Reply to François Anctil's (Referee #2) comment on manuscript "Using multi-model averaging to improve the reliability of catchment scale nitrogen predictions" by J.-F. Exbrayat et al.

In the following we provide answers to referee #2's points of discussion. For clarity, the comments of the referee were copied in black while our answers are in blue.

This paper, which promotes multi-model averaging, provides an interesting perspective of the pros and cons of simple ensemble averaging (SEA) and of reliability ensemble averaging (REA) for catchment scale nitrogen predictions based on four models of various complexities. The paper is well written, easy to read, and informative. Suggestions for improvement follow.
We thank Professor Anctil for his positive comments and general interest in the present work. We provide answers to the discussion points he raised and will address them in a revised version of the manuscript.

Principal comments

The pool of proposed models is rather limited for an ensemble study. I understand that some catchment scale nitrogen prediction systems are quite time consuming to implement and that the number of plausible options are limited, but I do not think that we can produce generalizable conclusions from four model and a series of scenarios reducing the fertiliser application on a single watershed. Still, the authors advocate, from such a set-up, that REA is superior to SEA. But is it true for all possible models and watersheds? Authors need to be more modest in their discussion and conclusion.

There are not many freely available nutrient mobilisation and transport models developed for mesoscale catchments (100 – 10,000 km²). A recent review by Breuer et al. (2008) listed a total of 8 model approaches that are used to simulate the N cycle in catchments. Among these 8 model structures, several are actually modifications of the same common ancestor; hence they share parts of their parameterisations. Our ensemble seems in fact to cover a large part of the available modelling philosophies in terms of simulated N-species and turnover processes as well as spatial distribution.

Of course, 4 members constitute a tiny ensemble in comparison to studies published in other fields (e.g. IPCC simulations, DMIP, etc…). Nevertheless, ensemble studies focusing on water quality and nutrient losses are still rare in the literature and this contribution is a further step in the innovative direction adopted by our working group as documented in previous contributions (Exbrayat et al., 2010 and 2011). We agree however that our results need to remain in a more local context and we will correct places of the text where we previously extrapolated their significance more than justified. The conclusion will be changed to:

> Through our straightforward example of fertilization rate reduction we demonstrated the potential advantage of using a multi-model ensemble to lower the risk of relying on a single, maybe subjectively chosen, model structure. This is a real advantage in our application case since the actual effects of different changes (management, climate) are not yet known,

making the evaluation of model quality impossible. So far, REA and similar averaging schemes have been primarily applied in climate and hydrological sciences and more work is still required in this direction to address their effect on predictions. We therefore see some potential of this technique in the ensemble approach in other fields of environmental modelling where the structural uncertainty of models used for predictions is large and rarely addressed.

I do not understand why the authors limited their ensemble experiment to the fertiliser application scenarios. They could also have used some of the metrics in Table 2 to compare REA and SEA. For instance, I would have liked to know if REA and/or SEA, when applied to the four models at hand, produce better RMSE values than the LASCAM model.

We wanted to focus our results on the scenario part, as it is where the most of uncertainties exist, and why predictive model are usually developed. The reviewer is however right in advocating the presentation of results obtained with the averaging techniques for the validation period, i.e. without changes in fertiliser application, as it constitutes a plausible scenario as well as the baseline to which the other scenarios are compared.

We have done the corresponding calculations for the validation period only as it constitutes the true 'prediction' period and a corresponding paragraph will be added to the manuscript:

> The validation period also corresponds to the control scenario. We therefore present corresponding results for a simple average of the predictions and the REA average in Table 2. Here, the REA average is only calculated with the reliability criterion as no perturbations have yet been made to our system. The simple average performs with a RMSE equal to 8.6 g N ha^{-1} d^{-1} which is worse than LASCAM but better than the other three models. However, the corresponding average export on sampled days is, at 0.42 t N d^{-1}, closer to the observed 0.41 t N d^{-1} than any of the single models. Meanwhile, the REA average outperforms all the ensemble members with a value of 6.5 g N ha^{-1} d^{-1}. This represents an improvement of about 10% compared to LASCAM, the best performing single model. The simulated mean export on sampled days equals the observed mean.

This allows us to better trust the REA procedure in scenario analyses for which we do not have any data to compare simulations with. This point will be further acknowledged in the discussion part:

> Interestingly, the REA average outperforms any of the other simulations in the control case for which we have data to compare with, therefore giving more credit to the approach.

Page 2292, line 13: we can read ". . .in spite of the demonstrated improvement in pre-diction reliability." Multimodel averaging is not magic; it does not always work. The authors need to nuance their statement.

See our answer to the principal comment.

A somewhat more nuanced statement is found page 2296, line 22: “Previous studies on multi-model averaging techniques set in a variety of environmental modelling contexts have demonstrated that the simple mean of a MME usually outperforms its members taken separately in terms of goodness-of-fit metrics.” The problem is that when people find a situation for which ensemble averaging does not work; they tend to not publishing their findings.

This is a very interesting comment. As stated in the cited part of the manuscript and according to previous studies model averaging is usually expected to outperform single members. In the case it would not, the averaging does not represent an improvement of any sort and may not be considered worth publishing by other researchers although this could contribute to some interesting benchmark studies on averaging. Averaging results however depend on the members of the ensemble and if they are already ‘too good’, there is not always enough space left for improvement (e.g. Viney et al., 2009).

Other comment

Page 2293, line 7: “Because of the sandy nature of the soils, evaporation is high (2000 mm yr−1)”. The authors are certainly discussing of the potential evapotranspiration. Evaporation cannot surpass precipitation.

This sentence will be rephrased to:

> Pan evaporation is high (~2,000 mm yr-1) and because of the sandy nature of the soils, runoff is mostly generated as a quick and peaky response to rainfall events which explains a five-fold difference between minimum and maximum annual discharge over the study period.

References

Breuer, L., Vaché, K. B., Julich, S. and Frede, H.-G.: Current concepts in nitrogen dynamics for mesoscale catchments, Hydrolog. Sci. J., 53, 1059–1074, doi:10.1623/hysj.53.5.1059, 2008.

Exbrayat, J.-F., Viney, N. R., Frede, H.-G. and Breuer, L.: Probabilistic multi-model ensemble predictions of nitrogen concentrations in river systems, Geophys. Res. Lett., 38, L12401, doi:10.1029/2011GL047522, 2011.

Exbrayat, J.-F., Viney, N. R., Seibert, J., Wrede, S., Frede, H.-G. and Breuer, L.: Ensemble modelling of nitrogen fluxes: data fusion for a Swedish meso-scale catchment, Hydrol. Earth Syst. Sci., 14(12), 2383–2397, doi:10.5194/hess-14-2383-2010, 2010.

Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H.-G., Graff, T., Hubrechts, L., Huisman, J. A., Jakeman, A. J., Kite, G. W., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Willems, P.: 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., Kite, G. W., et al.: 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, doi:10.1016/j.advwatres.2008.05.006, 2009.