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 #1: 1) p6, l4: pi_i is only a probability when y_i=0, otherwise a likelihood.

Response to reviewer #1: We agree with the reviewer and have accordingly changed the text in the revised manuscript to read as follows: "For this, the predicted probability or likelihood π_i of the i^{th} observed outcome is determined as..."

Comment from Reviewer #1: 2) P7, l15: 'smallest mean CRPS is selected': I don't fully understand how this works. Apparently c_{i+1} changes over time, so what exactly is minimized here? The CRPS over some training data with a rolling training window? Please add some more explanation.

Response to reviewer #1: The postprocessor is implemented following a leave-one-out approach, which consists of using 7 years for training (i.e., to estimate c_{i+1}) and the 2 remaining years for verification purposes. This is done separately at each lead time until the entire 9 years have been verified independently from the training period. Thus, we determine a different value of c_{i+1} for each 7-year training period and lead time.

To select the value of c_{i+1} for each 7-year training period and lead time, we first generate ten equally spaced values of c_{i+1}. For each value of c_{i+1}, the ARX(1,1) model is trained and used to generate ensemble streamflow forecasts, which are in turn used to compute the mean continuous ranked probability score (CRPS) for the 7-year training period under consideration. Thus, the mean CRPS is computed for each value of c_{i+1}, and the value of c_{i+1} that produces the smallest mean CRPS is then selected for use in the 2-year verification period under consideration. This is repeated until all the years (2004-2012) have been postprocessed and verified independently of the training period. To address the reviewer's comment, we have now incorporated this explanation in the revised manuscript.

3) p8, l15-16: '... is focused on flood events ... by choosing flow amounts greater than ...': This kind of subsetting is very problematic and can lead to false conclusions about the relative predictive performance of different methods, see Lerch et al. (2017). Bellier et al (2017) give a discussion of pitfalls of sample stratification and make suggestions how one can stratify samples in a way that avoids these pitfalls.

Response to reviewer #1: We are thankful to the reviewer for this constructive comment. We have read the suggested papers and decided to use the entire flow values, as opposed to using a sample stratification approach, when computing the different verification metrics, with the exception of the Brier skill score. Accordingly, we revised Figures 3-7 in the new version of the manuscript. The revised figures are qualitatively similar to the previous ones. However, the revised figures are more consistent in showing the scenario involving both preprocessing and

postprocessing (scenario 6) as having better performance than the other scenarios. In addition, there are now clear differences between the warm and cool season, where the warm season shows the different scenarios, particularly S4-S6, as being more similar to each other, while the cool season results remained similar to the ones in the original manuscript. We have now modified the original manuscript in several locations to reflect the differences noted in the revised figures.

4) Section 4.4.1: I'm not sure if this part of the analysis makes sense. In addition to the stratification issue (which demonstrably entails a bias), it is also known that the ensemble mean does not necessarily yield the best/appropriate point forecast when a relative error statistic is considered (see Gneiting 2011). I suggest either considering the mean error (over the entire verification data), or omitting this subsection entirely and maybe replace it by a subsection that studies reliability of threshold exceedance.

Response to reviewer #1: We again thank the reviewer for this constructive comment. As suggested by the reviewer, we have now removed the relative mean error statistic and this sub-section from the revised manuscript.

5) P6, l31: hourly

Response to reviewer #1: Thanks for catching this. The typo has been corrected in the revised manuscript.

6) P7, eq (7): xi_{I+1} -> xi_{i+1}

Response to reviewer #1: Thanks for catching this. We incorporated this modification into the revised manuscript.

7) P9, l15: It sounds weird to say that one basin outperforms the other, please reformulate

Response to reviewer #1: We have now revised the text following the reviewer's suggestion. The revised sentence reads as follows: "Further, the performance of the calibrated simulation runs is similar across the four selected basins, although the largest size basin, WVYN6, shows slightly higher performance with Rm, NSE, and PB values of 0.85, 0.82, and -3%, respectively."

8) p10, l24: Replace 'While' by 'The gains ..., on the other hand,'

Response to reviewer #1: Following the reviewer's suggestion, we incorporated this modification into the revised manuscript.