"Fuzzy committees of specialised rainfall-runoff models: further enhancements" by N. Kayastha et al.

RESPONSE TO Referee 3 (Interactive comment)

The paper presents an application of a multi-model combination approach known as "fuzzy committee models'. Several case studies are presented that examine the utility of the methodology. I had several main comments, mostly related to the depth of analysis and the study design.

> The authors are grateful to the reviewer the valuable comments and advices. We tried to address all these comments in this answer.

1. A concern I had is related to the calibration/verification approach. There is little discussion about how these periods were selected (and indeed the various data lengths are very short).

> ANSWER. We agree we have not mentioned this. Ideally, we have to try to split data into statistically similar sets (coverage of seasons, number and size of peaks, variance, mean, etc). Of course in this type of splits of hydrological one is constrained by the wish to keep data in contiguous blocks (to be able to plot the time series data such as hydrographs). So with the data available we did not have much choice, this choice was made and it is presented in Table 1.
> The corresponding explanation is added in the beginning of Section 3.

There is very little discussion about what the verification results revealed, which is disappointing as the ranking of the approaches changed between calibration/verification suggesting potential overfitting.

> ANSWER. Nobody can guarantee that a particular parameterization of some model which is best in calibration will be also best in verification – simply because data sets are different, contain noise, etc. Yet another reason could be that the calibrated models overfit the data, so in verification the models are not very accurate – and we admit we have not used sophisticated stopping rules in optimization to ensure minimum error on cross-validation rather than calibration set. In the new version of the manuscript we added a recommendation for the future work to perform cross-validation during calibration (for its early stopping) to prevent overfitting.
> Please note however that in all case studies the best committee model (identified by calibration) outperforms the best single model (identified by calibration) on verification data. In the new manuscript the conclusions now clearly state this.

2. The authors present results for both RMSE and NSE. Including both of these statistics is redundant as they target exactly the same flow characteristics (they minimize squared differences between observed/simulated flows). Only one should be used. Given the emphasis on squared flows, these statistics are also biased toward optimizing high flow simulations, so the results are not surprising. I would suggest selecting a variety of statistics that capture different elements of the hydrograph rather than a single summary statistic. This would make the discussion/assessment richer.

ANSWER. We use NSE because this is a traditional for hydrology measure, along with RMSE. Models can be optimized on one of them, and we have chosen to minimize RMSE (it could have been maximization of NSE as well).
To address this comment, we added a new table (after Table 3) with the *RMSE* calculated separately for low and high flows.

3. Somewhat related to this, the discussion of the results was rather shallow. Table 3 is not really discussed at all, and to me this provided the most food-for-thought regarding the results. I think the paper would be enhanced by expanding this table to better examine the differences between the models (using other statistics) and then providing a deeper analysis of what this table reveals.

ANSWER. We agree, and in the new version provide more observations based on Table 3, and we updated Discussion and Conclusions sections. In Conclusions we added the following observations and conclusive statements:
"• In calibration a committee model is always better than the single model, independent of the values of parameters MFtype and WStype (however we have to optimize δ and γ).
• When tested on verification data, the best committee model (identified by calibration) outperforms the best single model (identified by calibration) on all case studies.
• We cannot suggest the "universal" best set of parameters MFtype and WStype applicable for any case study: in calibration all of them were good, but in verification performance of models using different MFtype and N slightly differs for different cases."

4. One thing I found lacking in the paper was a discussion of the importance of recognizing uncertainty in the modeling process. Indeed, the multi-model approach is often favored as it addresses the idea of 'structural uncertainty' or the potential error/bias when relying on a single model structure. Here, the authors have not addressed uncertainty at all in their analysis, and I think this is a real weakness. The analysis of uncertainty is by now routine in hydrologic modeling studies and should form part of the basis by which different models are compared. In some cases, there are only small differences between the model simulations, which makes relying on a single summary statistic troublesome. If the results were expanded to consider predictive uncertainty then the comparison of models would be more convincing.

ANSWER. Indeed, we agree we have not covered this issue. We realize the importance of such analysis, but we would like to leave it to be addressed in the future studies. To address this comment, we have added the observation about sensitivity: shape of the membership function and weighting function in RMSE do not influence the choice of the best model in calibration. At the same time parameters (δ, γ) do influence the model output and its performance. We also added a recommendation to undertake explicit analysis of sensitivity and uncertainty, and suggested also to analyse dependencies between parameters (*MFtype*, *WStype*) and to employ more robust optimization methods used in model calibration.