Geosci. Model Dev. Discuss., 5, C748–C750, 2012
www.geosci-model-dev-discuss.net/5/C748/2012/

Geoscientific
Model Development
Discussions

Interactive comment on "Using multi-model averaging to improve the reliability of catchment scale nitrogen predictions" *by* J.-F. Exbrayat et al.

Anonymous Referee #1

Received and published: 3 October 2012

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. 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. 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.

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?

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.

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.

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?

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

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.

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?

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.

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.

Technical Corrections

P2290 l23 – would normally use "and or" rather than "or and"

P2292 l8 – delete "simulation". In this case the "models" themselves are also different (even if the same model "code" is used)

Interactive comment on Geosci. Model Dev. Discuss., 5, 2289, 2012.