Answer to referee #1

Referee comments are in italic, answers of the authors are in normal font.

This paper is an interesting contribution in the field of biogeochemical modelling at the catchment scale. It is appealing to see attempts to bring some light into the issue beyond the Manichaean arguments about model capabilities, usually only a reflection of modelers' preferences (frequently acquired just by contingency). I liked a lot the starting point of the paper, because it explicitly accounts for one of the most basic (and sometimes forgotten) limitation of a mathematical model: it is just maths!!, including lots of assumptions that may or may not be appropriate. With no doubt, ensemble modelling is one of the best shortcuts to cope with model's incompleteness also in catchment biogeochemistry, as is the case in some other areas where modelling is the core methodology (e.g. climate prediction). However, there are some assumptions and reasoning in the paper that in my opinion deserve discussion, because they are central for the conclusions reached. I'm particularly worried about the modelling target (nutrient load), because this is at odds with one of the basic objectives of the paper (to focus on N dynamics). Also, although the presentation is good in general terms, the explanation of some of the methodological aspects could be improved.

We thank the anonymous referee #1 for her/his interesting comments. We provide answers to her/his concerns in the following paragraphs and will consider implementing the remarks in the revised version of the manuscript.

1. Why are you using nutrient loads as modelling target if your aim is "to focus the study on the stochastic uncertainty linked to the nitrogen algorithms only"? I think nutrient concentration is a best option because: a. If streamflow and nutrient concentration are related, then is quite easy to fit the nitrogen modules against a corrected nitrogen concentration trace. I mean, if you have previously fit the streamflow modules, you can calculate the mismatch between modelled and observed streamflow. Then, you can correct your observed nutrient concentration by this mismatch considering the empirical relationship between nutrient concentration and streamflow. Fitting the model against this corrected nutrient concentration avoids biases in the nutrient parameters determined by poor fit to streamflow. b. If streamflow and nutrient concentration are not related at all, then you can assume that streamflow is not dramatically affecting your nutrient dynamics (or at least it is not the main driver). Then, work with raw nutrient concentrations. I think this is quite important in a paper where different models showing different fiis to streamflow are detected. I think in your case the importance of loads for management or other considerations are of secondary meaning, because you present here a pure technical issue.

Reply: We agree and see the point that such a procedure could improve N predictions for single models. However, this was not the scope of our paper. The aim of the work is to get good predictions based on existing model structures. We do not want to improve model structures, nor do we intend to "judge" which model structures are superior. The idea of ensemble modelling and the subsequent fusion of results are to get the best out of a set of simulations. The reason why we are not focusing on concentrations is simply that the existing model structures we are using have been developed for N load predictions as they are based on mass-balances calculations. As indicated in the discussion paper (P5303 L26-28), we considered this study a methodological one. Ensemble modelling and data fusion has not been applied before to hydro-biogeochemical models and we see this as a great step forward in covering part of the structural uncertainty in these models (see also reply to next comment). We will better introduce our aim of the study in the revised paper.

2. I was a bit confused by the way you used the term "uncertainty" in the paper. In your work you only used ensemble modelling to improve prediction, but in many places you talk about "prediction uncertainty" and the like. Although in the last sentences of the paper you put clear what you missed, I think you must be cautious when talking about uncertainty in your paper, because your numerical experiment is related to uncertainty only in a theoretical sense. In my opinion, your results are only about prediction (fit).

Reply: Model uncertainty (predictive uncertainty) is generally composed of stochastic and structural uncertainty. Whereas a lot of work has been put into the investigation of stochastic uncertainty in recent years, less effort was put into the investigation of structural uncertainty. We see our work as a contribution in that direction. The only other work that we are aware of that deals with model structure uncertainty is presented in the Euroharp project (Kronvang et al., 2009a, 2009b). However, their work mainly focuses on a model intercomparison and not so much on different fusion techniques.

3. There is reasoning in the discussion that clearly violates your own statements about the limitations of your work. You stated that because model results themselves were not a target, individual model efficiency may not be maximized.

Reply: No, we do not agree. Of course model results were a target and that is why we selected the best model runs for further analysis.

However, in the discussion you used particular model results to raise some conclusions (page 5318, lines 20 to 24). In this paragraph you argue that since improved water description did not result in better nitrogen dynamics description when comparing LASCAM and LASCAM-S, then you concluded that improved water description does not necessarily provide better nutrient export prediction. A part from being at odds with your own statements about your work limitations (if you acknowledge that the calibration of individual models may not be totally efficient then you must be cautious when comparing performance, as you wisely stated fithe. I do not understand why you did not follow your own wise recommendations later), in my opinion your reasoning is totally wrong. First, it is nonsense to say that improved streamflow description does not lead to improved nutrient flde calculation. If nothing else changes, by definition improved water routing description leads to better flux prediction. It is just maths.

Reply: We agree that it would be nonsense to say that improved routing did not provide better flux predictions – but only if the hydrological cycle in the models are tightly connected to the nutrient cycle. It is also well known in mathematics, that if one of two uncorrelated parameters is changing, the other just does not react. As you can see in our results improved hydrological predictions of LASCAM-S even lead to worse predictions in nitrogen, indicating that hydrology and biogeochemistry are only tightly linked. We will improve the description and discussion of these results in a revised version of the paper.

Second, in your case you are comparing two models that are identical but in some of the water routing routines. That is, the formulation of nutrient dynamics is the same. Then, if the model that performs *better with hydrology is performing worse for the nutrients, this only means that the Monte Carlo scheme you applied to find the best parameterizations are not optimal. I mean, probably 40000 realizations were not enough to catch the global minimum for these two models. This poses in doubt your statements in lines 26-29. Then, you cannot compare model fit to raise those conclusions, because your models can be used as heuristic tools only in case of fit to an almost global minimum.*

Reply: True, 40,000 model runs are not sufficient for parameter intensive models to find the global optimum. But again, this is not the objective of the paper. We selected the ten best model runs of the entire parameter set and further took advantage of the equifinality concept to obtain global better predictions by using model ensembles and fusion.

4. This leads me to a rather philosophical question. In the discussion, you seem to advocate in favour of applying models in an ensemble fashion without regard of the internal conceptualization inside the models (page 5320, lines 1 to 5). However, I wonder which the value of models is if not as heuristic tools (see papers by Naomi Oreskes). If you forget the conceptualizations, are not we loosing all the science behind? Do you mean that ensemble modelling is better that model design in terms of knowledge generation?

Reply: As stated in the introduction of the manuscript, models are just mathematical assumptions of the real world and they just give estimations of the state of the system they simulate. Ensemble modelling, and more precisely model fusion, has been proved to usually provide estimates which fits better the actual data than single model predictions. If a model user is only interested in getting better predictions, then s/he is likely to consider the models, or ensembles, as black-boxes and not consider the description of the underlying processes. Therefore, s/he would benefit from using an ensemble with a global better predictive quality. This was the main point of the somewhat empirical approach that has been used in our article.

Of course, a different user may be interested in the description of the underlying processes (flow partitioning, residence times etc...) and just use the prediction of discharge as a fitting parameter to assess the global evolution of the state of the system. From this point of view, the conceptualisation is really important and one would benefit from using a more ‘correct’ model structure. Still, the quality of the ensemble relies on the design of its members. We can therefore assume that if data-fusion is realised by merging already good results the final prediction will be getting better.

Finally, science should not be forgotten at all in the sense that we still need to evaluate, and presumably improve, the description of processes which are conceptualised and implemented in our models. In a revised version we will include some points of this discussion in the conclusion.

5. With no doubt, you need to explain better contents in section 2.3.2. I still do not understand why you have 55 MME predictions and not 75. The whole section is hard to follow.

Reply: We used 4 models per nitrogen species as CHIMP cannot simulate total N and HBV-N-D only simulates total N. From 4 models you can derive 1 combination of 4 models, 4 combinations of 3 models and 6 combinations of 2 models; a total of 11 combinations. As we used 5 merging schemes, it results in a total of 55 MMEs. Given that the median and the average for a MME using 2 models is the same, we could even have only considered only 44 MMEs. But we see that this is poorly described in the paper and we will improve this in the revised manuscript.

Minor comments:

In the Introduction, the scale of your work is not clearly stated. You start with broad statements about global biogeochemistry, but your work is about the catchment scale.

Reply: We work on the catchment scale. The introduction will be revised before re-submission.

I do not think you must work with loads, but if you want to defend this you must explain how did you calculate them (page 5306).

Reply: As written in the discussion paper (P5306 L10), we considered the available measured concentrations as representative of the mean daily concentrations. Daily loads are then simply computed by multiplying the concentrations by the water discharge of the corresponding day.

Page 5316, line 8. Figure 4????

Reply: Figure number is wrong and will be changed.

Reference

Kronvang, B., Behrendt, H., Andersen, H. E., Arheimer, B., Barr, A., Borgvang, S. A., Bouraoui, F., Granlund, K., Grizzetti, B., Groenendijk, P., Schwaiger, E. et al.: Ensemble modelling of nutrient loads and nutrient load partitioning in 17 European catchments, J. Environ. Monit., 11(3), 572-583, doi:10.1039/b900101h, 2009a.

Kronvang, B., Borgvang, S. A. and Barkved, L. J.: Towards European harmonised procedures for quantification of nutrient losses from diffuse sources--the EUROHARP project, J. Environ. Monit., 11(3), 503-505, doi:10.1039/b902869m, 2009b.