrams would be ok.) For the addition of one figure and perhaps a paragraph or two, the authors could (and in my view

should) have addressed this fundamental aspect of ensemble prediction. If the authors were concerned about having too many figures, in my view they could remove Fig 11, which is unlikely to be read in detail and (presumably) can be accessed through the GloFAS portal.

The aim of our paper is not a full-scale evaluation of GloFAS forecast skill for a comprehensive range of forecast aspects. Instead, the primary aim is to present to the community the model components and configuration used to generate operational global-scale forecasts and reforecasts. A secondary aim is to establish the science underpinning a new 'headline skill score' that is made available to users on the GloFAS Web Map Viewer (i.e. Figure 9). Our choice of CRPS as the metric was based on Pappenberger et al. (2015), among many others, who state that the most well-known overall summary metric used for a 'headline score' in operational ensemble forecasting is the CRPS. We also note that the CRPS, as an overall metric, does consider reliability implicitly; the CRPS is penalised for ensemble forecasts with lower reliability compared with forecasts with higher reliability (Hersbach, 2000). To comprehensively attribute which forecast aspects (there are many other than reliability) are leading to higher or lower overall forecast skill at locations under different hydroclimate regimes, times of year, etc., is a stand-alone paper itself.

Therefore, we strongly believe such an analysis does not fit into the current scope of our paper and standby our original comment. Future work to expand the number of evaluation metrics offered through GloFAS is underway, and we fully agree with and thank you for your suggestion to ensure reliability information is part of this work going forward, but we defend our current choice of CPRS (and hence CRPSS) as the GloFAS headline skill score in this first step.

References

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Wea. Forecasting, 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, Journal of Hydrology, 522, 697–713, https://doi.org/10.1016/j.jhydrol.2015.01.024, 2015.

Response to R#4

Anonymous Referee #4 (R#4)'s original text in black *with our response in* blue.

This is a review of "Daily ensemble river discharge reforecasts and real-time forecasts from the operational Global Flood Awareness System". I was asked to review this paper following a first round of revisions and have had the opportunity to evaluate the author responses to previous reviewers' comments. I think the authors have provided a thorough and convincing revision to these comments in the submitted manuscript.

As for my own personal review: I have read the paper and I think that this paper should be published promptly. This advancement seems to me as a fantastic contribution to the hydrological sciences community. I applaud the endeavor and think that HESS is a suitable journal for this type of contribution. I think the science (GloFAS model evaluation methods, figures and description of available forecast datasets) is sound and the end-product will be useful to many researchers. The structure is clear, concise and the text very well written.

We thank the reviewer for taking the time to review our paper, and especially going through previous review comments and our revisions, and for your positive words. Much appreciated.

I have but a few comments that I think the authors can easily address and which can be handled at the editorial board level.

1- Line 41: Australian --> Australia

Now changed in the revised manuscript

2- Lines 56-57: Here the word "global/globe" is used 3 times in the span of 21 words. I suggest varying a bit as to not distract the reader.

Now modified to: "GloFAS can be used for providing daily assessments of potential upcoming flood events for the whole globe, such a spatio-temporal consistent overview is required by several users" in the revised manuscript.

3- Line 91: Reference to a future date that is actually in the past. Actually, most reforecast dates, IFS model cycles, versions, etc. need to be updated. I acknowledge that this is due to the review process, however I just want to highlight that this should be updated prior to publication.

We have now been through the manuscript and have updated all necessary dates, links, and references to model versions such as the IFS that have advances since the paper was first submitted. Note that the reforecast dates are the same.

4- General question: GloFAS-ERA5 uses ERA5 since it is updated near-real-time. However, ERA5-Land will soon also be near-real-time and will have the same native resolution as GloFAS (although the authors are better placed than I am to talk about this!). I wonder if there is any plan to update GloFAS-ERA5 to GloFAS-ERA5Land once it becomes near-real-time? If so, perhaps add a sentence on this in the concluding remarks or discussion?

Yes indeed, we have already assessed the potential gains in moving to ERA5-Land and if you are interested have an initial experiment for river discharge published within Muñoz-Sabater et al. (2021) whereby we benchmark GloFAS-ERA5 against GloFAS-ERA5-Land at 1285 river discharge stations – spoiler is that ERA5-Land does show moderate improvements over ERA5 in the majority of catchments.

Reference

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, 1–50, https://doi.org/10.5194/essd-2021-82, 2021.

Congratulations once again on the, to me, excellent paper.

Thank you very much again.

Response to R#5

Anonymous Referee #5 (R#5)'s original text in black *with our initial response in* blue*.*

--- Summary

The manuscript by Harrigan et al. was in its second round of reviews when I first reviewed it. The paper is of high quality and will be undeniably valuable to future users of the openly available GloFAS hindcasts. My

review takes into account the replies of the authors to the previous reviews, replies which I found thorough and well supported. Nevertheless, after reading the manuscript, I still had questions similar to some of the comments previously raised, and therefore think some of these well justified choices deserve to be mentioned, even briefly, in the manuscript. Hereafter, I list some recommendations for improving explanations, some minor points, as well as a concern about Figure 5.

We thank the reviewer for taking the time to review our paper, and especially going through previous review comments and our revisions, and for your positive words. Much appreciated.

--- General comments

The authors have replied to comments made by previous reviewers on the subject, but I believe there should be in the paper an explanation for why the evaluation of a flood awareness system is not based on its capacity to forecast floods.

We stand by our decision to choose the CRPS (and CRPSS) as choice of headline skill score, it is the most well-known and scientifically validated in the operational ensemble forecasting literature, as we highlight in L248-250. It was critical that we have such a headline skill score that can be used to track overall baseline ensemble skill against persistence and climatology. This is essential information for forecasters to understand the fundamental scientific quality of GloFAS in their area of interest. It must also be noted that GloFAS is not only just used for flood forecasting, the ensemble forecasts cover the full flow regime and GloFAS has seasonal products for both high and low flow (https://confluence.ecmwf.int/display/COPSRV/GloFAS+Hydrological+Products+Overview). As mentioned in Sect. 4.4, this is just assessment of overall forecast skill and we highlight (L417-423) that future work should indeed look at other aspects of forecast quality, such as performance during extremes.

I wish the authors would better explain their choice of changing benchmark depending on the lead time, the reasoning behind the choice of the 0.5 threshold, and the expected influence on the results presented in Sections 4.2.2 and 4.2.3. For instance, why not choose the toughest benchmark to beat, designing for each catchment a benchmark based on both persistence and climatology that would resemble a locally optimized system switching from persistence to climatology when the CRPS of persistence and climatology cross?

There is some guidance in the literature on which is the most appropriate benchmark to use, depending on the lead time. We cite Pappenberger et al. (2015) in L232-236 who advise persistence benchmarks for short and medium range lead times and climatology benchmarks for longer lead times. GloFAS covers all three ranges from short- (1-3 days), medium (4-15 days) and extended (from 15-30 days), therefore there is no single most appropriate benchmark for all lead times assessed. There was a lack of guidance in the literature on out to exactly which lead time is persistence the most appropriate benchmark, and when climatology should be used from for global scale forecast models. This is exactly the analysis presented in Figure 5. We produce the experiment against both persistence and climatology for all lead times, and so can find the lead time, when averaged for all stations across the globe, at which persistence and climatology should be used. This result then informs the choice of benchmark in Section 4.2.2 and 4.2.3, which we believe is well justified from Figure 5.

We thought it best in Figure 6 and Figure 7 to not blend benchmarks when assessing the spatial distribution of skill. Essentially, adding a blended benchmark comprised of both persistence and climatology would add an additional degree of freedom in terms of interpreting the results. For example, is the GloFAS forecast skill different between two catchments because the one uses climatology and one uses persistence?

However, for the purpose of the new headline skill score layer on the Web Map Viewer (i.e. Section 4.2 & Figure 9), we do consider both persistence or climatology independent of lead time. If for example the CRPSS

against a climatology forecast drops below the 0.5 'high skill' headline score threshold at a 2 day lead time, and skill against persistence remains above, then we define the station is only 'highly skilful' out to 2-days.

Regarding the choice of CRPSS = 0.5 for our threshold, we categorise forecasts above this threshold as 'highly skilful' compared to the benchmark (i.e. GloFAS is 50% more accurate than the benchmark). There is no 'correct' choice of threshold, it is arbitrary, but if too low then on the map (Fig. 9) all stations would appear 'highly skilful' and would not be useful for users to see quickly at a global view, which regions tend to have higher or lower skill. Once a user clicks on an individual station then we show the raw CRPSS and CRPS for against both benchmarks, thus giving users all the information to make their own assessments.

I also have some concerns about the maps and some statements presented in Section 4.2.2. More specifically, to which extent does your choice of benchmark influence these spatial patterns? You chose to select the benchmark based on the lead time. Therefore, when looking at spatial skill patterns for lead times shorter than 15 days, won't the highest skills tend to be found in catchments where persistence is not adequate, i.e. catchments where seasonality dominates (hence the low skill in polar latitudes), rather than in places where GloFAS hindcasts are indeed of good quality? Reversely, at lead times longer than 15 days, the highest skills should be found in catchments where serial correlation dominates.

See our comments above, you are exactly right: but we think blending the benchmarks would make it even more difficult to understand the differences in skill (is it due to poorer forecast skill, strong serial correlation OR a climatologically stable location).

You mention L.294-295 that "The regions of highest skill are similar to those for short- and medium-range", but this is not obvious when comparing areas without skill in Figure 6d and Figure 7a and it seems to me that what we observe are artifacts from this strict split in benchmark at 15 days lead. For instance, parts of Russia go from negative skills at lead time 10 to positive skills at lead time 15 days. This is counter-intuitive, and I think a comment from the authors would be necessary.

This is an interesting point and yes is due to the choice of benchmark. We fully agree this deserved a comment in the manuscript, thank you for highlighting this fascinating example. We have therefore added the following paragraph to L305-311 in the revised manuscript:

"The choice of benchmark forecast used for short- to medium-range (i.e., Figure 6) and extended-range (i.e., Figure 7) maps was based on the global median of all stations in Figure 5a. However, there is spatial variability in the choice of best benchmark according to unique hydroclimate properties. For example, in northern latitudes around Russia and northern Scandinavia GloFAS was shown to be negatively skilful against persistence at a 10-day lead time (Figure 6d). However, GloFAS is shown to be skilful against a climatology benchmark in the same region at lead time 15 days. This shows that persistence is a much tougher benchmark to beat in these catchments compared to climatology, likely due to the high degree of serial correlation from snow processes"

Finally and in line with these comments, I question the use of "should be" L.432-433 in the conclusion: "The analysis shows that on average across the world forecasts should be benchmarked against persistence for lead times up to about two weeks and against climatology for longer lead times." It currently sounds as a recommendation, but it seems that this analysis should be performed and fine-tuned for use at non-global scale.

We agree that this sentence sounds like a recommendation and is not the main intention of the analysis, we have therefore removed this sentence.

In several of your replies to R2, you mentioned that a novelty from this work was that the reforecasts, along with other data types, are open. Opportunities for further research based on this data are well mentioned, but I expected a deeper discussion on the incentives for opening up and on encountered challenges. I believe your experience may be valuable for other data providers choosing to switch to open access, and highlighting overcome challenges would only highlight the value of this work.

We had not considered this point but are very happy to share our experience if it could be useful for other groups. We have therefore added the following paragraph to Section 4.3 L411-427 in the revised manuscript:

"While producing large sets of reforecasts and providing data free and open to the community has many benefits, it comes with challenges and key considerations. One of the main considerations is the data storage and delivery infrastructure. A full set of 20-year GloFAS reforecasts is ~23 TB in size. For each new major model upgrade a new set of reforecasts are generated, together with ~35 GB of raw data generated every day for the real-time forecast stream. It is clear that the size of data is a barrier for many users to use. Most users do not require the full temporal range of data and are usually interested in a sub-domain, for example their study region or country. It is simply not practical for every user to download ~23 TB of global data to their computing infrastructure if they only want data for their individual catchment, not to mention if a standard laptop is the only computer available to them. Our solution was to store GloFAS data on the ECMWF Meteorological Archival and Retrieval System (MARS; https://confluence.ecmwf.int/display/UDOC/MARS+user+documentation, last accessed: 25 March 2022) – MARS offers the functionality for users to choose temporal and/or spatial subsets (among others) and the heavy data handling and computation happens on ECMWF infrastructure so the user can download a smaller and more manageable subset of data. The CDS is the public facing front end for users to access GloFAS data and metadata, and communicates with MARS in the backend. A further consideration is producing sufficient documentation for users to interact with the data and provision of a support service whereby users can get in contact with GloFAS data and domain experts for queries: https://confluence.ecmwf.int/site/support (last accessed: 25 March 2022)."

There seems to be an inconsistency between Figures 5a and 5b. It is unclear why the intersection between the median CRPS of persistence and the median CRPS of climatology (~18 days) does not correspond to the intersection of the median CRPSS (~14 days). I could not find any reasonable reason why this would happen.

They are not expected to be the same. In Fig. 5b the median CRPS of persistence (blue line) and median CRPS of climatology (red) line cross at day 18, but this does not factor in the CRPS of the GloFAS forecast (black line), whereas in Fig. 5a, the **CRPSS** is expressed using the CRPS of GloFAS AND both persistence and climatology benchmarks.

--- Minor comments

L.14 Can we really say that global-scale hydrological forecast systems are widely used ?

Have removed "widely" in the revised manuscript.

L. 80-81 'reforecasts' (as ?) consistent as possible

Have added 'as' in the revised manuscript.

L.92 'of with' seems strange

Changed to "with increased data access availability" in the revised manuscript.

L.140-143 The parenthesis count is off.

Now fixed.

L.165 The second incentive for using hindcasts is not as universal as incentives 1) and 3). In fact, the reader only links IFS updates and influence on the hindcasts in the next paragraph. I suggest moving incentive 2) later in the text, and use this information not as a common incentive to using hindcasts, but rather as additional information on the choice of 2019 as reference year for the system versions.

Done.

Figure 3 For the sake of showing a self-explanatory figure, I suggest explaining « ens=11 » in the caption or explicitly writing « 11 members ».

Updated the caption as suggested.

Figure 4 KGESS should be briefly explained when presenting this figure. The authors offered a short and clear explanation in their reply to previous comments which could very well fit in the paper and ease understanding.

We realised we did not add the abbreviation of KGESS, now added. You can see where the KGESS is explained from L199-204 in the revised manuscript.

L.271-272 “At day 15…” This sentence is unclear.

Now reads “At day 15 the CRPSS is…” in the revised manuscript.

L.276 “The aid the readers…” seems incorrect. Did you mean “To aid the readers…”?

Yes, thank you. Now corrected.

L.287 “across at” sounds strange.

“at” has been removed.

Figure 8a Please add an explanation in the caption for the red and blue-shaded areas.

Now reads: “…at 5997 diagnostic river points by degree latitude of the river point with polar (tropics) climate region shaded in blue (red) (a)”

L. 352 Remove “be”

Done.

L.434 “… results show...”

Done.