Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system

Sanjib Sharma, Ridwan Siddique, Seann Reed, Peter Ahnert, Pablo Mendoza, Alfonso Mejia

Response to the reviewers' comments

We thank you very much for your thorough review of manuscript hess-2017-514. Below we provide a point-by-point response to each of the comments. The reviewers' comments are shown in blue font and our response follows immediately after that.

Comment from Reviewer #3: 1) the pre-processor is evaluated for 6 hourly 95th percentile events but is not evaluated for aggregated period events, which ultimately drive to floods. There is therefore a disconnection between the "value" of the post processor when evaluated independently, and the "value" of the pre –processor when verifying floods. The pre-processor has not been evaluated for the same "events".

Response to reviewer #3: The reviewer makes a good point. As we indicated before in our response to reviewer #1 and #2, we now use in the revised manuscript all the verification values when computing the verification metrics in Figures 3-7, i.e., we do not use any threshold or stratified sample. This means that all the preprocessed precipitation values and all the postprocessed flow values are used to compute the verification metrics.

We also note that we use 6-hourly accumulations since this is the resolution of the GEFSRv2 data after day 4 and since this is a temporal resolution commonly used in operational forecasting in the U.S. In Fig. 3, we want simply to illustrate the performance of S1 and S2 relative to each other, for this purpose using 6-hourly accumulations seems reasonable (i.e., the relative comparison between S1 and S2 is similar for 6-houlry or daily accumulations). Further, we use the 6-hourly precipitation accumulations to force the hydrological model and generate 6-hourly flows. Since the observed flow data are mean daily, we compute the mean daily flow forecast from the 6-hourly flows. The postprocessor is applied to the mean daily values since this is the resolution of the observations. But there is no mismatch between precipitation and flood events.

2) The conclusion that post processing only is needed to improve the skill of flow forecast seems to be based on statistics only and therefore you might get the right answer for the wrong reasons. The post processor maybe have the largest "value" but it does not mean that pre-processing steps should be skipped. I strongly recommend the authors to modify the conclusion to reflect that nuance.

Response to reviewer #3: We agree with the reviewer. As suggested by the reviewer's comment, we have now modified the conclusion to read as follows: "The scenario involving both preprocessing and postprocessing consistently outperforms the other scenarios. In some cases, however, the differences between the scenario involving preprocessing and postprocessing, and the scenario with postprocessing alone, are not as significant, suggesting for those cases that postprocessing alone can be effective in removing systematic biases."

3) Literature review and contribution of the paper and conclusion: A HEPEX blog by Boucher A. M. (2015) provides a summary of the contribution of previous papers. She refers to the papers also

mentioned below. 1) The literature and the insight provided by this experiment should be put in perspective with what has been done and found by others before.

Response to reviewer #3: Thanks for pointing us to this blog. We were indeed aware of the blog by Boucher A. M. (2015) (https://hepex.irstea.fr/pre-post-processing-or-both/), which summarizes different papers (e.g., Kang et al. (2010), Zalachori et al. (2012), Verkade et al. (2013), and Roulin and Vannitsem (2015)) related to preprocessing and postprocessing in streamflow forecasting. In fact, we have already discussed these paper/studies and their major findings in the original manuscript. Furthermore, our research questions and experimental set-up for the manuscript were designed in part to address concerns raised in the blog.

4) The fact that spatially disaggregated modeling is used might not be enough because there is no insight related to that modeling structure to the results. I would suggest framing the contribution differently.

Response to reviewer #3: We agree with the reviewer. However, we do not frame the contribution in terms of going from lumped to distributed hydrological modeling. This was not our intention and it is not what we say in the original manuscript. However, we do note in the original manuscript that this is clearly one aspect of the present study that differs from previous one. It was indeed surprising to us that most previous pre/postprocessing studies have been done with lumped models. Beyond the issue of model structure indicated by the reviewer, we think it is important to mention this aspect of the study because computationally the problem becomes very different when a distributed model is used. Further, the application of the preprocessor is also very different, hence worth mentioning in our opinion that a distributed model is used.

5) Study domain – this corresponds to the Susquehanna Basin – why use MAR instead of the Susquehanna River Basin?

Response to reviewer #3: We agree with the reviewer and have now incorporated this modification into the revised manuscript.

6) Warm and cold seasons: can you describe the type of events expected in both seasons?

Response to reviewer #3: To address the reviewer's comment, we added the following information to the revised manuscript: "The climate in the upper Mid-Atlantic Region can be classified as warm, humid summers and snowy, cold winters with frozen precipitation. During the cool season, a positive North Atlantic Oscillation phase generally results in increased precipitation amounts and occurrence of heavy snow. Thus, flooding in the cool season is dominated by heavy precipitation events accompanied by snowmelt runoff. While in the summer season, convective thunderstorms with increased intensity may lead to greater variability in streamflow."

7) PG 6 L31: change to "hourly"

Response to reviewer #3: Thanks for catching this. We incorporated this modification into the revised manuscript.

8) PG9 L4: add "observed" to "gridded precipitation"

Response to reviewer #3: We incorporated this modification into the revised manuscript.

9) PG9 L4: please specify the source of the gridded observed precipitation

Response to reviewer #3: The information requested by the reviewer is already included in the original manuscript. The text in the original manuscript reads: "Both the MPEs and gridded near-surface air temperature data at 4 x 4 km^2 resolution were provided by the NOAA's Middle Atlantic River Forecast Center (MARFC)".

10) PG9 L24: confusing; you mean "high precipitation events defined as 6-hourly accumulated precipitation events with a .95 non exceedance probability"? Also – see comment for the need to evaluate aggregated events

Response to reviewer #3: We have now modified the original manuscript to reflect the fact that we no longer use the 0.95 threshold but instead use all the verification data. We believe this change made the sentence more clear.

11) PG10 – Line 35: how do you specify flood events? Are those also 6 hourly discharge event with a .95th non exceedance probability? Please clarify

Response to reviewer #3: We believe that our previous answer to the reviewer helps to address this question as well.

12) Basins are not independent, could you add one comment how this might affect the results? In the result section at PG11 L34 it looks like you could see consistent results. It did not seem to be the case on the previous section.

Response to reviewer #3: We believe the results will be similar if we had selected basins that are geographically close to each other and of similar size to the ones we selected. In fact, we initially selected nested sub-basins in order to investigate the forecast performance with respect to basin size or, in other words, the scaling of verification metrics with basin size. However, we found that, although there is some tendency for the larger basins to show better forecast skill than the small ones, the scaling is rather mild and not consistent. The scaling tends to show significant variability so that it is not necessarily evident for the conditions considered (e.g., lead times and seasons). This information is now mentioned in the revised manuscript.