Initial response to RC1

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

Summary
This paper describes the most recent version of the GloFAS ensemble streamflow forecasting system. While there are no major advanced in methods used to generate forecasts, GloFAS is a system of international significance, and highly relevant to readers of HESS. The manuscript is well structured, admirably clear and succinct, and was a pleasure to read. Figures are well presented, and while references are sparse (especially in the introduction), as this paper is essentially focused on presenting an operational system this is ok. As the authors note, a major development is the availability of GloFAS forecast outputs in near-real time, and this is well-explained and documented.

I therefore believe the study ultimately deserves publication. I nonetheless had two major issues with this paper, listed below. I therefore recommend the paper be revised before it can be published.

We thank the reviewer for their positive words about our manuscript and constructive comments below.

Major comments
1) There appeared to me to be an error in the calculation of CRPSS with respect to (wrt) persistence - see specific comments below. If this is not due to an error, I would like the authors to explain what to me were counterintuitive results.

We provide a detailed explanation of the pattern of CRPSS wrt persistence below at your specific comment and show the results are as expected.

2) The authors earmark the assessment of reliability to future work. I do not think this is good enough, given 1) that reliability is a key attribute - in my view at least as important as skill - of ensemble forecasts and 2) their statement in the introduction that "not also having direct access to the raw data precludes the use in further downstream applications (e.g. impact modelling, multi-model forecast systems, production of value-added products for specific sectors such as river transport and hydropower industries, and advancement in techniques requiring large-scale datasets such as machine learning)." This statement implies that the authors expect the outputs in the ways specified - i.e. as direct inputs to impact assessment models of some kind or other. In my experience such models very often require reliable ensembles wrt to observations (or at least unbiased ensembles) as inputs. As GloFAS does not treat hydrological uncertainty, it is highly likely that ensembles are overconfident, particularly at short lead times (e.g. Bennett et al. 2014). I think this is information that users of these outputs, and therefore readers of this paper, would want to know. I therefore would like to see the authors present an assessment of reliability as well as skill, and the ramifications of this assessment discussed. Given the forecasts are likely to be treated as continuous variables in impact models, I suggest using probability integral transforms (PIT, e.g. Gneiting and Katzfuss 2014) to assess reliability (noting the need to generate 'pseudo'-PIT values in cases where streamflow observations can equal zero). If the authors prefer, PIT values can then be summarised with either the alpha-index (Renard et al. 2010) or the beta-score (Keller et al. 2011) (whichever is more suitable) for presentation in plots similar to Figure 5 or 6.

We agree that reliability is indeed an important aspect of hydrological forecast quality, but there are many important aspects of forecast quality relevant for GloFAS users (for example, forecast skill for extreme events). The focus of this paper (as outlined in L70-72) is to provide a detailed description of how the GloFAS forecast datasets (both real-time and reforecasts) are generated and present a first global assessment of ensemble forecast skill against two of the most common benchmarks in the hydrological literature; thus defining the new GloFAS "headline skill score" that is added as a new layer on the GloFAS web interface. We do not claim that this first evaluation covers all forecast quality aspects, nor do we believe this specific paper is the right

place to add the evaluation of just reliability and not other aspects that might be important for the diverse range of users. By providing the large-sample reforecasts dataset openly, we strongly encourage users of GloFAS foreacsts to conduct their own specific and local evaluation. We see this paper as the first step, and as part of the ongoing GloFAS evolution will expand global-scale evaluation efforts to other forecast quality aspects. Thank you for your suggestion on the method for evaluation of reliability, we will certainly take this into consideration going forward.

Specific comments
L88-97 Please provide the model time step at some point in this paragraph.

The ECMWF ENS is run at a 6-hourly forecast time step and for ingestion into the GloFAS hydrological modelling chain, data from the 00 UTC run is extracted and aggregated to 24-houly time step. This information will be added to the end of this paragraph in the resubmitted manuscript.

L125 "https://www.globalfloods.eu/" the hyperlink associated with this text 1) differs from the text and 2) returns a 404 error.

Thank you for noticing this. The hyperlink should have pointed to the main GloFAS web site: "https://www.globalfloods.eu/" and will be corrected in the resubmitted manuscript.

L250 Figure 5. To me, there's something very counterintuitive (and perhaps erroneous?) about the persistence skill plot. The accuracy of persistence (the benchmark, and the demoninator in eq 1) is often very high at short lead times and then declines with lead time - often rapidly. In my experience, this decline is usually much faster than the decline in the accuracy of forecasts. So I would expect CRPSS wrt to persistence to be very low perhaps even close to 0 - at very short lead times, and then to rise with lead time. But Fig 5 shows the opposite of these trends - i.e. CRPSS wrt persistence starts high and falls with lead time. I can't see how this can occur without a calculation error - though perhaps I've missed something? Even if this is not due to an error, these results at least requires some discussion/explanation. CRPSS calculated wrt to climatology looks sensible to me, which makes the persistence results even more puzzling.

We do not think the results are counterintuitive, but agree it is very useful to have access to the individual components of the skill score equation 1 (i.e. $CRPS_{fc}$ and $CRPS_{bench}$) to better interpret the accuracy of the GloFAS forecasts *relative* to the accuracy of both persistence and climatology benchmark forecasts, as a function of lead time. We therefore show in Fig. A (below) the individual CRPS accuracy components in equation 1 (i.e. CRPS of GloFAS forecasts (black line), persistence benchmark (blue line) and climatology (red line)) as a median across n=5997 river points. This plot helps interpret the global median CRPSS results (solid lines) presented in Figure 5 in the original manuscript. While the CRPSS in Fig. 5 is dimensionless with an optimum value of 1, the CRPS error is measured in units of the variable being evaluated (here m^3/s) and so has an optimum of 0 (i.e. perfect accuracy against the reanalysis).

It is clear from Fig. A and consistent with your comment: the accuracy of persistence (blue line) is highest at short lead times then gets rapidly less accurate as lead time increases. But it is also the case that the accuracy of GloFAS forecasts (black line) are highest at short lead times, and get less accurate as lead time increases. The decline of accuracy of persistence is however faster than the decline of the accuracy of GloFAS forecasts; this is also consistent with your comment. The persistence benchmark is very simple (use the reanalysis river discharge value from day before the forecast for all lead times), whereas the GloFAS forecast includes both information on the initial conditions as well as meteorological forecast information. Therefore, the accuracy of the GloFAS forecasts is (on average) higher than persistence, even for short lead times. In our opinion this is intuitive, and it would be more surprising if the considerably more sophisticated GloFAS forecasts were only as accurate or marginally more accurate than a simple persistence forecast, including at short lead times. Further, the core meteorological variables that drive GloFAS forecasts (e.g. precipitation and temperature) are

most accurate at short lead times (Haiden et al., 2019). We also include the CRPS for climatology (red line) in Fig. A for completeness.

To help users better interpret the CRPSS for their stations of interest, we will include CRPSS as well as CRPS plots as a clickable "pop out" window as part of the new GloFAS "forecast skill" layer on the GloFAS Web Map Viewer. We will update Sect. 4.2 and Fig. 8 to reflect this change in the resubmitted manuscript.

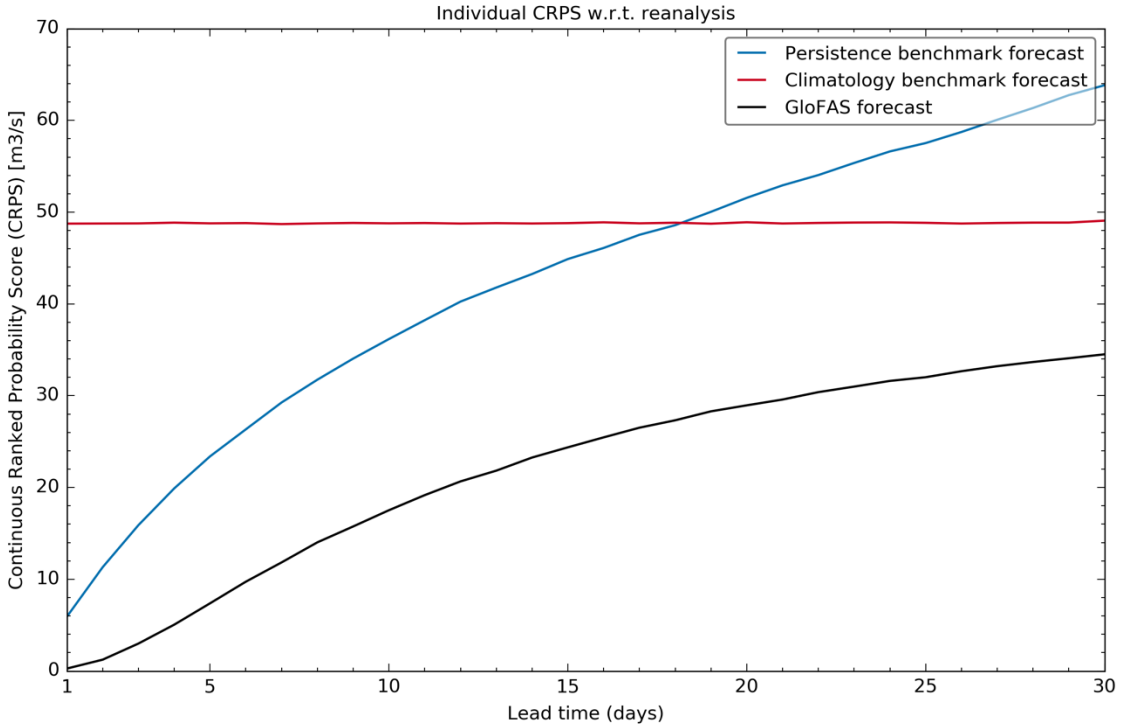

Figure A.: Global median Continuous Ranked Probability Score (CRPS) for GloFAS forecasts (black line) and both persistence (blue line) and climatology (red line) benchmark forecasts from 1- to 30-day lead times with respect to GloFAS-ERA5 river discharge reanalysis across 5997 diagnostic river points.

L271 Figure 6 As with Fig 5, I would expect skill wrt to persistence to rise with lead time, not to fall.

As per the response to comment above, the accuracy of both the persistence benchmark forecast and GloFAS forecasts themselves decrease with lead time, but the accuracy of GloFAS forecasts decrease at a slower rate (Figure A above).

L343-345 "Future work should assess other aspects of forecast quality such as reliability (Robertson et al., 2013), value (Cloke et al., 2017) or performance during extreme events (Bischiniotis et al., 2019)." Not suggesting any change here, but the authors may also like to consider calculating the skill/reliability of accumulated volume forecasts (e.g. accumulated 30-day streamflows), as this may well be of interest to reservoir operators and others. The ability to simply sum streamflows of individual ensemble members over various lead times is a major benefit of ensemble streamflow forecasting systems such as this one (as opposed to probabilistic forecasts generated at discrete lead times).

We currently we use weekly river discharge averages within GloFAS-Seasonal operationally and for seasonal forecast evaluation (see Emerton et al., 2018). However, providing forecast products (and forecast evaluation) at a range of different accumulations is something we will take on board and seek feedback from GloFAS users, thank you for your comment.

Typos/grammar/style

L79 "descripted" - ’decribed’?

This should have been “described” and will be updated in the resubmitted manuscript

L140-141 "see for https://confluence.ecmwf.int/display/COPSRV/01.+GloFAS+operational+system a description" should be "see https://confluence.ecmwf.int/display/COPSRV/01.+GloFAS+operational+system for a description

This will be updated in the resubmitted manuscript.

L152-155 "Twice per week ... as real time (Vitart 2014)." Suggest breaking this long sentence in two at the comma.

We have broken this long sentence into two and will rephrase to the following in the resubmission:

“A reforecast task is run twice per week (on Mondays and Thursdays) in parallel to the real-time forecast, using ERA5 atmospheric reanalysis (Hersbach et al., 2020) for initial conditions of past dates. A reforecast of the corresponding date for the previous 20 years is produced with a reduced number of 11 ensemble members but using the same model version as real-time (Vitart, 2014).”

Thank you again for your insight and we appreciate your time to review our manuscript,

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

References

Emerton, R., Zsoter, E., Arnal, L., Cloke, H. L., Muraro, D., Prudhomme, C., Stephens, E. M., Salamon, P. and Pappenberger, F.: Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1.0, Geoscientific Model Development, 11(8), 3327–3346, doi:https://doi.org/10.5194/gmd-11-3327-2018, 2018.

Haiden, T., Janousek, M., Vitart, F., Ferranti, L., Prates, F. and Prates, F.: Evaluation of ECMWF forecasts, including the 2019 upgrade, 2019.