Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/nhess-2021-25-RC2, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on nhess-2021-25

Anonymous Referee #2

Referee comment on "Leveraging multi-model season-ahead streamflow forecasts to trigger advanced flood preparedness in Peru" by Colin Keating et al., Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-25-RC2, 2021

This paper describes evaluation of seasonal flood forecasts over Peru. The results are a key part of developing forecast-based early warning systems, where a robust understanding of model skill is crucial. The paper addresses an important question, the results are interesting and the manuscript is well structured and clear.

I am happy to recommend publication, after the authors address a few comments, below.

1. False alarm ratio (FAR) is calculated by counting #false alarms and #triggers and dividing the former by the latter. This is fine and is the standard method to do so. However it does mean that the sample is relatively low (particularly for seasonal reforecasts), leading to high uncertainty on the values. e.g. The sample size at Maranon is 19, but there are many fewer flood events and triggers than this. It is possible to partially address with uncertainty ranges on the verification statistics, calculated through a standard bootstrap resampling method (i.e. pick a 19 years with replacement from Maranon, recalculate FAR/ HR/ POD, and repeat). I would like to see this uncertainty added to figure 8, and its implications discussed.

2. In addition to point 1 above. The sample size means there is an inevitable an aspect of forecast behaviour which is not captured in the reforecast - and bootstrap resampling across years is unable to quantify this. To explain with an example: in 2013 the statistical model predicts 94% probability of exceedence, which is observed. In the evaluation this year will always turn up as a 'hit' for any threshold less than 94%. However if we take the probability as a reliable representation of likelihood, there is still a chance that it would have been a false alarm (i.e. 6%). Similarly, there is a chance that every probability which resulted in a trigger was a false alarm (as long as that probability wasn't 100%). This is an unavoidable result of small sample size - and one which bootstrapping will not quantify - so I am not suggesting any change. However I suggest the authors reconsider their conclusion "L483 Detection of additional high flow events is possible by lowering the forecast probability ... while maintaining a low false alarm ratio". This is only true for this particular realisation of the reforecast. If you lower the probability threshold, there will

always be more chance of false alarms when you trigger, by definition. You might get lucky, but then again you might not. It is important to be clear about this otherwise misleading conclusions may be reached, e.g. L427 suggests a lowering of the trigger to 50% may capture many more events, "without additional false positives". This is highly contingent on the particular realisation of the reforecast. A decision-maker may read this paper and decide to take action when the forecast probability is 50%, as they understand that this has an FAR of 0%. But, the chance that action on a forecast of exactly 50% will be in vain is ... 50% (assuming the probability is reliable). So there is a good chance they may be in for a nasty shock! I suggests the authors rethink their advice on lowering the trigger without consequence.

3. The statistical model uses antecedent SST as a predictor (capturing ENSO activity). It also uses a precipitation forecast from the NMME. But what about using the SST forecast from the NMME? If ENSO state is a strong forcing of rainfall/streamflow, then I would imagine that the FMA SST is more strongly related to streamflow than DJF SST? Possibly the precipitation forecast may capture some of this future signal - although precipitation errors are well known. I hope that the authors might consider adding this, as it may increase the skill even further and lead to a better early warning.

4. Can you show the weightings for the statistical model? The results are shown from cross-validation leave-one-out (which is appropriate). But if you built the model again using all years, this would be useful to show the relative importance of each predictor.

5. The GloFAS seasonal forecasts are publicly available on the 10th of every month - not the first, as is stated in L274 (see https://www.globalfloods.eu/technical-information/glofas-seasonal/ - NB they are initialised on the 1st but there is a lag until they are available, which may be where the confusion arises). Does this change the potential for early action, as the first month is almost half over before the GloFAS forecast is available? There are a few possibilities:
- if the action is strictly constrained to the start of the month, the GloFAS run from the previous month is the only available run, so this should be used instead in the comparison
- if it is OK that no forecast is available until the 10th, then the statistical model could (in theory) include additional information on the streamflow/SST/precip in the first few days.
The authors may want to follow either (or neither) of these ideas. But at least please comment on this issue of forecast timeliness in the text.

Minor comments
L41 Was FbA originally applied to droughts? As far as I am aware it is only now being developed for drought/food insecurity. Please clarify.
L69 There is a bit of a logical jump from the previous paragraph, consider adding a linking sentence.
L111 Slightly long sentence, could be split for readability.
L160 What do the colours represent in Figure 1? Satellite image, topography? If the latter then it needs a colorbar.
L200 Table 2: Piura has correlation of 0.84 between J streamflow and FMA streamflow. However in L148 it states that there is no significant monthly autocorrelation in Piura streamflow. This seems to be inconsistent.
L200 Table 2: Maranon GCM precipitation forecast is not included as a predictor,

presumably because the correlation with MAM streamflow is not sufficiently high. I wonder: is this because (a) there is low correlation between seasonal rainfall and seasonal streamflow at Maranon or (b) the GCM precipitation forecast at Maranon is not particularly good? It would be good to include this information. If the answer is (b), then see point 3 above: it may be that SST is a more valuable predictor to take from the GCM forecast.
L226 I am unsure what " n.d." means in this context.
L228 A 3-phase ENSO model is used at Piura, although a 2-phase model does not affect material performance. Given the favouring of parsimonious models (L257), why do you retain the 3-phase model?
L272 Requires some more info on GloFAS: what is the reforecast period, which model version used, has the model been calibrated for these basins (where streamflow data has been shared with the GloFAS team, the model has been calibrated).
L315 It would be useful to explicitly note how many upper tercile events are present for each site.
L399 What is meant by 'observed trigger'? From the context I think it should read 'event'? 'Trigger' only applies in context of the forecast, not the observations (similarly used in L448).
L450 I am not sure what is meant by "TS is maximised".
L463 Another thing to consider with these close-to-threshold events is that the difference in streamflow between may very well be within the margin of observational error - particularly if the seasonal average is based on daily data (i.e. an accumulation of systematic/random errors over 90 days).
L468 "two events of similar magnitude...are likely to produce similar impacts with early actions likely to yield similar benefits". I am not sure it is reasonable to say this. Two seasons with similar average seasonal streamflow may have highly different subseasonal variability. For instance season A: all season just below the overtopping level without breaching, season B: a little way below season A average for the first month, but then increasing and repeatedly flooding in the next two months. A & B may have very similar average streamflow - but very different impacts.