Reply to Reviewer #2

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 would like to thank Reviewer #2 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.

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.

We thank the Reviewer for bringing up this important issue regarding the uncertainty associated with a small sample size. We appreciate this suggestion to add uncertainty ranges to Figure 8 and have revised the figure accordingly (reproduced below). Additionally, we have discussed the implications of this uncertainty in section 5.2. Lines 471-496 have been revised to:

> *Skill in detecting events is highly dependent on the threshold probability required to trigger early action. In general, a lower threshold for action will result in instances of worthy action but also more actions in vain. Conversely, a higher threshold for action will prevent false positives yet will reduce the likelihood that early actions will be taken when needed. This tolerance for false positives when implementing early action is an open question for decision makers and may depend on numerous technical, institutional and political factors outside the scope of this study. Here, the trigger mechanism for early action, which requires a 75% probability of streamflow above the 80th percentile, suggests a tolerance for a FAR of 0.25 for an unbiased forecast. Crucially, the small number of events when each forecast triggers early action (4 for San Regis and 7 for Puente*

Sánchez Cerro), creates significant uncertainty in the POD, FAR, and TS values calculated for the hindcast period (Figure 8). However, notwithstanding sources of model-related uncertainty, achieving an acceptably low FAR at the 75% probability level with 95% confidence is possible for Piura with the GloFAS and multi-model forecasts (Figure 8d), although no forecast achieves this for Marañón (Figure 8c). Importantly, uncertainty in these metrics is generally reduced in the statistical and multi-model forecasts compared to GloFAS (e.g., Figure 8a from 30% to 65% probability). The confidence intervals for the statistical and multi-model forecasts also tend to be offset in the more skillful direction compared to GloFAS This is particularly the case for Threat Score (TS), a validation metric that describes the degree to which observed events correspond to forecast events, and is useful for evaluating the benefits of additional true positives against the costs of additional false positives when true positives are relatively rare (Figure 8e and 8f). However, there are notable exceptions to this trend, such as the large uncertainty in FAR for the statistical model at Piura above a 55% probability. While these results do not highlight an optimal probability threshold for decision makers, the statistical and multi-model forecasts generally appear more skillful across most probability levels. In addition, false positives incurred by reducing the trigger probability may also be offset by a stopping mechanism in which action is halted if the forecast is not confirmed 30 days later (IFRC, 2019).

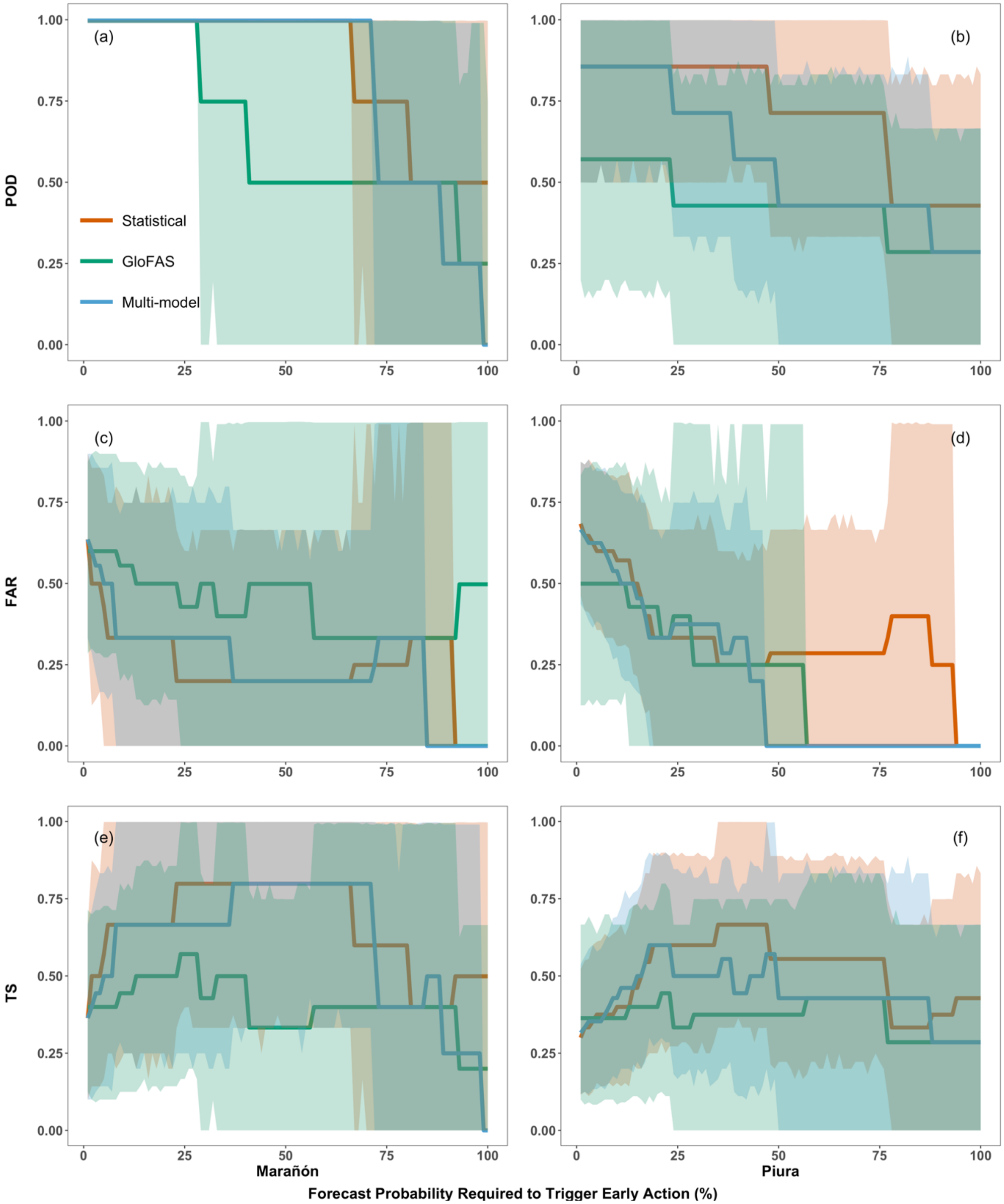

Figure 8: Probability of detection (POD), false alarm ratio (FAR) and threat score (TS) as a function of the threshold probability required to trigger early action for each location and forecast approach. Ribbons represent sample size-associated uncertainty at the 95% level, as calculated via bootstrap resampling of the hindcast period (n=1000).

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 suggest the authors rethink their advice on lowering the trigger without consequence.

We thank the Reviewer for this comment and in consideration of this have removed the conclusion on line 483. We have also revised lines 461-465 to the following:

> *A modified trigger mechanism captures some lower-magnitude events at San Regis; if early action is triggered based on just a 50% probability of exceeding the 80th percentile, the statistical model also triggers in 2009 and the multi-model triggers in 2009 and 2013 (thus each capturing all four observed events). However, caution is advised when reducing this threshold probability in practice as it will likely result in additional false positives.*

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.

We thank the Reviewer for this suggestion on how we might further improve forecast skill. We had initially explored the use of NMME forecasts of SST as potential predictors but found that this yielded no improvement in skill (in terms of correlation and RPSS) for predicting Piura streamflow. Specifically, we created a predictor from the average of the two NMME models identified in this paper (GEOSS2S and CFSv2), over the Niño1+2 region (80-90W; 0-10S) over the FMA season, issued Feb 1st. We selected SSTs in the Niño 1+2 region because correlation between observed FMA Niño 1+2 anomaly and Piura streamflow is very strong (0.82, compared with the correlation between observed FMA average SST anomaly in the Niño 3.4 region and Piura streamflow, at 0.30). However, correlation between Piura streamflow and predicted Niño 1+2 SSTs (from the two NMME models) is 0.74, less than the correlation between Piura streamflow and predicted NMME precipitation (0.84).

Additionally, we do not observe any significant improvements in prediction skill for years critical for flood preparedness. Given this result, we have chosen to retain our original inclusion of NMME precipitation predictions here, although we acknowledge that the NMME SST forecast performs almost equally well for Piura.

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.

We appreciate the Reviewer's interest in the statistical model weightings. During principal component regression, the set of predictors are transformed by PCA and a subset of the resulting principal components are retained for a multiple linear regression. Thus, the model coefficients are based upon the PCs (which contain information from multiple predictors) rather than the predictors themselves, which are highly correlated. One way to extract the relative importance of predictors is through assessing their individual correlation with streamflow, as presented in Table 2, reproduced below. From this perspective, it appears that pre-season SSTs and precipitation are most important for Piura, closely followed by antecedent streamflow, NMME precipitation forecast, and SLPs. For Marañón, pre-season SSTs, antecedent streamflow, and SLPs are relatively more critical, followed by soil moisture and precipitation.

<table>
<tr><th rowspan="2">Potential Predictor</th><th rowspan="2">Abbreviation</th><th rowspan="2">Spatial Region</th><th colspan="2">Time Frame</th><th colspan="5">Pearson Correlation with Streamflow</th></tr>
<tr><th>Piura</th><th>Marañón</th><th colspan="3">Piura</th><th colspan="2">Marañón</th></tr>
<tr><td>Streamflow</td><td>SF</td><td>-</td><td>J</td><td>F</td><td colspan="3">0.84*</td><td colspan="2">0.84*</td></tr>
<tr><td>Precipitation</td><td>P</td><td>Basin-Avg</td><td>J</td><td>JF</td><td colspan="3">0.88*</td><td colspan="2">0.68*</td></tr>
<tr><td>Soil Moisture</td><td>SM</td><td>1st PC of statistically significant (p < 0.05) regions within 12N to 23S, 35W to 81.5W</td><td>J</td><td>F</td><td colspan="3">0.69*</td><td colspan="2">0.74*</td></tr>
<tr><td>Air Temperature</td><td>T</td><td>Basin-Avg</td><td>J</td><td>F</td><td colspan="3">0.26</td><td colspan="2">0.11</td></tr>
<tr><td>GCM Precipitation Forecast</td><td>P(GCM)</td><td>4.5S to 5.5S, 79.5W to 80.5W</td><td>FMA</td><td>-</td><td colspan="3">0.84*</td><td colspan="2">-</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td>El Niño</td><td>Neutral</td><td>La Niña</td><td>El Niño</td><td>La Niña</td></tr>
<tr><td>Sea Surface Temperature</td><td>SST</td><td>1st PC of NIPA-identified regions</td><td>NDJ</td><td>DJF</td><td>-0.79*</td><td>-0.90*</td><td>0.85*</td><td>-0.93*</td><td>-0.80*</td></tr>
<tr><td>Sea Level Pressure</td><td>SLP</td><td>1st PC of NIPA-identified regions</td><td>J</td><td>F</td><td>-0.82*</td><td>-0.74*</td><td>0.79*</td><td>0.90*</td><td>-0.72*</td></tr>
</table>

Another way to indirectly assess the significance of each predictor would be to test the correlation strength between each predictor and the first PC of all predictors. Through some additional analysis outside this paper, we note that, for Piura, this first PC correlates most strongly with soil moisture in the negative phase (La Niña years) and precipitation in the neutral and positive phases (neutral and El Niño years). For Marañón, the first PC is most highly correlated with SLP in the negative phase and precipitation in the positive phase.

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.

Thank you for bringing this point to our attention. In an operational setting, according to the Peruvian Red Cross flood early action protocol, forecasts are issued on a rolling basis, with early actions taken any time streamflow forecasts are above the threshold. While for simplicity we issue our forecasts at a fixed date annually, it would be acceptable to issue the forecast on the 10th of the month (when GloFAS seasonal forecasts become available). We therefore opt to modify our statistical and multi-model forecast issue date accordingly. We have revised lines 105-110 to reflect this:

> *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. Ideally, 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. In practice, due to lags in data availability and for purposes of direct comparison with a physically-based model, forecasts developed in this paper are issued on the 10th day into the three-month season.*

We have also revised lines 196-199 to reflect this change:

> *Predictions of seasonal (three month) average streamflow (m^3/s) are issued on the 10th day into the three-month high flow season identified in Sect. 2, leveraging predictors based on values in the preceding months. Practically, issuing the forecast ten days into the forecast season allows time for large-scale climate data to be made available online, while also fostering a more direct comparison with GloFAS as described in Sect. 3.4.*

Lastly we have also revised lines 301-302 accordingly:

> *GloFAS forecasts are initialized on the first day of every month and become publicly available on the 10th day of the month.*

We note that this revision has the additional advantage of allowing a buffer window for other large-scale climate data sources to be made available online, and thus may be a more realistic issue date from an operational standpoint. We also acknowledge that additional predictor data from the first few days of the month could in theory be used, which may provide some additional forecast skill. We opt not to pursue this path because we expect any additional skill to be marginal due to the length of the forecast season and the slowly evolving nature of SSTs – a key predictor. This choice allows a more direct comparison with GloFAS seasonal because it is also initialized on the 1st of the month.

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.

We thank the Reviewer for this comment and would like to clarify that to the best of our knowledge FbA has only been applied to droughts more recently. Our prior confusion likely stemmed from a report by Cabot Venton et al. (2012) which modeled the costs of early response versus late response for drought in Kenya and Ethiopia but did not involve the implementation of a forecast based early action schema. We have updated lines 40-42:

> *While FbA was initially applied to acute and slowly evolving threats like tropical cyclones, more recent efforts have targeted hydrological threats including extreme rainfall and flooding (e.g., Gros et al., 2019).*

L69 There is a bit of a logical jump from the previous paragraph, consider adding a linking sentence.

Thank you for bringing this to our attention, we have revised lines 69-72 to:

> *Improvement in the skill of hydrologic models over the last several decades has aided the development of FbA systems for flooding. Among hydrologic models, those that are physically based (or dynamical) simulate physical processes such as infiltration and runoff to produce streamflow predictions and are often forced with climate predictions downscaled from general circulations models (GCMs) or numerical weather models.*

L111 Slightly long sentence, could be split for readability.

We have revised this sentence (now lines 138-139) to:

> *In the Amazon basin, the influence of climate variables on flood risk remains understudied (Towner et al., 2020) as a result of the nonlinear relationship between precipitation and streamflow (Stephens et al., 2015).*

L160 What do the colours represent in Figure 1? Satellite image, topography? If the latter then it needs a colorbar.

The coloring in Figure 1 represents idealized land cover. (This map layer was obtained from https://www.naturalearthdata.com/downloads/10m-raster-data/10m-natural-earth-2/.) We have updated the Figure 1 caption on lines 165-166 to clarify this:

> *Case study locations with catchment boundaries delimited in red. Shading represents idealized land cover. Made with Natural Earth (naturalearthdata.com).*

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.

We thank the Reviewer for this comment and would like to clarify that there is significant monthly autocorrelation in Piura streamflow. 174-177 in the revised manuscript now read:

> *Monthly mean streamflow at Marañón exhibits a sinusoidal autocorrelation structure, with statistically significant autocorrelation at one- and two-month lags as well as at interannual timescales. In contrast, streamflow at Piura exhibits significant autocorrelation at up to a three month lag yet minimal autocorrelation at interannual timescales, indicating a greater degree of variability in successive years.*

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.

Our initial goal for the statistical model was to forecast streamflow using three main classes of observed, pre-season variables: large-scale climate, precipitation, and (antecedent) streamflow. We deviated from this approach by including the NMME forecast for in-season precipitation for Piura, largely motivated by the relatively small basin size; this characteristic results in flashy-type floods and relatively limited watershed memory as streamflow moves quickly through the basin. On the other hand, the Marañón watershed, at 362,000 km^2, is significantly larger and preseason precipitation, particularly in the upper parts of the basin, correlates well with streamflow (0.68), due to travel times on the order of weeks to months. While including an NMME precipitation of prediction did not improve model skill, we agree with the Reviewer that including an SST forecast from NMME may further improve the skill of the statistical model. This would require further analysis, complicated by the fact that significant SST regions differ by phase as shown in Figure 3b). We suggest that these additions should for now remain an avenue for future work and stress that this paper's goal is to provide an illustration of how statistical forecasts may complement operational physical models for improved preparedness, which we believe the Marañón case achieves at present.

L226 I am unsure what " n.d." means in this context.

We have updated this citation to:

> *(NOAA, 2020)*

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?

We appreciate the Reviewer's question here. Aggregate model performance does not differ drastically, though is slightly improved in the 3-phase version (RPSS of 0.43 vs 0.39; correlation of 0.91 vs 0.88). One key reason for selecting the 3-phase model was its improved performance in key years for flood preparedness. For example, the two-phase version underpredicts 2017 streamflow by 35% compared to 12% in the 3-phase. Additionally, the 3-phase version reduces the spread of model residuals: on average, the standard deviation of residuals in the 2-phase model is 94.6 while the 3-phase lowers this to 79.8. We have thus rephrased lines 254-255 to better reflect our rationale for choosing the 3-phase model:

> *(While a two-phase model for Piura was also tested, the 3-phase model improves performance, including in years critical for disaster preparedness.)*

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).

We have updated lines 299-301 to incorporate information on the reforecast period, model version and calibration:

> *Monthly hindcasts over the period 1981-2017 from the physically based GloFAS Seasonal model (version 2.0) for the two study locations are available from ECMWF (https://www.globalfloods.eu/general-information/data-and-services/). Both study locations were used for model calibration (E. Zsoter, personal communication, May 6, 2021).*

L315 It would be useful to explicitly note how many upper tercile events are present for each site.

We have included this suggestion by amending lines 342-344 to:

> *As previously stated, the extreme category is classified as seasonal streamflow values in the top 20% (80th percentile) of observations – four events for Marañón and seven events for Piura.*

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).

We agree with this suggestion and have updated "observed trigger" to "event" on L399 and L448.

L450 I am not sure what is meant by "TS is maximised".

We have revised this section (see Comment 1) and have eliminated this phrasing.

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).

We agree and have revised line 495 to reflect this possibility:

> *It is also possible that observational error in streamflow measurements exceeds these differences.*

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.

We agree with the Reviewer's logic here and clarify that our intent was to illustrate the likely impacts due to instantaneous streamflow values. Therefore, we have revised lines 495-502 to the following:

> *From an operational standpoint, such edge cases beg the question: should some amount of early action still occur? An observed seasonal mean near the early action threshold, especially at the more variable Piura River, may contain much larger instantaneous discharge values and thus true flood risk may be obscured. Operationally, a trigger mechanism for early action at the Piura River should account for increased with-season variability of flows, perhaps by lowering the action threshold. Aside from these issues, a sharply defined threshold allows a*

potentially improper distinction between "worthy actions" and "actions in vain." In practice, absent a physical basis underpinning the action threshold, the difference in benefits resultant from early action may be negligible for instantaneous discharge just above and below the threshold.