Comment on nhess-2021-25

Anonymous Referee #1

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-RC1, 2021

General comments

The paper under review addresses an important topic within the scope of the journal, is generally well written and structured. Figures are visually appealing (especially Fig.6 & 7). Datasets used are adequate for the purpose of the study. Methods are rather traditional statistics (fairly old-fashioned), mainly a linear regression on principal components, but presumably also quite robust. No non-linear transformations, no unconventional predictors. The multi-model approach mentioned in the title is interesting. The chosen performance metrics for validation are also suitable. According to the authors, the developed model is an improvement to the current operational methods in Peru.

I suggest adding "in Peru" to the title of the manuscript – or any other spatial restriction the authors consider appropriate – as the method was only tested for two rivers in this specific country, and includes predictors that might not be suitable in other areas of the world (e.g. sea surface temperature for ENSO condition). If the authors want to claim that their method is in general better than operational practices worldwide, this claim would have to be substantiated by additional model runs in different places.

The authors made their code available to review via a GitLab repository, which is much appreciated! The provided R scripts are well readable (although not entirely in agreement with modern style guides, e.g. https://style.tidyverse.org/) and seem to cover all steps mentioned in the manuscript, from data preparation to model building and plotting. I did not try to run the code, as the raw data is not provided, but the scripts make the conducted research transparent.

Specific comments

About the manuscript, I request the following clarifications and modifications:

- Please clearly define the term "season-ahead prediction". The term could be interpreted as predicting one season from the previous season, but I assume that the authors mean to predict one season from just before the start of that very season, as the 1-month-ahead streamflow appears to be included as predictor. Does the model only predict the maximum streamflow at some point during the season, or also a timing? 3 months is still quite an uncertain timeframe.
- In the introduction and discussion there should be an additional paragraph putting the used methods in context of what is state of the art in international scientific literature – not only in Peru. The last two sentences of the conclusion are: "*(…) because the statistical model developed here is optimized for performance across all years, further refinement prioritizing the detection of appropriate trigger levels for early action in high flow years may be warranted. Such efforts could involve alternative modeling frameworks (e.g. logistic regression), additional predictors, and evaluation of category selection applied in the prediction process.*" - But that is not enough and should appear earlier in the paper. Also, an additional paragraph about ensemble theory / multi-model studies would be adequate.
- Data: The authors should make very clear for the reader which data was used to fit the statistical models, i.e. how many observations, where does the target variable (y) come from and how certain is it, what exactly are the explanatory variables and how have they been treated (scaling etc.). Most of that information is somewhere in the manuscript, but it is not as clear as it should be on first reading. Table 3 could be a good place to collect this information.
- "*There are numerous methods for selecting the appropriate number of PCs to retain; here, the first two PCs are retained unless the model has two or fewer predictors, and then only the first PC is retained.*" (254-256). How is the selection of only 2 PCs motivated? Contributions may differ during the seasons or per region, but at least some sort of check should be presented, e.g. by plotting the cumulative explained variance for El Niño and La Niña (or any other method the authors prefer to make this point that 2 PCs are sufficient). According to Table 3, only in one case have there been 2 PCs used – in all other cases only 1, so it is only linear regression with 1 predictor? Or 1 PC plus the streamflow before the start of the season?
- A critical point, acknowledged by the authors, is the selection of a threshold to issue an emergency. In my opinion this problem could be communicated better to the decision makers if a full probability distribution of expected streamflow were predicted, rather than a point estimate. Bayesian regression would be the adequate tool, then. As the statistical model presented by the authors appears to be very simple (linear regression with 1 or 2 predictors), implementing this in a Bayesian framework should be feasible. In that case, also Bayesian decision theory could be applied for the threshold selection. Apparently the authors create an error distribution by sampling the model residuals 1000x with replacement, which might end up in similar estimates, although with slightly different interpretation. At least the authors should discuss the probabilistic output in more detail, and also discuss how this probabilistic output can be used in risk communication and decision making for the problem at hand.
- The multi-model seems to be dominated by the linear regression model. If this is the case, the authors could discuss which other models might be suitable to include in future multi-model ensembles.

Technical corrections

- All tables would benefit from some formatting.
- In Table 2, the letters J and F are used without explanation. I assume it is January and February, respectively, as the authors write in the text that the high streamflow seasons in the basins are FMA and MAM, respectively. January and February would therefore correspond to a 1-month-ahead value. However, that should be stated explicitly in the text and above the table – or more clear abbreviations like "Jan" and "Feb" should be used.
- Especially the very important "predictors" column in Table 3 consists of abbreviations with distracting line breaks. As the columns of that table are repeated, consider arranging the "negative phase" "positive phase" and "neutral phase" in rows rather than columns, and use the free space to add more columns giving detailed information on the models, like the number of observations, PC2 explained variance, maybe even the cross-validation score. Consider removing the bold rectangle and make the font of the column/row names bold instead.
- There is a LICENSE file included in the GitLab repository, but no README and CITATION files. I would like to encourage the authors to add these two missing components, although it is not a criterion for acceptance of the manuscript.