Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-532-RC3, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Daily ensemble river discharge reforecasts and real-time forecasts from the operational Global Flood Awareness System" *by* Shaun Harrigan et al.

Anonymous Referee #3

Received and published: 23 December 2020

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.

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?

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.

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).

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

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.

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

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?

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'?

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?

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?

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?

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-532, 2020.