Answer to referee #2

Referee comments were copied in italic, answer of the authors are in normal font.

It is a very interesting paper, well worked and clean presented. There are several works in hydrological literature trying to reduce simulation uncertainty by models combination (especially in real time flood forecasting) and few ones in quality modelling (as it has been well reviewed by authors). I recommend its publication, but some additional explanations should be done in the final version:

The authors would first like to thank the 2nd referee for her/his positive comments and interest in our work. Answers to comment are in the following paragraphs and revision of the manuscript is ongoing.

1.- The single models and the ensembles have been calibrated in the period Jan 2000 to Dec 2004. However, I don't see any type of model validation (temporal and/or spatial), which is important in general (see the details for example in the DMIP project, which is mentioned in the paper), but crucial when using models without physical meaning. And the linear regressions between models are not. So, two important questions arise immediately:

- Why there is not a temporal validation period, splitting the 5 observed years into 3 years for calibration and 2 for validation?

- The results will be the same in a different year than calibration period?

Reply: We agree that validation procedures of conceptual models are commonly used to demonstrate the robustness of a model (i.e. structure and parameters) by applying it in other spatial and/or temporal conditions. However, if we adopt a split-sample approach to our data set, the calibration and validation periods would become really short with very few observations to 'train' the models. The focus of the article was to demonstrate the applicability of different model fusion methods to nutrient predictions and show the advantage of using different, usually highly uncertain, model structures rather than one. Given this objective, we still think that disregarding a validation and calibration phase is acceptable. However, we will include this point of discussion in the conclusion of the paper.

2.- P5314 L15-19. From my point of view, this is the key part of the paper and I think more explanations are needed. For example:

- I don't fully understand why are you using monthly measurements and daily simulations for regressions

Reply: In our case study, one measurement per month (i.e. grabbed sample) was available for each of the nitrogen species, usually around the 15th of the month. In order to check the goodness-of-fit between the models and the observed data, we compared the observed values corresponding to these samples with the value predicted for the same day. Still, the models also provided daily simulations of the loads of the different species they consider for the un-monitored days. In order to check the correctness of the fusion methods, we extrapolated the different weights to these un-monitored periods as well and dismissed the predictions providing unrealistic negative values.

- I must assume the same types of regression are done for MMEs than SMEs, or donot I?

- What are the a priori implications of using constrained or unconstrained?

- Or more general, why are you using linear regressions for ensemble constructions? Remark any precedent in literature if any. Or in other words, justify this type of ensembling.

- Can you explain with more detail Table 2?

Basically, the same methods were applied to build SMEs and MMEs, the only differences being the input members (i.e. 10 best members of the MC procedure for the SMEs, best SMEs for the MMEs). Linear regression techniques have been used several times before in different modelling contexts: meteorological forecasts (Krishnamurti et al., 2000), sea surface temperature (Fraedrich and Smith, 1989) or rainfall-runoff (Ajami et al., 2006; Shamseldin et al., 1997; Viney et al., 2009). As explained by Viney et al. (2009) multiple linear regressions might include a non-zero intercept leading to predict flow even if all the members predict a zero flow. They might also assign negative coefficients and this can result in negative flow predictions. In order to circumnavigate these problems, the constrained regression through the origin is used.

We will include a better description of the regression method in the revised paper version.

3.- P5311 L17. It is not clear if, for the paper case study, there is only hydrological calibration (and N parameters are obtained by Monte Carlo simulations) or there is a two step calibration (including N submodel calibration).

Reply: The hydrological modules were calibrated against the discharge records of the two stations. The N predictions were obtained via a MC uncertainty procedure. This point will be clarified in the revised manuscript.

4.- I like very much the Discussion Section, but I find the conclusions are very short. Why not merge both in a single "Discussion and conclusions" section?

Reply: Thanks! The suggestion of merging the discussion with the conclusion part will be considered, though we have several other items raised by the reviewers that might be included in the revised version of the conclusion.

MINOR CORRECTIONS/SUGGESTIONS

Unfortunately different people understand differently the scale "mesoscale". Stress in the Introduction that in your case "mesoscale" is basin scale (I think!).

Reply: This will be corrected in the revised manuscript.

Models combination is more frequent in real time forecasting world. Authors cite Abrahart and See (2002), but there are more interesting works for a literature review in this topic. Just two examples (I have not published any related paper, by the way!):

- Regonda, S. K., B. Rajagopalan, M. Clark, and E. Zagona (2006), A multimodel ensemble forecast framework: Application to spring seasonal flens in the Gunnison River Basin, Water Resour. Res., 42, W09404, doi:10.1029/2005WR004653.

- Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, Water Resour. Res., 43, W01403, doi:10.1029/2005WR004745.

Reply: We agree that model combination is mostly used for real-time / short-term forecasting and thank the referee #2 for sharing her/his knowledge about ensemble modelling. We will consider some more references about ensemble modelling in the revised version of our paper.

P5306 L25. There is an incompatibility between subcatchments division and HRUs and land uses. If I have understood well, the last are nested. Explain it here.

Reply: The semi-distributed LASCAM, CHIMP and SWAT consider the catchment as a succession of sub-catchments. LASCAM computes the water and N balance at the sub-catchment scale. In CHIMP, each sub-catchment is divided into several land-use classes for each of which the water and N balance is computed. Output of each land-use classes are weighted by their respective area to compute the total runoff and N flux discharging from the sub-catchment. In SWAT, the delineation design is a bit more complicated. HRUs in SWAT are lumped representations of similar land use and soils, and they are not spatially explicit. Similarly, the HRU outputs are weighted by their respective area.

A clearer explanation will be implemented in the text.

Table 1 caption. Better "Main model characteristics"

Reply: This will be corrected in the final text.

Table 1. Authors have described SWAT explaining it has three different runoff fluxes surface runoff, lateral flow and baseflow. It will be interesting to have the same equivalent description for the other models, in this table 1 and within the text. In the text, I suggest also to underline the implications of different hydrological modelling in the N modelling.

Reply: This will be implemented in table 1.

Table 3 and 4. Are results during the calibration period? Mention it.

Reply: Yes, this will be highlighted in the revised version.

Figure 1. Names in the figure are confusing: there are two "Vattholma", names have different colors without explanation, Uppsala is clearly a city, but the rest, are they cities, basins or stations?

Reply: Blue names are those of the discharge stations (blue markers) whilst red names are those of the monitoring stations (red markers). Explanation will be implemented in the figure caption.

Figure 3. It is not needed "METHODOLOGY" within the figure, because it is already in the caption.

Reply: This will be corrected in the revised manuscript.

References

Ajami, N. K., Duan, Q., Gao, X. and Sorooshian, S.: Multimodel Combination Techniques for Analysis of Hydrological Simulations: Application to Distributed Model Intercomparison Project Results, J Hydrometeorology, 7(4), 755, doi:10.1175/JHM519.1, 2006.

Fraedrich, K. and Smith, N. R.: Combining Predictive Schemes in Long-Range Forecasting, J. Climate, 2(3), 291-294, 1989.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S. and Surendran, S.: Multimodel Ensemble Forecasts for Weather and Seasonal Climate, J. Climate, 13(23), 4196-4216, doi:10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2, 2000.

Shamseldin, A. Y., O'Connor, K. M. and Liang, G.: Methods for combining the outputs of different rainfall–runoff models, J: Hydrol., 197(1-4), 203-229, doi:10.1016/S0022-1694(96)03259-3, 1997.

Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H., Gräff, T., Hubrechts, L., Huisman, J. A., Jakeman, A. J. and others: Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions, Adv. Water Resour., 32(2), 147–158, 2009.