Authors reply to Referee#2, 20 April 2021

Dear Referee#2

Thank you for the feedback on our article. We appreciate the valuable comments that are helpful in order to improve the manuscript. Replies and corrections are done as follows: the Author response (AR) is marked with red text, while the author's suggestions to corrections (AC) are marked with blue text. All Referee comments are kept in a black.

On behalf of the authors
Trine Jahr Hegdahl

This review is for Manuscript No.: hess-2021-13, entitled “The benefits of pre- and postprocessing streamflow forecasts for an operational flood-forecasting system of 119 Norwegian catchments”, authored by Trine J. Hegdahl, Kolbjørn Engeland, Ingelin Steinsland, and Andrew Singleton. With this manuscript, the authors examine the benefits of preprocessing metrological forcing and/or postprocessing streamflow forecast in improving the quality of ensemble streamflow forecasts. I believe the topic is of interest to the hydrometeorological community. However, I have major comments that needs to be addressed before this manuscript is ready for publication.

1. Abstract is quite long. For one, third statement (First Paragraph) convey the same message as the First and Second statements. Results in the Abstract Section need to be well summarized (focus on important findings, rather than mentioning all of your results).

   AC: We will avoid repetitions and reduce the length of the abstract. Further, we will focus on the main findings of the study.

2. The authors are using different training period (2014, 2006-2011, 45 previous days) for different preprocessing schemes. To ensure a fair comparison of methods, they need to each be assessed using an equivalent systematic method to determine the optimal training set for each method.

   AR: In the study we used the processing method used by the Norwegian meteorological institute for temperature and precipitation, including the set of tuned parameters in the pre-processing model (i.e. grid calibration). In addition, we performed our own processing using BMA. The grid-calibration and the BMA calibration differs in several ways, and reflects the difference in processing apporaches:

   - BMA is tuned to each catchment individually whereas the grid calibration is tuned globally to all Norway
   - BMA uses a sliding window of 45 days for training the model, whereas the grid-calibration used previous years to train the model. The parameters of the BMA model change for each issue date, whereas the grid calibration has parameters that depend on season.

   The idea of using an already exciting pre-processing method, was two-fold. Firstly, we wanted to assess if the existing grid-calibration improved the forecast compared to using the raw ensembles. Secondly, we wanted to assess the potential improvement by catchment specific calibration of the post-processing model. The results from this study show that the grid calibration, even though not optimized for the catchments, improved the skill of the operational hydrological forecasts for most cases. However, using BMA where the forecasts were tuned to each individual catchment gave the best performance.

To define the optimal training length for BMA, we chose to use the same training length for P, T and Q. This was done to ensure a fair comparison of the methods. We did however see that optimal training length depend on catchment, lead time and variable. The dataset is as pointed out limited, and we assume that optimal training length to a large degree will vary depending on the availability and representativeness of events in the training period.

AC: We think that since the two processing approaches differs in many aspects, it is challenging to discuss and explain why the outcome is different. To simplify both figures presenting the results and the discussion, we will exclude the results from the grid-calibration and only use BMA.

3. The authors are preprocessing and/or postprocessing the flood events. How realistic is it to preprocess (postprocess) large and rare precipitation events (flood events) by using just 45 previous days of training period? For flood events, it seems like the longer training period (multiple years, if available) is generally advantageous.

AR: We agree that the choice of training data for the pre- and processing algorithm is a critical issue. In this study we use a sliding window, as suggested originally by Rafetry et al. (2005) and later on used in several papers where the BMA approach is used for post-processing (e.g. Sloughter et al., 2007, Li et al., 2020). Typical sizes for the training window are 30 to 60 days. One argument for using a sliding window for training, is that the calibration adjusts to seasonal variations in model biases and easily adjusts to new model versions. In Raftery et al (2005) a window size of 60 days was used. Sloughter et al. (2007) used a window size of 30 days and found that increasing the training period beyond that did not further improve the sill of the forecasts. The choice of window size is a trade of between having enough data for estimation and obtaining a flexible post-processing approach that adapts to the most recent data. We tested several training lengths (A-Figure 1 and A-Figure 2 in supplement), and we needed at least 45 days to have enough days with no-zero precipitation, and we used the same training length for all forecasted variables. We acknowledge that pre- and post-processing extreme precipitation and floods using BMA is difficult. The reason is that the forecast performance of extreme event might be very different from the forecast performance for the training window. We think that an alternative approach to select training data might be to look for similar events.

AC: We will evaluate how sensitive our results are to the size of the training window and summarize the outcome. We will add a paragraph in the discussion about the choice of training data for the BMA method, the size of the sliding window and alternative ways to select training data.

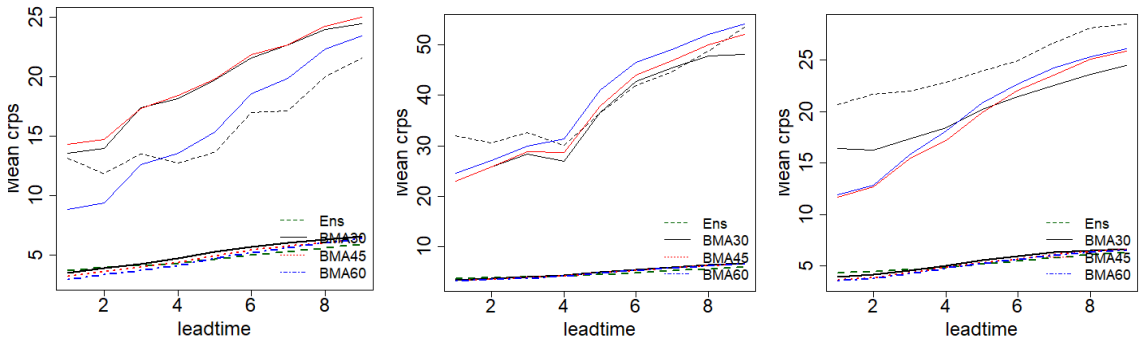

Figure 1Example of CRPS (vertical axis, small is better) for different BMA training lengths (30, 45 and 60 days) for precipitation. Lead times on the horizontal axis. The upper thinner lines are evaluated on floods only. Catchments from left: Bulken, Røykenes and Møska.

4. The authors use cubed root transformation in their BMA model for precipitation. How were this transformation chosen? Did the authors choose cubed root without testing alternative transformations?

   AR: Different power law transformations of precipitation within a BMA framework has been investigated in several papers, and the cube root transformation has been shown to give the best result (see e.g. Sloughter et al, 2007, and Li et al., 2020) and is often used as a standard transformation when processing precipitation. This is also the transformation tested and used by The Norwegian meteorological institute Norway (personal communication). In this study we chose to use this standard transformation.

   AC: We will add a sentence that inform about alterative power law transformations and use the papers by Sloughter et al (2007) and Li et al., (2020) to explain why the cubic root transformation was used.

5. Most of the Figures (Figure 5 - Figure 14; and Appendix-Figures) are difficult to follow. This really creates difficulty in reading the Result Section. Please make all the Figures simple and easy to follow.

   AR: It is challenging to find a good way to summarize results from a large dataset, and we will look for alternative approaches to present the results.

   AC: We will simplify the figures by reducing number of lead times, (present only every second lead time) and methods (present 7 instead of 14) presented, by

excluding the grid-calibration. We will keep the box-plots in Figures 6,7,10 and 11, but make the interpretation of the results simpler, and look at how we can e.g., change the colors to improve the figures.

Citation: https://doi.org/10.5194/hess-2021-13-RC2

Li, X, Chen, J, Xu, C-Y, Chen, H, Guo, S. Intercomparison of multiple statistical methods in post-processing ensemble precipitation and temperature forecasts. Meteorol Appl. 2020; 27:e1935. https://doi-org.ezproxy.uio.no/10.1002/met.1935

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M.: Using Bayesian model averaging to calibrate 6 forecast ensembles. Monthly Weather Review, 133(5), 1155-1174, 2005.