Response to reviewer 1

General remarks

This paper describes results from a hindcast study of a multimodel approach for ensemble streamflow forecasting. While the method has been described in an earlier paper, this paper focuses on the performance evaluation of two variations: MEads and MEhds. The skill metrics are well explained and results show a clear improvement of skill compared to the historical ensemble (HE, or classical ESP) method.

The text is well-written and clear. Below are a few minor comments and questions.

Questions and comments

Page 3, line 30: It surprises me that no weighting scheme was applied. I would expect you have knowledge about the relative performance of the different model chains from experience of from earlier studies (Olson et al, 2016). Is this something you intend to investigate in the future?

Yes, we understand your surprise here. The wording used (page 4, lines 1-2) implies we did not test weighting schemes which is not the case. We did in fact test two types of weighting, a simple arithmetic weighting system similar to that used by Olsson et al. (2016) and a linear regression based approach. In both cases the sharpness of the multi-model ensemble was improved significantly but there was no improvement in the multi-model's general performance. The pooling approach was still able to hold a small advantage over the weighted versions.

The following changes were made to correct and clarify what we did:

page 4, lines 1-2 was changed to read,

"The simple weighting scheme used by Olsson et al. (2016) was tested but, other than improving the ensemble sharpness, did not offer an improvement over the pooling approach."

The following paragraph has been added to the conclusion as well:

"How the individual model ensembles are combined to give the multi-model output needs to be revisited. When we applied the asymmetric weighting scheme proposed by Olsson et al. (2016) we did not find that it improved the multi-model performance in general across all stations and forecasts and so did not use it. However, we do believe that more work should be done to find a more appropriate weighting scheme than simple pooling. Perhaps by better understanding how the performance of the different modelling chains are affected by the initial conditions and lead-time it will shed more light on how to best approach this issue. Further development and testing along these lines are planned for the future."

We use the mean or average index values for the period. We have reworded the sentences (page 4, 29) to clarify this,

"The teleconnection indices they identified are the Arctic oscillation (AO) and the Scandinavian pattern (SCA) and the periods of persistence for these indices, expressed as the index mean for the identified period, are the seven and eight months leading up to the spring flood respectively."

This is a typo. The reason that the entire data series cannot be used is that the hindcasts of the driving data used in the multi-model are only available from 1981.

We have changed it to read, "…the data used in our work are for the period 1981-2015 due to some of the other datasets used in this work only being available from 1981".

Grammar and spelling

- To achieve this, operators …
- In practice, there are …

The text has been reviewed and punctuation added when using adjuncts as suggested.

The second hyphenated word has been removed.

Changed to read, "We focus on forecasts of the accumulated streamflow volume during this period or SFV."

The word 'the' has been removed so that it now reads, "… to perform better than…"

The 'an' has been changed to an 'a' to read, "a mean".

The phrase has been replaced with the 'at' so the sentence reads, "Out of the three terciles the prototype shows the least skill over HE at discriminating between NN events and non-NN events."

The word 'the' has been added as suggested.

Table 3: LT, MT and UT are not introduced. Do you mean BN, NN, AN?

Yes, this is a typo and we did mean BN, NN, AN. The change has been made.