Response to reviews

Title: Coupling global models for hydrology and nutrient loading to simulate nitrogen and phosphorus retention in surface water. Description of IMAGE-GNM and analysis of performance

Authors: A.H.W. Beusen, L.P.H. Van Beek, A.F. Bouwman, J.M. Mogollón, J.J. Middelburg

We are very grateful to the two reviewers for their constructive feedback. The suggestions for better-input data from reviewer 2 will definitely lead to significant improvement of next versions of the model. Reviewer 1 had a concern about the validation data used for the Mississippi, which we will address below and in the revised manuscript. Below are the **reviewer comments in bold**, our response is in regular text, *new text that will be included in the revision of our paper is in italics*.

REVIEWER 1
The authors introduce the IMAGE-GNM model, which builds in hydrology-based N and P loading and retention into the existing IMAGE model. The model is a great improvement over the existing Global-NEWS model, in that it resolves to 0.5º x 0.5º grid cell size, rather than lumping processes together in regression equations that can only be resolved at the watershed scale. The model is also set up for future mechanistic improvements that can delineate the behaviour of different N and P species. Their modelling approach is well described and presented in a logical, transparent manner. There are a few minor details in the model validation/discussion (see below) that can be improved upon, but overall I recommend this manuscript be accepted for publication in GMD.

Specific comments:
- While the model is developed at the 0.5 x 0.5 grid cell size, it is unclear at what scale the model's output is actually valid. The discussion in section 3 comparing model results with data from the Mississippi, Meuse, and Rhine Rivers seems to rely on data from a single monitoring station (at least for the Mississippi; the number of locations used for the Meuse and Rhine is less clear). The Mississippi is a huge river, so I'm wondering how this one particular monitoring location was chosen for model comparison. It seems to me that, given the number of monitoring locations on the river, any number of sites will yield good correlation with model output (and also any number will yield poor output) just based on the variability of the river and the landscape. This discussion needs to be developed a lot more with comparison to additional stations in the river, or at least a justification for why this one particular site in St. Francisville, LA was used.

Response: The Mississippi station St. Francisville was chosen for validation due to its widespread usage in scientific studies, for example the USGS Nutrient Trends in Streams and Rivers of the United States, 1993–2003. National water Quality Assessment Program (U.S. Geological Survey, 2009). Since it is quite close to the river mouth, it encapsulates the integrated effects of the whole river basin. In the revision we include 10 more stations located throughout the Mississippi. The locations are those selected by USGS in their 2007 open file report (U.S. Geological Survey, 2007). For the 11 stations in total (including St. Francisville) we calculated the RMSE values and added figures to the supporting information showing the comparison for concentrations of N and P, the load of N and P and the discharge (see new

Table 4 below). Results confirm the reviewer's concern, i.e. there are some stations where the model is poorly simulating the N or P concentrations.

We added the following text to the discussion in the first paragraph of section 3.1:

We first compared the IMAGE-GNM model results with observed concentrations for two stations (rivers Rhine and Meuse) in The Netherlands and at 11 stations in the Mississippi, USA (see SI1). Stations near the river mouth (Lobith at the Rhine, Eysden at the Meuse, and St. Francisville, Louisiana for the Mississippi) are shown first. The latter station was selected for comparison with the U.S. Geological Survey analysis of water quality (U.S. Geological Survey, 2009). The measured concentrations were aggregated to annual discharge-weighed concentrations, whereby for the U.S. data years with <6 observations were excluded.

The following references will be added to the list of literature:

U.S. Geological Survey: Streamflow and nutrient fluxes of the Mississippi-Atchafalya river basin and subbasins for the period of record through 2005. Monitoring network for nine major subbasins comprising the Mississippi-Atachafalaya river basin. USGS Open-File Report 2007-1080 (http://toxics.usgs.gov/pubs/of-2007-1080/major_sites_net.html) (accessed 6 November 2015), 2007.

U.S. Geological Survey: Nutrient Trends in Streams and Rivers of the United States, 1993–2003. National water Quality Assessment Program, in, edited by: Sprague, L. A., Mueller, D. K., Schwarz, G. E., and Lorenz, D. L., 196 p., 2009.

Then, after the 4th paragraph in section 3.1 we inserted the following text about the model comparison for the 10 additional stations:

We also investigated the model performance for 10 more stations in various states within the Mississippi river basin (Table 4). These stations, along with the St. Francisville station, form the monitoring network for nine subbasins in the Mississippi (U.S._Geological_Survey, 2007). The plotted data for all 11 stations in Mississippi river basin are available as separate graphs in the SI. The model performance is acceptable (RMSE<50%) for 8 stations for N concentrations and 5 stations for P concentrations. There are some stations where the model poorly simulates the N concentrations such as Arkansas river and Red river (Table 4). Such high RMSE values do not occur for P. In general, simulated P concentrations are closer to observed values than N concentrations.

One of the reasons for poor agreement is the large fluctuation of discharge, load and concentration at some stations. Apparently, these peaks are associated with periods of high rainfall. We do not know if these peak values represent the full period of the measurement interval. For example, a peak value that represents two months (in the case there are 6 measurements per year) also yields a peak in the aggregated annual value. However, it is not known if this peak actually represents 1 day (with a much lower aggregated annual value) or two months. In contrast to St. Francisville, P concentrations (and N concentrations) at the other stations are not consistently underestimated or overestimated. Furthermore, at this level of comparison, the spatial data for land use and wastewater discharge locations in urban areas may not be realistic. For example, our wastewater discharge occurs in all grid cells with urban population, while in reality discharge may take place in discrete locations with wastewater treatment plants.

And Table 4 will be added, and the original Table 4 and 5 will be 5 and 6:

Table 4. RMSE for simulated versus measured N concentrations, N load, discharge, P concentration and P load for 11 stations in the Mississippi river, Ohio river, Red river, Missouri river and Arkansas river. Measurement frequency ranges from 28 per year to 3. Years with less than 6 observations were excluded.

Station id	*Name*	*RMSE (%)*				
		Discharge	*N concentration.*	*N load*	*P concentration.*	*P load*
5420500	*Mississippi River at Clinton, IA.*	*60*	*36*	*72*	*23*	*66*
3612500	*Ohio river at dam 53 near Grand Chain, ILL.*	*32*	*19*	*44*	*48*	*53*
5587550	*Mississippi river below Alton, Ill.*	*56*	*48*	*47*	*53*	*71*
7355500	*Red river near Alexandria, LA.*	*18*	*119*	*152*	*69*	*72*
7022000	*Mississippi river at Thebes, ILL.*	*67*	*49*	*34*	*64*	*52*
5587455	*Mississippi river below Grafton, ILL.*	*51*	*46*	*27*	*44*	*26*
3303280	*Ohio river at Cannelton dam, KY.*	*56*	*10*	*59*	*58*	*89*
6610000	*Missouri river at Omaha, NE.*	*35*	*74*	*76*	*88*	*78*
6934500	*Missouri river at Hermann, MO.*	*19*	*53*	*56*	*73*	*82*
7263620	*Arkansas river at David D. Terry L&D BL Little Rock, AR.*	*53*	*244*	*369*	*52*	*92*
7373420	*Mississippi river near St. Francisville, LA.*	*19*	*23*	*26*	*51*	*44*

- The discussion relating the model output to European rivers seems much more valid, as many monitoring stations on each river are compared. Here the authors also briefly mention that the model has problems when modelling individual stations on small rivers. Is it possible to elaborate on this statement in a more quantitative way? How small?

Response: An arbitrary choice has been made to exclude river basins with less than 4 grid cells (<10,000 km^2) because of poor spatial representation (land use, urban areas, etc.). Nevertheless, river basins with somewhat larger areas (4-10 grid cells) may also have this problem.

Although also mentioned in the SI, for clarity we will add the following explanation to the 5th paragraph of section 3.1:

River basins with less than 4 grid cells, of ~2,500 km^2 each, were removed because river basin areas of <10,000 km^2 do not have adequate spatial data representation. This is an arbitrary choice, and probably many river basins with 4-10 grid cells also suffer the problem of poor spatial data.

Technical comments: - in the readme file, “The python script for the N model can be started with:” is stated twice. The second time it should read P model.

Response: Technical comments: in the readme file, “The python script for the N model can be started with:” is stated twice. The second time it should read P model. This has been corrected.

Are the ratios on page 16, line 9-10 mass ratios or molar ratios? I assume mass, but maybe clarify so the reader does not need to go to the citations to double check.

Response: The ratio on page 16 is a mass ratio. It will be added to text.

Grammar error on page 4, line 28-29: “This global scale model focuses is on: : :”

Response: Grammar error on page 4, line 28-29 will be corrected.