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

RESPONSE TO Referee 1 (Interactive comment)

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

The paper presents an analysis and development of a theory that explores the improvement of predictive models by a soft combination of modules or models. The concept of fuzzy committee have been well conceptualized by some of the authors of this paper in their previous publications, however few papers have developed comparative examples with optimized models. Although it is a very interesting study I have a couple of concerns.

The main conclusion is taken from calibration and not from the verification data set.

ANSWER: It is not the case, and we recognize why this Reviewer made this conclusion: our formulations at places were not clear and did not reflect the results shown in Table 3 correctly. In fact, from Table 3 it can be seen that for all case studies a committee model has higher performance than a single model both on calibration and verification data sets. In the new version of the manuscript we reformulated our conclusions to make them clearer.

In the new version in Conclusions we now state:
"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.".

There is also an important difference in performance between calibration and verification (Figure 3 and 4). I would suggest extending the verification results on only low flows or high flow regions, and to check where is the "gain or loss" of performance. With this, it should be possible to detect whether there is improvement or not of the fuzzy committee.

ANSWER. There is no surprise that performance on verification set is lower. Concerning the comment on the performance on low-high flows, we followed the recommendation of the reviewer and in the new version added a table showing the performance (*RMSE*) calculated for low and high flows separately.

I should mention a number of important points that I believe should be addressed.

1. The fact that RMSE is an error measure that squares errors and provides implicitly more weight to high values makes it not suitable to compare two different regimes at the same time. May be a normalized value could provide more information.

ANSWER. Indeed *RMSE* provides implicitly more weight to high values but our idea was to amplify this difference even more. It worked: we found that our model with the

explicitly reinforced accent on high flows by using weight w_{HF} performs better on high flows than the Single model using standard (non-weighted) *RMSE*.
It is suggested to "compare two different regimes at the same time"; to answer this comment, we added a new table after Table 3 with performances shown separately for high and low flows.

2. It is interesting to see that the high flows in Leaf catchment (fig. 3) have less RMSE than the low flows. I believe this is due to the RMSE used in the optimization of the models, that RMSE includes weights in its operation, the graph is misleading. The graph should show the normal RMSE to be able to provide information about the observed value and not the distorted reality balanced by the weights of hypothetical flow regimes determined by the arbitrary parameter Alpha.

ANSWER. We understand the essence of this comment, and struggled ourselves on how to present results best with minimum confusion. We do not think it is reasonable to change the mentioned graph. We hope a reader would be able to understand that $RMSE_{LF}$ and $RMSE_{HF}$ cannot be compared since they use different formulas. $RMSE_{LF}$ values happened to be even higher than of $RMSE_{HF}$ – the reason is that the number of low flows is much higher than of high flows, and the denominator (total number of observations) in both formulas is the same. However, to take this comment into account, we added an explanatory sentence in the manuscript in Sec 2.2.

3. On the other hand, it is well known that due to the random generation of some of the parameters the overall RMSE variability in the calibration imply performance values in verification. Therefore, most of the models might have RMSE lower than the committee error improvement presented in the paper, if so the conclusion is not really possible to be made out of such results (check table 3, low difference in values of errors). It is important to make either an analysis of the variability of the models used with each data set (verification) or either do a ten fold cross validation process.

ANSWER. We agree that there is variability in *RMSE*, and indeed differences in performance between various models are not large. To address the issue of possible sensitivity of results to optimization scheme, we used two different global optimization (calibration) methods (which showed similar results), and of course RMSE is an average across thousands of records for both calibration and verification, so we hope the effect of variability is not substantial. Ten-fold cross-validation is a good idea, but unfortunately we could not allocate more time for this, so we added a recommendation to do it in the future. We still think that the results reported in Table 3 are valid and by comparing *RMSE* for various models we can state that using committee models brings improvement (albeit not substantial).

4. The actual pareto front seems to show only calibration values, may be is better to show pareto graphs only with verification results; if the goal is to conclude something about performance. If the goal is to conclude on the calibration capabilities and its relation with verification samples, this should be identified in the same graphs. The Figure 3 and 4 does not separate the pareto local models used in calibration and verification.

ANSWER. Fully agree. In the new version we have added a new plot ($RMSE_{LF}$ vs $RMSE_{HF}$) showing the verification results. However we cannot put all results on one plots since in calibration and verification the (relative) weights used are different (e.g.

note in Eq 2-7 normalized Q (denoted as l or h) depends on Q_{max} which could be quite different in calibration and verification).

We also changed notation “local model” to “single specialized model” to be consistent with the rest of the text.

Note: We found that enough projections of the model parameterizations of Leaf catchment for calibration and verification in updated Figure 3 and the plots for Bagmati and Alzette are similar as Leaf catchment. Adding more graphics would not be helpful to improve clarity this work and the reasons of limited space we decided not to present in this paper.

5. It is important to provide the reader with a figure that allows him to visualize the hydrograph and highlights what is considered as low and what is high flow, according to the fuzzy parameters selected (Related to the conclusion Page 683 line 25). I believe that the difference in regimes is the most probable reason of the improvement difference.

ANSWER. We understand the nature of this comment, but to address it would be difficult: there is no threshold separating the low and high flows. In most types of weighting schemes we use two smooth weighting functions that force a model to be more accurate for lower or higher flows. Indeed, we agree that “the difference in regimes is the most probable reason of the improvement”.

Aside of this, it is possible to see a number of English mistakes that would be important to correct to improve the readability.

ANSWER. We agreed. We updated the text.

Page 677 line 19 to 22, please divide and explain better, is not clear.

ANSWER. This text is now rewritten: "*In the present paper we tested the performance of committee models that use several weighting schemes in objective functions for calibration of specialized models and different membership functions to combine models. We also tested their performance on test data sets*".

Page 680 line 14, what is viva versa? sentence is not clear

ANSWER. Corrected.

Page 681 line 5, Check sentence.

ANSWER. Corrected. Now it reads: *" First the two optimal specialized models: model 1 for low-flow ($Q_{LF,i}$) and model 2 for high-flow ($Q_{HF,i}$) are sought using optimization (minimizing $RMSE_{LF}$ for model 1 and $RMSE_{HF}$ for model 2); this can be done by solving a single-objective optimization problem separately for these two models, or by multi-objective optimization for two objective functions $RMSE_{LF}$ and $RMSE_{HF}$)*."

Page 681 line 10, why to use NSE coefficient if your targets have been built with RMSE weights. Did you check NSE per regime of the data?

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). NSE is not used for each regime of data but only for the whole time of simulation.
To address the comment about regimes, we also added a new table (after Table 3) with the RMSE calculated separately for low and high flows – this allows for more detailed analysis of model performance in different regimes.

Page 681 line 24, check sentence.

ANSWER. Corrected as " *The identified best sets of parameters for different models are given in Table 4 (Appendix B).*"

Page 683 line 17, this contradicts your 3rd conclusion.

ANSWER. Agreed. We updated the text.