Initial response to RC3

Anonymous Referee #3 (R#3)'s original text in black *with our initial response in* blue*.*

This paper described the development, application and user service upgrade of GLOFAS to include reforecasts and reforecast-based skill calculations. It probably contains more engineering-related detail than is typical in a HESS paper, but it does advance the science as well in providing a working example of the applied science concept that reforecasts can help to usefully complement the information available from forecasts alone – thus I think it is appropriate for HESS. It would be nice to see a bit more framing on where a forecast effort such as GLOFAS fits in overall field of hydrologic forecasting, but the authors may deem that after some years of running GLOFAS, it is now a widely understood / accepted approach, and that the overall GLOFAS rationale is adequately addressed in prior papers. I don't actually think it is as well understood outside of Europe, where EFAS served as an introduction to this type of service. Partly for that reason, I'd suggest that the authors do more to highlight some expected limitations of GLOFAS relative to a local/regional forecast (suggestions below), particularly given the use of perfect model benchmarks. On the plus side, the paper could strengthen the context framing by noting that it represents one of the first large operational scale effort at reforecasting in hydrology, in contrast to the introduction of reforecasting more broadly for weather and climate over 10 years ago. Overall, however, I find the paper to be a high-quality, very readable effort, presenting a range of accessible and useful information, hence I recommend publication with relatively minor clarifications and adjustments listed below.

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

We agree very much with you on the need to frame the justification/reason for a system such as GloFAS in the context of local systems and within the overall field of hydrological forecasting - thanks for the suggestion! GloFAS aims to provide complementary information in addition to, rather to instead of, locally calibrated catchment forecast systems. In the first instance a global system is useful for providing a hydrological forecast in regions where there is currently no operational local system, or when a local system covers only part of a larger, often transboundary basin. Further, there are users that require a global overview of potential upcoming extreme events, such as international disaster and humanitarian agencies. We will expand the introduction of GloFAS in the introduction section to highlight the type of application GloFAS is designed for, and also the limitations relative to a local/regional forecast as you suggest.

Specific comments:
53: 'originally designed for large/transboundary river basins' – this is surprising because those would almost all be regulated and impaired, yet glofas does not represent such effects. could this statement be sharpened? if glofas can't be expected to forecast mainstem flows in such basins accurately, what was glofas designed to do more specifically? eg forecast runoff anomalies in such basins? or smaller tributaries across large basin domains? or natural flow changes and risks?

As mentioned in L114-115, 667 of the largest reservoirs are represented in GloFAS. Nevertheless, the reservoir scheme is of course a simplification of reality and the actual real-time release schedules of individual reservoirs is sensitive information and typically not publicly available. However, to help guide forecasters, the locations of the reservoirs explicitly modelled in GloFAS together with the ratio of reservoir volume to mean annual discharge for all downstream river cells are provided to users as supporting map layers within the GloFAS Web Map Viewer (https://www.globalfloods.eu/). While GloFAS generates raw river discharge magnitudes, the nature of a global-scale system means it must rely openly available datasets and is run typically at coarser resolution than locally calibrated models, thus provides varying degrees of accuracy with significant biases, as documented in Harrigan et al. (2020). Nevertheless, GloFAS forecasts are compared relative to thresholds derived from the same model, and therefore can provide awareness of anomalously high river discharge. For

example, if 80 % of forecasted ensemble members exceeded the 1 in 20-year modelled threshold then this would signify an extreme forecast 'signal' irrespective to any systematic biases in the hydrological model.

58: there are more service-related gaps, I think. at the end of this paragraph could another sentence be added to suggest that 'Other concerns include : : :' where I could imagine that the lack of information about where glofas reflects upstream management effects might be another one. ideally a user could mask out reaches where it's thought that what is shown cannot represent the reality of the river flow, perhaps because it is 50% determined by a reservoir release or major diversion.

We agree that improving the representation of human controls on the hydrology is a key gap in global hydrological modelling and forecasting. We will highlight this more strongly in the revised manuscript and think it would best fit in "Sect. 4.4 Future directions". As per the response to your above comment, there is already layer called "Reservoir impact" within the GloFAS Web Map Viewer (Fig. A). We think this information is critical for forecast users as this layer informs the degree of control by reservoir operations on a hydrological forecast for a particular location, and therefore on the possible resulting increased uncertainty in the published forecasts.

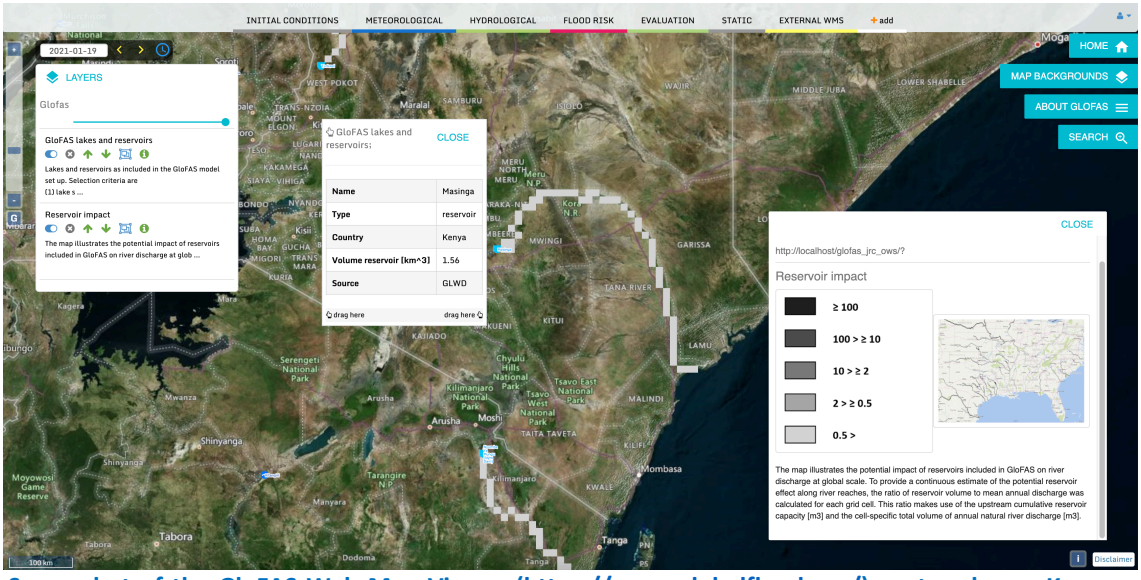

Fig. A.: Screenshot of the GloFAS Web Map Viewer (https://www.globalfloods.eu/) centered over Kenya with the "GloFAS Lakes and reservoirs" and "Reservoir Impact" layers activated, showing the example of the overall potential impact the Masinga reservoir has on the annual natural river discharge or the Tana River.

Fig1: The label 'hydrological model' is interesting because most of the hydrological components of LISFLOOD are not used – really it is the empirical groundwater attenuation and the channel routing of runoff. A more descriptive label for this component might be 'Catchment and channel routing' (whether GW or channel it's all a kind of routing, conceptually). Not a big issue but it may be confusing given that a lot of hydrology (runoff generation, snow accum/melt, et) is done by the LSM. btw, LISFLOOD could conceivably also offer a hillslope routing function (gamma or UH distribution).

Thank you for this useful suggestion, we will change the label to 'Catchment and channel routing' in the resubmitted manuscript.

114: Would be helpful to add a sentence to clarify how these reservoirs are represented and their realism – eg level-pool scheme, fixed rule curve

We will add the following sentence to the resubmitted manuscript: "Reservoir outflow is calculated with a set of four rules depending on the current reservoir filling level (see Burek et al., 2013; Zajac et al., 2017)".

174: Perhaps add a sentence or two here or in the later discussion if the reanalysis latency impacts the skill of real-time forecasts (2-5 days is a lot of lag for a flood forecast!) or means that the skill of the hindcasts may be systematically higher than that of the real-time forecasts, since presumably the reanalysis initialization in the past doesn't have latency issues.

It is practically impossible to create identical forecast initialisation for reforecasts versus real-time forecast systems due to the constraint on availability of operational data streams. In the case of GloFAS the 'fill up' (see Figure 2 in the manuscript) period from the last available GloFAS-ERA5 until real-time is based on the short-range (i.e. 1-day) ECMWF-ENS forecast control member and is only needed for the real-time forecast and not the reforecast. This introduces a small difference in initialisation between the reforecasts and the real-time forecast. As both ERA5 and the control member forecast are based on the ECMWF IFS weather model and data assimilation scheme, this ensures consistency between the datasets. We will add a sentence as you suggest in the discussion to highlight this point.

178: Can you say what determines 'skillful'? eg 5% kge above benchmark in the SS?

Hydrological skill in Harrigan et al. (2020) is determined using the modified Kling-Gupta Efficiency Skill Score (KGESS) against a mean flow benchmark, with KGESS > 0 defined as skilful.

199, 235: Perhaps add a sentence explaining what fraction of these sites are impaired/ unimpaired and what 'synthetic' means in the context of an observation (eg labeled in the table). In general, it should be quite clear in the paper when you are verifying against the real versus perfect model world. The initial description of verification against in situ gage stations may slightly obscure the later default to benchmarking against the reanalysis discharge. Could a sentence be added at 235 to state what is gained/lost in the interpretation of skill through comparing to the perfect model benchmark?

In the Supplementary tables, the label 'synthetic' under the 'Provider' column refers to a GloFAS diagnostic river point where no in situ observed time-series river discharge is available at this station location. We will update these tables with expanded metadata when resubmitted to allow readers to better interpret these column labels.

Your comment regarding what is gained/lost calculating skill from the perfect model benchmark is similar to R#2. This approach is common practice in forecast evaluation, but for the benefit of a broader audience we will outline the justification of this approach in Sect. 3 of the resubmitted manuscript - i.e. that it allows for the prediction range to be determined regardless of systematic hydrological model error and that forecast skill can be calculated for any location in the world.

239: again could you quantify in a sentence what you consider 'skillful'. without going fully statistical (ie significance level of a skill scores X% above 0), what is the rule of thumb used by the authors to consider a forecast 'skillful'?

As defined in L230-231, a CRPSS > 0 shows forecasts are more skilful than the benchmark.

Fig 5: It's curious that the persistence-benchmarked forecast SS is so high at timestep 1. If anything I would expect it to start slightly lower than its peak SS because as you move from longer leads to approach T0 the persistence forecast, in theory at least, approaches zero error, and the persistence and actual forecasts approach each other.

There must be something in practice here that means this is not the case, ie persistence has an offset at T0 that is not present in the actual forecast. Especially in medium to large rivers, much of the time by day 1 the

flow from T0 has not changed much. Is it correct that the actual forecast is 95% better than persistence even at day1? Can/should this behavior be explained in the paper?

Your comments regarding interpretation of the forecast skill relative to the benchmark forecast is similar to that made by R#1. We found it useful to look at the CRPS values of individual components of the CRPSS skill score in Equation 1 to interpret the pattern as a function of lead time (Fig. B). 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 the case that the persistence benchmark forecast is most accurate at a 1-day lead time and gets increasingly less accurate with longer lead times. But it is also the case that the GloFAS forecast is most accurate at a 1-day lead time with accuracy decaying with increasing lead time, but at a slower rate than the persistence benchmark, hence the CRPSS score comparing both CRPS.

To help users better interpret the CRPSS for their stations of interest, we now include CRPSS as well as CRPS plots as a clickable "pop out" windows 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.

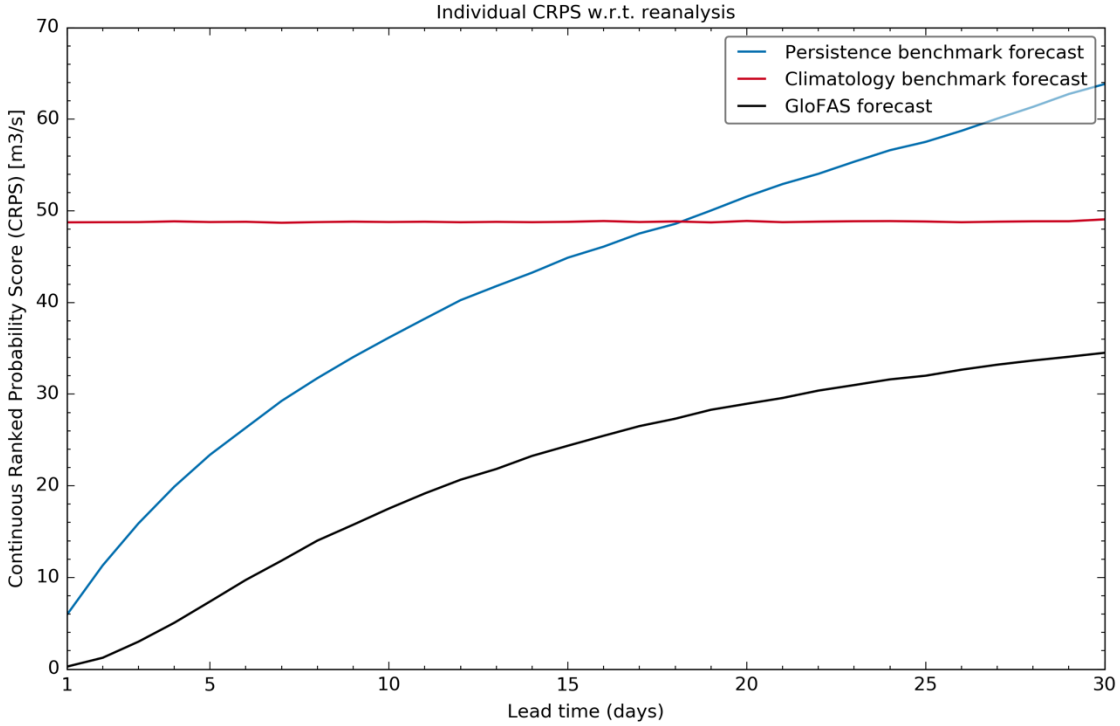

Figure B.: Global median Continuous Ranked Probability Score (CRPS) for GloFAS forecasts (black line) and both persistence (blue line) and climatology (red line) benchmark forecasts from 1- to 30-day lead times with respect to GloFAS-ERA5 river discharge reanalysis across 5997 diagnostic river points.

343: I don't quite agree, as I think of bias & mean error-based scores reflecting only accuracy. The use of an integrated score such as crps means the results are sensitive not just to accuracy but to forecast spread (hence reliability). Perhaps rephrase here?

What we tried to say was that we focused on the overall correspondence between forecasts and the verifying reanalyses, and that there are many other aspects of forecast quality such as reliability (and spread) etc. that could be unpacked in future evaluation work. This could be confused with the technical definition of the CRPS and so we will rephase as you suggest in the resubmitted manuscript.

354: It would be helpful to add a discussion returning to some of the major caveats on the applicability of the analysis. I strongly support the overall message of the paper and commend the effort for having generated reanalyses and used them to demonstrate the kind of information one can derive from them, but I also recognize that in many areas the skill / usability estimates are compromised by use of perfect model benchmarks, and the lack of representation of real-world impairments in GLOFAS. Could the authors walk us through some of the possible limitations, eg in a paragraph, discuss a few cases that users may face in interpreting this kind of skill or using the reforecasts? eg, in the best case, through the model may have some biases it generally represents catchment variability such that the perfect model skill analysis is more or less directly transferrable to real world conditions. A use could infer usability qualitatively or even apply quantitative post processing methods for their own site observations to adjust the skill scores accordingly. In a medium case, the model is badly biased (say magnitudes by 50-100% and seasonal timing off by 30 days), but still represents observed variability in high/low flow conditions – what more would be required of a user then? And in the worst case, say on highly regulated rivers, perhaps glofas cannot be used except in the most extreme situations, eg when there is so much or so little water that management effects are secondary. Would this suggest any other future directions, eg providing guidance or tools on such user-based post-processing and analysis?

This is a very good suggestion to add a further discussion point on the 'limitations' of both the current hydrological forecast model system (e.g. areas with poor hydrological model and forecast skill) and in terms of how the reader/users of GloFAS may interpret the evaluation in light of perfect model/proxy observations. It is very important that reanalyses are not confused with in situ observations and readers are aware of the advantages and disadvantages of our approach. We will add a new section before Sect. 4.4 titled "Limitations" in the resubmitted manuscript.

Your last point is probably one of the biggest challenges in large-scale hydrological forecasting. While we can provide information on whether GloFAS is on average skilful or not for a particular location, this becomes more challenging in areas with very little information on river management and/or a lack of situ observations. However, when faced with an emergency situation due to very extreme hydrometeorological conditions, for example tropical cyclone Idai that devastated Mozambique in March 2019, operational experience has shown that despite the uncertainty associated with GloFAS forecasts, useful information on the future evolution of flood risk can be made, when balanced with many additional sources of information (Emerton et al., 2020).

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

Burek, P., van der Knijff, J. M. and de Roo, A. P. J. D.: LISFLOOD - Distributed Water Balance and Flood Simulation Model - Revised User Manual, Publications Office of the European Union, doi: 10.2788/24719, 2013

Emerton, R., Cloke, H., Ficchi, A., Hawker, L., de Wit, S., Speight, L., Prudhomme, C., Rundell, P., West, R., Neal, J., Cuna, J., Harrigan, S., Titley, H., Magnusson, L., Pappenberger, F., Klingaman, N. and Stephens, E.: Emergency flood bulletins for cyclones Idai and Kenneth: A critical evaluation of the use of global flood forecasts for international humanitarian preparedness and response, Int. J. Disaster Risk Reduct., 101811, doi:10.1016/j.ijdrr.2020.101811, 2020.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H. and Pappenberger, F.: GloFAS-ERA5 operational global river discharge reanalysis 1979–present, Earth Syst. Sci. Data, 12(3), 2043–2060, doi:https://doi.org/10.5194/essd-12-2043-2020, 2020.

Zajac, Z., Revilla-Romero, B., Salamon, P., Burek, P., Hirpa, F. A. and Beck, H.: The impact of lake and reservoir parameterization on global streamflow simulation, J. Hydrol., 548, 552–568, doi:10.1016/j.jhydrol.2017.03.022, 2017