Reply to Anonymous Referee #1 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 #1's points of discussion. For clarity, the comments of the referee were copied in black while our answers are in blue.

General Comments

I enjoyed reading this paper which presented an interesting comparison between 4 different models for simulating catchment N losses. A key point of interest to me was that whilst one of the models apparently gave a good simulation of the baseline condition and might be considered one of the better models on this basis, it's response to a simple management change was highly inconsistent with the other three models. To me this highlights many of the problems that arise from the use of inadequate data to calibrate and validate models.

The referee raises here an interesting point of discussion: hydro-biogeochemical modelling generally suffers from the lack of high frequency and high quality datasets over extended periods of time. Moreover, most of the impact studies addressing catchment scale nutrient losses have done so using a single model structure and a single parameter set, hence disregarding the uncertainty existing in parameterisations and parameter values.

Related to this, I was surprised that the calibrated models could end up with such widely varying estimates of TN loads exported as I would expect this to be a basic tuning component of the model calibration.

Differences in behaviour between models forced by the same data are actually quite common and can be found in literature for a large variety of environmental modelling contexts (e.g. IPCC and SRES scenarios). In the case of nutrient losses, a previous study by Kronvang et al. (2009) shows differences of up to an order of magnitude in simulated N losses for some of the catchment they studied. This can be explained by the lack of consensus in the way catchment scale N cycle is represented in hydro-biogeochemical models (Breuer et al., 2008). Simulation of the water balance greatly impacts nutrient losses and RMSE is more sensitive to the correct timing of peak events. As shown in Table 2, SWAT and HBV-N-D runoff predictions are of quality comparable to CHIMP during validation. However, Figure 2 and Table 3 clearly suggest that SWAT globally overestimates and HBV-N-D underestimates the TN losses, i.e. SWAT good matching of peak events is accompanied by a constant high discharge while HBV-N-D simulates lower flows. This has been integrated in the revised manuscript such as:

> Simulation of the water balance greatly impacts nutrient losses and RMSE is more sensitive to the correct timing of peak events. As shown in Table 2, SWAT and HBV-N-D runoff predictions are of quality comparable to CHIMP during validation. However, Figure 2 and Table 3 clearly suggest that SWAT globally overestimates and HBV-N-D underestimates the TN

losses, i.e. SWAT good matching of peak events is accompanied by a constant high discharge while HBV-N-D simulates lower flows.

In many ways I found these aspects of more interest than the stated objective of demonstrating that multi-model averages improve the reliability of catchment predictions. Indeed, I was left more with the feeling that I would want to interrogate the four model simulations in greater detail, in order to establish what their strengths and weaknesses are. Ultimately this might lead me to reject one or more of the models on the basis that its structure inadequately represents the processes that are believed to govern nitrogen losses in this catchment

We agree that there is value in benchmarking the performance of each single model, but this falls beyond the scope of our paper. We further argue that the ensemble contains a certain amount of information to which each model, whether best or worst performer, participates. Model averaging techniques aim at extracting this information in the best way and an apparently "bad" model may still contribute to the improvement in the final prediction. Finally, the way to discriminate between good and bad models is often too subjective (threshold on objective function, quantile of 'best' simulations according to one objective function, etc…).

I would very much like to see some revisions to this paper that present more analysis of the individual model behaviour. Even simple comparisons of time-series simulations of N losses would be advantageous as they would demonstrate some of the differences in temporal dynamics that the different models simulate. Inter model comparison of the different N species (where simulated) would also be interesting. In addition, I wonder what influence intra-model parameter uncertainty could have on the results presented – e.g. would a different selected parameter set for the CHIMP model give the same response to changing fertiliser usage?

We agree that addressing the stochastic uncertainty linked to parameter values is another point of interest. Different parameter sets are likely to provide different responses to changes in fertiliser application. However, we made our selection of parameter sets based on the calibration / validation process. Since SCE-UA has been demonstrated to successfully find optimal (or near-optimal) parameter solutions in a large range of non-linear problems, we argue that the four single models we introduced in our ensemble constitute the best (or near-best) outcome their structure and parameterisations can produce for the studied period.

In general terms I found the paper easy to read and understand. In places the grammar is slightly strange, but is usually understandable. I recommend publication of the paper following inclusion of some additional analysis as outline above.

We hope that our replies satisfy the referee's concerns and we will integrate them in a revised version of the manuscript.

Specific Comments

P2290 l18 – I disagree that is "is always sensible" to avoid disqualifying any of the ensemble

members. Rather, if the model simulations clearly conflict with known behaviour then it seems unsound to retain them in the final analysis.

In a true prediction exercise like our scenario analyses, the behaviour may not be known beforehand and we have to consider each model as being an equally good estimate of reality. Furthermore, there probably exists a part of the prediction where even the overall less well-performing model will outperform the other ones. The philosophy surrounding the model averaging procedure is to extract the most relevant information from each of the ensemble members to produce a more robust final output. Therefore, we argue that it makes more sense to keep a model and assign it a low weight rather than totally disqualifying it and losing its information content. One should keep in mind that by averaging predictions, we do not explicitly consider the processes represented in each model, but just the prediction they provide.

P2293 l29 – It would be interesting and informative to use a broader set of objective functions for the model calibration procedure e.g. RMSE and Nash-Sutcliffe efficiency for N concentrations rather than just loads as these give much more information about the N dynamics rather than just the hydrological forcing (which dominates the load calculations). Also presentation of some objective functions for hydrological simulation would be informative. Is inadequate simulation of the hydrology a limitation in the capability of the HBV-N-D model to simulate runoff, or is it linked to the N budget?

We only provided the results of the calibration procedure itself, i.e. the reduction of the RMSE by the SCE-UA algorithm over the corresponding period. Providing concentrations results is not fully relevant to our problem as models were calibrated to predict decadal losses. Furthermore, calibrating to match concentrations is mathematically difficult in ephemeral streams where the denominator in the model's calculation of concentration (i.e., streamflow) goes to zero. Still, we agree that the quality of the hydrological calibration / validation is an important piece of information for the evaluation of the models and results will be implemented in a table in the manuscript.

P2294 l7 – I don't understand the rationale for why independence of predictions leads to errors cancelling out?

In their introduction, Abramowitz and Gupta (2008) state that different modelling groups produce "quite different models" that can be considered as independent from each other. Masson and Knutti (2011) further demonstrate that a genealogy of climate models can be constructed with a hierarchical clustering based on the Kullback-Leibler divergence between models by only considering a single predictive variable (either temperature or precipitation). They show that models developed at one institution are more similar to each other and may therefore share biases. By using more independent representations of a natural system, we obtain sets of simulations that are not biased in the same way and more likely to compensate each other.

P2296 l10 – Coming from an environment where inorganic N dominates the N load in most rivers I am interested in the processes governing such dominance of DON in the stream loads. Is this linked to the fact that the land was previously forested and that the soils have extremely high organic N

content as a result of this? If this is the case, then it seems unrealistic to expect that changing fertiliser inputs would lead to any significant change in the stream N response, at least in the short term? A better overview of the N dynamics of the catchment would be beneficial for an international audience.

The catchment has indeed been extensively cleared since European settlement in the 19th century. According to Petrone et al. (2009), the Ellen Brook catchment is located on a clear-cut, highly weathered alluvial plain with little clay or silt content and these sandy soils have a limited capacity to adsorb dissolved organic matter (DOM). The export of DOM is significant during the wetter season but DON that accumulates in the groundwater slowly discharges in high concentration to surface water during dryer periods (Donohue et al. 2001). This will be documented in the revised manuscript such as:

> Soil texture does not allow the adsorbtion of large quantities of dissolved organic matter (Petrone et al., 2009). Furthermore, dissolved organic nitrogen that accumulates in the groundwater slowly discharges in high concentration into surface water during the driest months (Donohue et al., 2001).

We are conscious that some action is needed in the region of Perth to reduce the frequency of occurrence of algal blooms. However, our study does not aim at finding a solution to this particular eutrophication problems met in the Ellen Brook and Swan-Canning Estuary. Due to the nature of perturbations, other measures with immediate effect have been adopted to improve the quality of the water flowing in the Ellen Brook like fencing the stream to avoid animals to enter it. However, Donohue et al. (2001) reported that the residence time of organic matter in soils over the whole Swan-Canning basin is rather long and that the concentrations of both N and P in surface waters may remain unchanged for a certain period of time despite management practices. Here, we take advantage of the extensive monitoring undertaken in this area to more confidently apply our models.

P2298 l15 – I am surprised that different models supplied with the same input data in terms of N inputs etc. can simulate such widely differing loads. How do the N budgets of the outputs stack up against the inputs?

See our answer to general comments.

P2300 l18- Although the outlying position of CHIMP decreases its reliability in the weighting scheme, it still shifts the ensemble average significantly. This may be less of an issue where many models are used within the ensemble, but for a small sample size of four its influence could be considered as inappropriate in these circumstances.

This point was already addressed in the first version of the manuscript:

> Further, because of the outlying position of CHIMP, the simple mean provides a final prediction equivalent to an almost 25 % reduction in nitrogen losses when no fertilization occurs. However, the trust we can put in this projection is questionable since it is not really in agreement with any of the single projections and that its intermediary position is merely a result of

very different but equally weighted projections.

The simple mean equally weights all the models, therefore the outlying position of CHIMP drags the average towards a value far from any of the ensemble members, around 25% of reduction (dashed line in Figure 3). By taking a measure of the convergence between models into account, we partially remove this problem and the average REA (solid line in Figure 3) predicts a reduction of about 10% in TN losses. This is well in agreement with LASCAM, HBV-N-D and SWAT while still using CHIMP in its calculation for which the convergence criterion decreases at with the reduction in fertiliser application.

P2300 l29 – To me this highlights the importance of testing model structures using high quality data sets in catchments typical of their intended application, rather than blindly transferring models which have been developed for other regions of the world.

High quality data sets are not freely available everywhere and we are not conducting a test of the models themselves, but of the REA method. The ensemble approach can be considered as the current state-of-the-art option to extract the relevant information from models. It can be used to show the place left for improvement and reduction of the uncertainty in the understanding we have of particular processes by also allowing a direct comparison of models. The overall gain of quality of the REA might be enhanced by modestly performing models.

Technical Corrections

P2290 l23 – would normally use “and or” rather than “or and”

We removed the “or”.

P2292 l8 – delete “simulation”. In this case the “models” themselves are also different

(even if the same model “code” is used)

This was done.

References

Abramowitz, G. and Gupta, H.: Toward a model space and model independence metric, Geophys. Res. Lett., 35, L05705, doi:10.1029/2007GL032834, 2008.

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.

Donohue, R., Davidson, W. A., Peters, N. E., Nelson, S. and Jakowyna, B.: Trends in total phosphorus and total nitrogen concentrations of tributaries to the Swan–Canning Estuary, 1987 to 1998, Hydrol. Process., 15(13), 2411–2434, doi:10.1002/hyp.300, 2001.

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, 2009.

Masson, D. and Knutti, R.: Climate model genealogy, Geophys. Res. Lett., 38(8), L08703, doi:10.1029/2011GL046864, 2011.

Petrone, K. C., Richards, J. S. and Grierson, P. F.: Bioavailability and composition of dissolved organic carbon and nitrogen in a near coastal catchment of south-western Australia, Biogeochemistry, 92(1-2), 27–40, doi:10.1007/s10533-008-9238-z, 2009.