Reply to Reviewer #1

Title: Leveraging multi-model season-ahead streamflow forecasts to trigger advanced flood preparedness in Peru

Author(s): Colin Keating, Donghoon Lee, Juan Bazo, Paul Block

MS No.: nhess-2021-25

The authors thank Reviewer #1 for the constructive comments and feedback on our manuscript. Our specific replies are denoted in blue color and revised manuscript text is denoted by italics.

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.

We thank the anonymous Reviewer for these comments which have led to further improvements in the quality of the manuscript. We agree with the Reviewer's suggestion to amend the title of the manuscript to include "in Peru," which now reads:

> *Leveraging multi-model season-ahead streamflow forecasts to trigger advanced flood preparedness in Peru*

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.

We thank the Reviewer for this comment. We note that for additional readability, we have restyled all scripts according to tidyverse formatting rules.

Specific comments

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

1. 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.

To clarify the term "season-ahead prediction," lines 105-108 now read:

> *In this paper, we use the term "season-ahead prediction" to describe forecasting the mean streamflow for an upcoming three-month season issued at the start of that season. For example, a season-ahead prediction of January-February-March streamflow would be issued on December 31st and represents a prediction of the average streamflow over the upcoming three months.*

2. 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.

We agree with the Reviewer and have subsequently added additional paragraphs in the introduction section to detail the range of current statistical modeling approaches in the literature. We have also added a paragraph providing background on multi-model techniques. Combined, lines 81-100 now read:

> *A common traditional approach for statistical hydrologic modeling is multiple linear regression (MLR), which relates a predictand to the linear combination of several predictor variables (Moradkhani and Meier, 2010). For categorical streamflow forecasts, logistic regression (for two categories) or multiple logistic regression (for three or more categories) has been used successfully (e.g., Wei and Watkins, 2011). Because these methods are prone to multicollinearity due to the overlapping signals present in many hydroclimate variables, techniques such as principal component regression (PCR; a combination of principal component analysis and MLR) and partial least squares regression (e.g., Lala et al., 2020) are employed to address this challenge. More recently, machine learning techniques, adept at capturing nonlinear relationships between predictors and a predictand, have been successfully applied to hydroclimate forecasting, including artificial neural networks (Zealand et al., 1999), random forest classification (Ali et al.,*

2020; Lala et al., 2020) and support-vector machines (Asefa et al., 2006; Shabri and Suhartono, 2012). There is also increasing recognition that hybrid approaches combining statistical and dynamical techniques can offer greater accuracy than even state-of-the-art dynamical models (Cohen et al., 2019).

Multi-model techniques have been developed based on the assumption that errors present in individual models may cancel out, thus providing a multi-model average with greater skill than any individual model, and to bound forecast uncertainty based on the spread of model predictions. Several methods of combining models include equal weighting, linear regression and Bayesian methods that assign weights according to the probability that the model in question has the highest skill (e.g., Gneiting and Raftery, 2005). In some cases, multi-model ensembles have been shown to significantly increase forecast skill over the best performing individual model (e.g., Regonda et al., 2006), while not in other cases. For example, Bohn et al. (2010) note only modest improvement when using a least-squares weighted multi-model.

3. 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.

We thank the Reviewer for pointing our attention to this. Regarding the target variable and its certainty, we obtained this dataset from the Peruvian Meteorological Agency, El Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI), and they have conducted appropriate quality assurance. Lines 170-172 have been revised to:

Daily streamflow data for each location (1999-2017 at San Regis, 1971-2017 at Puente Sánchez Cerro) was provided by the Peruvian Meteorological Agency, El Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI), who performed appropriate quality assurance.

We have also clarified where the target variable comes from and the treatment of explanatory variables (scaling to have unit variance but no transformations) at the beginning of Section 3.3. Lines 265-277 have been revised to:

A principal component regression (PCR; coupled principal component analysis and multiple linear regression) framework is adopted to predict seasonal (3-month) average seasonal streamflow derived from daily streamflow observations obtained from SENAMHI as described in Sect. 2.5. The forecast for each location is composed of sub-models (multiple linear regression) composed of years in a particular climate state, as represented by the preseason (3-month average) value of MEI. This produces two sub-models for the Marañón River at San Regis and three for the Piura River at Puente Sánchez Cerro. A hindcast assessment is

conducted by evaluating each year in the historical record using the appropriate sub-model to predict seasonal streamflow. For example, in 1998, the preseason (NDJ) average MEI value is 2.43, thus the positive phase sub-model is selected to predict Piura River FMA streamflow. Predictor variable types listed in Table 2 may be included in some sub-models and not others, subject to their correlation with streamflow in that phase (Table 3). To be included, the predictor in question must be both significantly correlated with streamflow across all years and significantly correlated with streamflow in the subset of phase-specific years. A principal component analysis is conducted on eligible predictors which are first scaled to have a unit variance. A subset of PCs is retained according to North's Rule-of-Thumb (North et al., 1982) for input into the multiple linear regression.

We have also revised Table 3 (reproduced below) to include the number of observations (years) for each sub-model and the subset of predictors retained for each sub-model.

Table 3: Final predictors included in each sub-model.

Site	Sub-model	Number of observations	Predictors retained from Table 2	PCs retained	PC1 % variance explained	PC2 % variance explained
Marañón	Negative Phase	12	SST, SLP, SF, SM	1	61	22
	Positive Phase	7	SST, SLP, SF, SM, P	1	87	9
Piura	Negative Phase	11	SST, SLP, SM, P(GCM)	1	74	15
	Positive Phase	14	SST, SLP, SF, SM, P, P(GCM)	1	78	13
	Neutral Phase	11	SST, SLP, SM, P, P(GCM)	1	68	15

4. "*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?

Thank you for this comment; we have revised the method by which we retain PCs in our model and have revised Table 3 (reproduced above) to include the percent variance explained by the first and second PC. The revised process by which PCs are retained for each phase's sub-model are described on lines 275-281 as follows:

A principal component analysis is conducted on eligible predictors which are first scaled to have a unit variance. A subset of PCs is retained according to North's

Rule-of-Thumb (North et al., 1982) for input into the multiple linear regression, given as:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_n x_{n,t} + e \,, \qquad (1)$$

where y_t is observed seasonal streamflow in year t, β_0 is a constant, $\beta_1 \ldots \beta_n$ are regression coefficients, $x_{1,t} \ldots x_{n,t}$ are the PCs retained, and e is the residual or error. If North's Rule-of-Thumb indicates that no PCs are non-overlapping then only the first PC is retained.

5. 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.

We thank the Reviewer for this comment and acknowledge the potential value of alternative modeling approaches (e.g. Bayesian regression/inference), especially for threshold selection. We emphasize that this study illustrates the potential for tailored statistical approaches to complement operational physical forecasts, and acknowledge that a range of alternative statistical approaches may offer enhanced skill. We agree these alternative approaches warrant consideration in future work, especially for developing specified guidance for stakeholders.

As mentioned by the Reviewer, we undertake a simplified ensemble generation process to create a probabilistic forecast distribution. Again, alternative approaches are available, however we are not focused on selecting the 'best' approach, particularly since that clearly differs by case study, disaster, region, etc. However the Reviewer's point regarding the need to better emphasize the importance of the probabilistic output in our methods is well received. We have revised lines 281-282 as follows:

The creation of probabilistic forecasts are essential as early action decisions are conditioned on the forecast likelihood of an extreme event exceeding the 80th percentile.

We have also added more detail about the probabilistic output to lines 421-432 which now read:

The primary focus of this study is to predict the occurrence of high flow conditions to initiate flood preparedness actions, based on a sufficient percentage of the probabilistic prediction surpassing a pre-defined threshold. The probabilistic statistical forecast model at each location effectively captures interannual

variability and extremes (Figs. 4 and 5). For the two most extreme years in the observed record (2012 and 2015 for Marañón; 1983 and 1998 for Piura), the full distribution of predicted streamflow falls above the 80th percentile of observed streamflow (black dashed line). In these years, decision-makers are highly certain of an impending extreme event. However, for the majority of years, some smaller fraction of the forecast distribution falls above the 80th percentile threshold, presenting a greater challenge (less certainty) in decision making.

We agree with the Reviewer that utilizing probabilistic outputs is important in risk communication and decision making. We specifically address these issues in Sections 5.1 and 5.2, however we also acknowledge that there is room for improvement with respect to integrating probabilistic forecast output into decision making. This may include optimizing trigger thresholds, the probability required to surpass this trigger to initiate action, and exploration of the tradeoff in forecast skill and increased lead time for actions available at a range of lead times, all in the context of stakeholder tolerance for false positives and expected benefits. Indeed this is an active line of research in our group, however moves beyond the scope of this paper.

6. 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.

We agree with the Reviewer's observation that the multi-model is dominated by the statistical model (we note that this is now the case for only one site in our revised analysis with an updated PC retention and predictor selection method). Ideally, members of a multi-model should each contribute skillfully such that errors in any single model are balanced by the other models. In our case, the global physical model (GloFAS; currently used for early warning decision-making) lacks sufficient skill at our study sites (for the lead time evaluated) to improve upon the statistical model or counter-act its errors. A calibrated basin-scale physical model may be better suited and more skillful than the bias-corrected GloFAS forecast when coupled to one or more GCMs with demonstrated predictive skill in the region (e.g. NCEP CFSv2 and NASA GEOS-S2S for coastal northern Peru, according to the work of our colleagues in atmospheric science). However, given that our statistical model at Piura is already forced using GCM precipitation predictions, it is not clear that additional skill would be realized in a multi-model. A challenge of modeling at our present study sites is data scarcity; however, machine learning techniques that leverage remotely sensed data (e.g. detecting antecedent soil moisture conditions or the state or direction of the atmospheric-oceanic system) could potentially offer avenues for improvement. To that end, we have added the following text to our conclusion (lines 569-571):

Future work could also consider machine learning techniques with the goal of leveraging remotely sensed data to detect antecedent conditions at a subbasin scale and the state of the climate system.

We also note that the relative skill of the statistical and physical models (and thus weighting in the multi-model) may also be dependent on lead time, seasonality, and antecedent conditions. For example, the global dynamical model may be relatively more skillful at shorter lead times due to

its ability to include the effects of recent precipitation. At our study site, skill at shorter lead times may inform early actions relevant to the disaster event.

In conversations with our colleagues in the social sciences, we have learned that stakeholder buy-in – a critical step for creating forecasts that add value – may be easier to achieve with a simple model compared to a model that is more opaque or complicated. Further, the simple statistical model presented here performs quite well overall, and while a more complex model may perform marginally better, the overall gains are likely minimal compared to efforts placed on proper forecast dissemination, communication or training of stakeholders, etc.

Technical corrections

1. All tables would benefit from some formatting.

We have re-formatted all tables to improve clarity and readability.

2. 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.

We have added the following clarifying text to the Table 2 caption:

> *"J (F) indicates January (February)."*

3. 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.

We thank the Reviewer for these helpful suggestions. We have revised Table 3 (reproduced above), switching rows and columns and adding two additional columns for PC2 variance explained and number of observations.

4. 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.

We thank the Reviewer for this suggestion and have added README and CITATION files to the GitLab repository.