First of all I must thank the authors for their reactions on all remarks and comments that were given. I think that these remarks have lead to a better manuscript. I have read the reactions and the comments and I think there are still three major concerns.

In this response, page, line, table, and figure numbers correspond to the revised manuscript.

1. Firstly I think that the coupling of the model with P is the new part of the model. I have my concerns on the quality of the implementation. The loss in-stream processes are about 56-59% for TN and about 6-10% for TP. This means that this model calculates on a global scale at least 6 times more N retention than P retention. This concerns and surprises me, and I think this is not consistent with the current literature. Common knowledge would be that there is equal or more P retention than N retention. See for example Maranger et al. (2018) but there are many other papers addressing this.

The study by Maranger et al. (2018) was acknowledged by the authors as "first attempt to consider coupled biogeochemical cycles of inland waters and their ecosystem level stoichiometry at the global scale" and further concluded that "not surprisingly a great deal more effort has been expended to understand the processing of C than in understanding the fates of N and P". They further note that: "Somewhat surprising was the fact that the N:P molar ratio did not change relative to the input ratio suggesting that these elements are processed similarly in freshwaters".

Given the authors' stated intention to provide a "first attempt", it is understandable that they did not delve too deeply into the uncertainties in this aspect of their conclusion. However, they acknowledge that the uncertainty in N inputs is large, citing values from 67 to 129 TgN yr^{-1} (Beusen, et al., 2016; Galloway, et al., 2004). A review of Galloway et al. (2004)'s analysis suggests that the upper bound should have been 118 TgN yr^{-1}, but this does not change the fundamental finding that the uncertainty is large. Maranger et al. (2018) find stronger agreement in the available P input estimates (9 TgP yr^{-1}). They opt for single estimates of riverine N and P exports to the ocean (43 TgN yr^{-1} and 4 TgP yr^{-1}, respectively), but the two estimates available from the literature suggest that these are also highly uncertain. Seitzinger et al. (2010)'s Global NEWS, for example, estimates an P export of 8.6 TgP yr^{-1}, which happens to be very similar to our model estimate 7 (7-8) TgP yr^{-1}. Furthermore, as the study of Harrison et al. (2019) suggests that "IMAGE-GNM-TP underestimates P export from high-exporting basins such as the Amazon, Ganges and Yangtze [44] rivers.", the 4 TgP yr^{-1} arising from Beusen et al. (2016)'s IMAGE-GNM tends to underestimate P exports from globally distributed large rivers.

As a simple exercise to illustrate the size of this uncertainty, if Maranger et al. (2018) had used the upper bound of the N inputs from Galloway et al. (2004, 118 TgN yr^{-1}) and P outputs from Seitzinger et al. (2010, 8.6 TgP yr^{-1}), they would have concluded that 4% of the P inputs was buried and 64% of the N inputs was lost to either burial or denitrification. While Maranger et al. (2018) did not conduct an uncertainty analysis, the very cautious language they used to describe

their first order conclusions clearly recognizes that their work was a first attempt at assessing a complex and highly uncertain set of processes operating on global scales, not a definitive analysis.

Also, we would like to emphasize that, by providing a direct comparison against observed exports of N and P to the ocean (i.e., Figs. 3-4, Table 4), we have tested the consistency of our solution with existing constraints. We do not claim that it is unique but, given the inputs asserted, the simulated losses generate outputs that are consistent with observations and fall within the range of published uncertainties. We now demonstrate this explicitly in a newly added Table 6.

In addition, we have expanded the discussion in pages 27-28 and lines 600-606 as follows:

Globally, TN inputs to rivers (N inputs hereafter) in LM3-FANSY are 85 (85-91) TgN yr^{-1}, of which about 59 (56-60)% are lost to the atmosphere or retained within freshwaters (N loss/retention hereafter) and the other 41 (40-43)% are exported to the coastal ocean (N exports hereafter) (Table 6). LM3-FANSY thus estimates that N loss/retention is 144 (130-148)% of the N exports, consistent with the 147% and 143% of Galloway et al. (2004) and Seitzinger et al. (2006), yet larger than the 73% estimated by IMAGE-GNM (Beusen et al., 2016). LM3-FANSY estimates of 59 (56-60)% of the N inputs are lost or retained in freshwaters, consistent with the 60% of Galloway et al. (2004), yet larger than the 42% of IMAGE-GNM (Beusen et al., 2015).

Finally, we have emphasized that the lack of reservoirs/hydraulic controls in our model may cause our estimated P burial to be on the lower end of uncertainty bounds, but it is not outside of the published uncertainty bounds (page 29 and lines 629-638).

Globally, TP inputs to rivers in LM3-FANSY are 8 (8-9) TgP yr-1, of which about 9 (6-10)% are stored within freshwaters and 91 (90-94)% are exported to the coastal ocean (Table 6). IMAGE-GNM estimates that ~56% (5 of 9 TgP yr-1) of the P inputs are stored within freshwaters (Beusen et al., 2016). This is a large difference from our estimate of 9 (6-10)% retention, but the difference is around a very uncertain number as the storage has not been directly measured. LM3-FANSY, which does not account for dams and reservoirs, likely underestimates global freshwater P retention by at least ~12% (Table 6, Maavara et al., 2015, see Sect. 4.5 for further discussion). The overall consistency of our SS, N, and P estimates with the observed cross-watershed constraints (Figs. 2-4, Table 4), however, suggests that the bias introduced by the lack of dams and reservoirs may not be large. In contrast, underestimates of P exports to the coastal ocean from high-exporting basins such as the Amazon, Ganges and Yangtze Rivers shown in Figure 2 of Harrison et al. (2019) imply that IMAGE-GNM likely overestimates global freshwater P retention.

The following references are newly added:

Maavara, T., Parsons, C. T., Ridenour, C., Stojanovic, S., D€urr, H. H., Powley, H. R., and Van Cappellen, P.: Global phosphorus retention by river damming. Proc. Natl. Acad. Sci. USA, 112, 15603–15608, 2015.

Seitzinger, S., Harrison, J. A., Böhlke, J. K., Bouwman, A. F., Lowrance, R., Peterson, B., Tobias, C., and Van Drecht G.: Denitrification across landscapes and waterscapes: a synthesis, Ecol Appl., 6, 2064-90, 2006.

2. Secondly, I do not agree with the calibration/validation method of this paper. To illustrate this, I give an example. Assume that the load (N or P) through time can be described as a straight line for a river i. The general equation is Load_i = slope_i * t + observation_i where t is the time. We choose the observation_i as the load of the river at time t=0. For time t=0 we do the calibration procedure as proposed by the authors. The R squared of this example is equal to one. Note that the slope_i is not chosen. This means that the trend of the lines is free to choose. It could go downwards, or upwards. This is exactly, why this calibration method is not good enough. The load could be correct for t=0, but the trend is not. The authors try to compensate this with a number of figures for the Mississippi, but this is not sufficient for a global model.

In the previous revision, our efforts of presenting 37 time series plots of measurement-based vs. simulated loads from large U. S. rivers (Mississippi, Columbia, and St. Lawrence) and rivers within the Mississippi River Basin (Figs. 8 and SI5) were our attempt to respond thoroughly to the reviewer's request to include time series analysis, not to compensate for any weakness of the analysis method with a large number of comparisons. To further expand our analysis, which focused on interannual variability and bias, we have now added an analysis of linear trends. Most trends (30 of 37) are insignificant over the observed time series at the 0.05 level (Table 7). In the 7 cases where the trend is significant, our simulation captures the sign of the trend, matching the significance level of 0.05 in 5 of the 7 cases.

All together, the time series analysis now evaluates the model's capacity to capture 1) interannual variability of the loads based on the Pearson correlation coefficients (r) between the measurement-based vs. simulated annual loads, 2) bias of the loads based on the prediction errors of the measurement-based vs. simulated mean loads over the time periods, and 3) linear trends of the loads over the time periods based on the significant tests and signs of the slopes.

The associated result and text have been added in Table 7 and on page 36, lines 747-749 as:

Lastly, the t-statistic shows 30 of the 37 measurement-based time series loads have no significant linear trends over the time periods at the 0.05 level (Table 7). LM3-FANSY captures 30 of the 37 measurement-based trends in terms of whether the trends are significant, and when significant, all of the slope signs demonstrating downward vs. upward trends.

Finally, we agree that, in the one river case the reviewer uses in their example, a model's capacity to match the mean condition would have little bearing on its capacity to predict the change. However, our cross-watershed comparison assesses the model capacity to recreate loads and concentrations across systems with different temperature, precipitation, and other climate characteristics. We would argue that a model's capacity to capture watershed differences across spatial climate gradients is relevant to its capacity to capture watershed responses to changes in climate forcing over time. While not a perfect analog, such comparisons are common practice in global climate and earth system model evaluation. We feel that the combination of this spatial comparison with the extensive time series analysis described above provides an overall evaluation that is as extensive as any literature for our global application.

3. The third remark is both for the editors and for the authors. The authors mention the following line: "We also note that this is a model description paper, not just a model evaluation paper." I don't agree with this statement. Assume that this is true, than this paper should not be published because the equations are all not new. I think the interesting part of this paper is the combination of model description and the implementation and/or validation. This means that a paper should give a presentation of the used equations and used inputs and whether this chosen approach is working or not (validation). This means that this model must have a good validation over the historical time period before they can use it in scenario analysis. I think this is not the case yet.

We agree with the referee and point out to the editor that this sentence was in response to one of the comments rather than in reference to the entire manuscript. Furthermore, we feel that we have failed to adequately convey to the referee the value of our "description" which has led to a communication breakdown with the referee. As an example, the referee's assertion that "the equations are all not new" suggests a lack of familiarity on how prognostic biogeochemical models for use in global coupled climate and Earth system applications, such as the Coupled Model Intercomparison Project (CMIP), are constructed. While the general equations are certainly not "new" insofar as they always conform to mass conservation and the principles of nutrient-light-phytoplankton-detritus interactions in existence for nearly a century, each implementation is unique in drawing from a wide range of functional representations and other structural decisions that require dedicated calibrations such that the parameters used are unique, even going from version to version of the same model family. This is what makes the commitment of model developers to explicit "description" of the model details so important.

This is a standard for public data provision critical for model intercomparison such as in CMIP to facilitate analysts to understand the reasons behind model differences both in comparison to historical observations and future change. GMD has played a critical role in this regard in filling the void of journal opportunities to focus on "description" as a critical element in knowledge sharing and community science advancement.

Lastly, as extensively described in the previous response to the referee, we emphasize that we have augmented the validations further with time series comparisons suggested by the referee and expansive sensitivity analyses. The level of time series evaluation presented in this study is equal to those presented in comparable model description papers in GMD (e.g., Beusen et al., 2015), and the cross-watershed evaluations of solids, N, and P, in different forms (SS, NO_3^-, NH_4^+, DIN, DON, TKN, PO_4^{3-}, DOP, and TP) and different units (yields, loads, and concentrations), across the 70 globally distributed rivers covering 55% of the land surface (excluding the Antarctic) exceeds any of other global watershed model evaluations presented in GMD and elsewhere (e.g., Beusen et al., 2015, He et al., 2011, Pelletier, 2012).

Reference

Maranger, R., Jones, S.E., and Cotner, J.B. (2018). Stoichiometry of carbon, nitrogen, and phosphorus through the freshwater pipe. Limnology and Oceanography Letters 3(3), 89-101. doi: doi:10.1002/lol2.10080.