Dear Referee#1

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. We have split up the text and answered out the referee comments/questions as they occur.

On behalf of the authors
Trine Jahr Hegdahl

AC: We start some general comments to the suggestions by Referee #1. The reviewer suggests us to write a different paper where we focus on evaluating different pre-processing approaches applied to meteorological ensembles for a limited set of catchments.

The reviewer asks why we "needed to look at 119 catchments. Why not pick few catchments and present definite answers to the two questions which Authors have summarized in the conclusion." We think that using a large sample of catchments is the main strength of the paper. In the introduction we have identified this as one of the unique contributions of this paper. One of the main results from our study is that the resulting performance of pre- and processing approaches varies substantially between catchments, and conclusions based on one or a few catchments are therefore not robust.

The reviewer further suggests to "limit the work to only pre-processors, that could be one option". We think that searching for the best combination of pre- and prost processing is an interesting topic. This is also identified as one of the novelties in the introduction and is also the essence of the first research question.

In summary, we want to keep the scope of the paper as it is, i.e., the large sample of 119 catchments and also evaluate the benefit of both pre- and post-processing in different combinations. We will, however, within this scope of the paper, follow up several of the suggestions from the reviewer. We will now continue to address the other issues raised by the reviewer.

...

However, from Section 2.3 onwards, I started noticing that there is one line of thinking that Authors are trying to pursue in this manuscript, which may be the serious limitation of the scope of their work.

For instance, why should one use only HBV model, why to use ECMWF ENS data only, how can one interpolates 25 km spatial resolution forecasts to 1 km observation data without any downscaling technique, how Authors estimate the aggregated average values for each catchment, why not one use log-sinh instead of Box-Cox transformation, how justified it is to use Ensemble Coupla-Coupling, why not use Schaake-Shuffle, aren't there other pre- and post- processors than CAL and BMA, etc. etc. There are many such questions which are not addressed here.

AR: There are certainly several research question that could be addressed, but for making a paper with a clear scope, we found it useful to focus on a few research questions and elsewhere make choices based on recommended methods from the literature.

We used the HBV model since this model is part of the operational flood forecasting setup, it is well known and widely used and is hence an interesting model to evaluate We think that adding one more hydrological model would increase the complexity of the study.

ECMWF provides the best medium-range weather forecasts for Norway, and is used operationally, both in weather and flood forecasting. Deterministic forecasts are other options, like ECMWF High resolution forecast, and the short-range forecast by AROME-MetCoOp (not available as an ensemble at the time of the study), or a super ensemble based in the TIGGE archive could have been used. Again, adding more ensemble forecast would increase the complexity of the study.

The 25 km grid was resampled to the 1x1 km grid size using nearest neighbor interpolation combined with elevation adjustment for the temperatures. The BMA and the CAL methods are used to account for the discrepancy between the resolution of the model and the observations and can be regarded as bias-correction or statistical downscaling approaches.

We used the methods described in the text to aggregate the forecasted variables to daily averages.

We have previous studies showing that the Box-Cox transformation performed well for the Norwegian catchments (Engeland et al. 2010) and since this was not the focus of the study, we did not consider a log-sinh transformation.

The Ensemble Copula Coupling and Schaake Shuffle are similar, where the aim is to provide a temporal (and spatial) consistency between the variables that are processed using empirical copulas. The Schaake-Shuffle is often used with a reference to observations/climatology, whereas in the ensemble Coupla Coupling the reference is the ensemble members given by the model and is used to keep the temporal consistency between temperature and precipitation for each of the ensemble members.

There are plenty of pre- and post-processing methods (also mentioned in the manuscript), we did however choose to compare two approaches. One approach is to make use of data provided by a weather service center, where the focus is on improved precipitation and temperature forecasts for larger areas, this is referred to as the gridded approach (CAL). The alternative approach (BMA) is adaptive, and the calibration is catchment specific. We did not include more techniques in this study.

There are several other interesting approaches, methods, and questions to be asked. We think, however, that including several models, forecasts, and methods will increase the complexity of the study and not be beneficial to improve the clarity of this study.

AC: We will address each of these issues raised by the reviewer in a revised manuscript, by further explain the choices we made with respect to datasets, models, transformations etc. and refer to recommendations from literature when suitable.

In other words, I couldn't find what is novel here, knowing very well that there are several papers on this topic already published. Practically, every month we find new publication on pre- or post-processing in different journals.

AR: We argue that the number of catchments and the inclusion of pre- and postprocessing for both daily streamflow, and floods are a novelty. However, if we were to reduce the study to a few catchments, and only focus on preprocessing we find that the novelty will vanish.

Why not pick few catchments and present definite answers to the two questions which Authors have summarized in the conclusion. Since the study tried to capture so many different aspects, physiography, seasonal, snow-melt vs. rainfall based flood, etc. etc. that lead to Authors having fairly standard conclusions.

Instead, I would look into few aspects but with rigorous analysis and try to derive some conclusions which can actually benefit the hydro-meteorological forecast community, not only in Norway but other places also.

Authors reply to Referee #1, 31. March 2021

AR: One of the main results from this study is that it is very difficult to be conclusive about transferring information from a single study. We do however provide some insights that is important to consider for the usefulness of processing.

It is very important to have plots which can be interpreted easily.
In this manuscript, almost all the results are shown through box-plots. A set of time series plots showing how good the pre- and post-processors are improving the forecast would be highly beneficial.

AC: We can provide time series plots for selected catchments, showing the median forecast for each method at e.g., lead time 1, 5, and 9 days for the study period.

Similarly, to show the improved flood forecasting, a time series plot would make things very clear.
But I can only imagine the difficulty one would face in summarizing the results of 119 catchments, 51 ensembles, 9 lead time etc.

AR: We will do an effort to find alternative ways of presenting the results from this study.

Therefore, a small number of catchments from different parts of the country may be the way forward. In one sense, Authors are already doing this by summarizing results of only 6 catchments.

Then please limit the work to only pre-processors, that could be one option. In summary, Authors have to find a way to focus on novelty of this study rather than trying to cover all possible aspects on this topic.

AR: Only focus on preprocessors for the flooding events.

Citation: https://doi.org/10.5194/hess-2021-13-RC1

Reference:
Engeland, K., Renard, B., Steinsland, I., and Kolberg, S.: Evaluation of statistical models for forecast errors from the HBV model. Journal of Hydrology, 384(1), 142-155, 2010.