This is the third time that I looked at this paper. The authors did not improve the model. So this new revision does not have new results from an improved model. I think that is a pity. The two major problems that still exists are:

We objectively responded to all of the reviewer's comments in good faith during both of the previous rounds. In our first revision, we did substantially modify and improve the model and manuscript to address both reviewers' concerns. Thanks in part to these reviews and thousands of years of additional simulations, the model now recreates observed river solid and nutrient loads and concentrations as well or better than other published global models and has been more thoroughly validated than any of them. This validation is based on comparisons against observations (Figs. 2-4, 8, and SI5-6, Tables 4 and 7), other models (e.g., Fig. SI2, Tables 5-6), and published uncertainty ranges (Tables 5-6). It includes the cross-watershed validation of temporal mean conditions across globally distributed rivers, the time series validation of temporal variability and trends, and the validation of global means. We relied on and, when requested, extended above and beyond community-wide practices for assessing global river models over the last several decades.

In our second revision, we demonstrated using a simple uncertainty analysis that, contrary to the reviewer's concern, our P losses due to retention and N losses due to the sum of retention and denitrification fall safely within the bounds of current published uncertainties (Table 6, pages 27-28, lines 605-613 and page 29, lines 636-647). We also directly and transparently addressed the reviewer's concern that "the load may be correct for t = 0, but the trend is not" by adding a comparison of time trends against available time series from large rivers. While the availability of climate-scale time series for such a comparison is limited, the model trends are generally consistent with those observed (Table 7, page 36, lines 756-758). Further model improvements were thus not required to address the reviewer's second round of comments. We are thus surprised to see more general elements of model skill raised as a major issue in our third revision.

In this third revision, we have again endeavored to directly and thoroughly address the concerns raised by the reviewer. We have clarified that P retention in freshwaters is indeed larger than N retention, and we have addressed the new questions of model skill the reviewer has now raised. Our detailed responses are provided below. We objectively show that our model meets or exceeds current published global river modeling standards set by widely used and successful models, including models published in GMD. This is discussed explicitly in this response and, if any doubt remains, can easily be further verified by the editor and reviewer with the references we provide. The reviewer's contention that our model is not up to GMD and community skill standards is thus verifiably false. Further reviews will not change this fact. Nor will they change the corollary that the most widely used and successful global river nutrient models to date would have never been published if were the reviewer's criteria applied.

We share the reviewer's desire for improvement. We are cognizant of global model skill challenges we must meet as a community in the years to come and now highlight these limitations more in the Abstract. In particular, we emphasize that while the model shows considerable skill in matching contrasts of nutrient and SS loads and concentrations across globally distributed rivers, misfits for any single river can be substantial. This is a ubiquitous limitation of models designed for global robustness over regional optimality and we have no qualms against calling the communities attention to it. We hope that this additional transparency helps address the reviewer's concern.

We are proud of the results herein, which reflect the hard won fruits of years of development. We hope to have the opportunity to present what is now a uniquely thorough documentation of a skillful model that includes frank discussion of the model strengths and weaknesses to the Geoscientific Model Development community in a manner consistent with Geoscientific Model Development's mission.

Our detailed responses to each of the reviewer's specific comments are provided below.

1. the RELATIVE difference between N and P retention is huge and mostly there is more P than N retention (in percentage). The new text does not answer this aspect.

Our response to the reviewer's previous comment focused on contrasting the relative freshwater N and P loss mechanisms that control how much N and P made it to the river mouths. This breakdown is more commonly discussed in the literature, allowing for comparisons with additional published estimates. We demonstrated, in our previous response, that the high total loss of N relative to P when expressed as a fraction of their respective inputs falls within current uncertainty bounds, as acknowledged by Maranger et al. (2018). A key to this result, however, is that N losses include both denitrification and retention/burial, while P losses are just retention/burial. Upon re-reading the reviewer's previous comment, we see that the language shifted from "loss" at the outset of the comment to "retention" at the end. We focus more specifically on retention here.

Our retention results are consistent with the reviewer's expectation: P retention is higher than N retention when expressed as a fraction of the inputs. The 6-10% of P inputs lost due to freshwater processes are all from retention/burial, mainly reflecting the sorption of inorganic P onto solids and its deposition to bottom sediments. While organic N and P can accumulate within and be removed from the bottom sediments in response to variations in supply and remineralization in FANSY, the model does not include net long-term organic burial. Thus, effectively all of the 56-60% of N inputs lost due to freshwater processes are from denitrification. The model thus

estimates that P retention is indeed larger than N retention as a fraction of the inputs. This is now clearly stated on pages 27-28, lines 605-613 and page 29, lines 636-647.

The degree of net long-term organic N and P burial on a global scale is not well constrained. Maranger et al. (2018) estimated that around 15% of N inputs are lost due to burial. This estimate, however, is subject to uncertainties of similarly large scale to all other aspects of Maranger's first order budget. As the authors themselves state:

"For N, freshwaters appear to bury around 15% of terrestrial inputs (16 Tg N yr-1 , Table 1), however as outlined above, this number is based on the difference of a relatively imprecise loading estimate."

Net long-term organic burial could be added to FANSY as a simple fractional efficiency but, in the absence of better constraints and mechanistic contrasts between regions, we would be blindly shifting a secondary loss process. Instead, we now discuss the lack of net long-term organic burial in the model and the need to address this in the future through improved observation-based constraints and more realistic models of sediment diagenesis (page 43, lines 847-857).

The simplification of sediment dynamics is indeed shared by all published global river nutrient models. While such advances are needed, they would require a multi-year effort beyond the scope of this already comprehensive paper, and ultimately still suffer from a paucity of constraints. This is supported by the fact that we find that a model with no net long-term organic N and P burial can reconcile basic observed contrasts between N and P inputs and those observed at the river mouths within their admittedly large uncertainty ranges.

The model estimates of N and P retention and their comparison is now clearly stated as:

On pages 27-28, lines 605-613 as:

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 does not include net long-term N burial to bottom sediments, as all organic N delivered to the sediments is ultimately remineralized. While year to year sediment N inventories may vary, effectively all long-term N losses are to the atmosphere via freshwater denitrification.** LM3-FANSY 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).

On page 29, lines 636-647 as:

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, SI6, 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. **Even though our freshwater P retention estimates are near the lower bound, our P retention estimates are far higher than those for N, mainly reflecting the sorption of PO_4^{3-} onto solids and its deposition to bottom sediments.**

Discussion of the lack of net long-term organic burial in the model and the need to address this in the future is stated on page 43, lines 847-857.

The Rouse number-dependent transport criterion from Pelletier (2012) was adapted to simulate the deposition/resuspension fluxes between the suspended matters (i.e., ISS, PON, POP and PIP) and benthic sediments (i.e., Sed, SedN, and SedP). The criterion was designed to primarily simulate suspended loads, typically accounting for > 80% of total (i.e., suspended and bed) loads from most large (> ~100 km^2) river basins (Pelletier, 2012; Turowski et al., 2010), without explicitly modeling benthic sediments. **We acknowledge that our simplified benthic sediment component resulted from adapting the Pelletier's approach drives uncertainties in modeling the suspended matters and benthic sediments, including important diagenesis, other biogeochemical transformations, and physical processes (e.g., mineralization, denitrification, and net long-term organic burial) that occur within the benthic sediments. An implementation of more sophisticated benthic sediment dynamics (Chapra et al., 2008; Di Toro, 2001) with improved observation-based constraints and bed load transport processes is thus subject to critical future work.**

2. The validation method.

We have broken up this comment to ensure that each part is addressed.

The one point validation is okay when you do a regression for a single time step. This model is dynamic and the authors first calibrate on this time point and then perform a validation on this same point.

The model was calibrated based on the 1990 year results, yet validated based on 11 years (1990-2000) of annual results for the cross-watershed analysis (Table 4) and ~38 years (~1963-2000) of annual results for the 37 time series analysis (Table 7, Fig. 8). These additional efforts were made in our first revision in response to the reviewer's request to evaluate dynamic results, and in our second revision to add a time trend analysis. Thus, the model was not validated based on the single time point.

We also emphasize that FANSY is not a regression model. We are not fitting a simple set of predictor variables to river loads using a series of additive linear and/or non-linear relationships. Rather, our model, like others of its kind, is formulated as a set of interacting processes constrained as much as possible by mechanistic relationships and parameter values drawn from the literature. This overall mechanistic structure and a universal parameter set – the same parameters for all the basins (i.e., without tuning of each basin) – is imposed. We then perform a limited calibration of highly uncertain yet influential processes (e.g., denitrification) to remove first order biases (pages 17-18, lines 375-397). While the regression analogy can be useful, it is imperfect. The foundational validation comes from determining whether the model formulation, composed by both its mechanistic structure and mechanistic parameter values, can explain global contrasts in solid and nutrient concentrations and loads. This is a foundational comparison for almost all global models throughout the published literature (e.g., Mayorga et al., 2010; Tian et al., 2023; Pelletier et al., 2012). We have no objection to raising the validation bar and we have now added extensive additional evaluations at the reviewer's request. We strongly argue, however, that constructing a mechanistic process-based model that can skillfully capture many aspects of the observed global distribution of river solid and nutrient concentrations and loads is not a trivial result as the reviewer seems to imply. The literature is firmly on our side in this regard.

In addition, the cross-watershed validation method (based on the Pearson correlation coefficient (r) or the Nash–Sutcliffe model efficiency coefficient (NSE) of log-transformed data) that we have used in this study is not something that we have created, but it is the method that has been shared by almost all global river modeling communities for the last several decades, no matter whether a model is dynamic (e.g., DLEM-TAC, Tian et al., 2023) or not (e.g., Global NEWS, Mayorga et al., 2010). For sound reasons which we describe below, it is the standard approach

used when one is attempting to determine if a model can capture observed patterns that vary over orders of magnitude. This will be discussed further below.

My hypothetical example of straight lines is an example of a set of models which will have a perfect fit. But this example could also be for parabolic or hyperbolic functions. However the authors react with another statistical method to prove that the model is different than a straight line (I think, because I could not fully understand what they were now adding to the manuscript).

As we articulate above, we are not performing a regression analysis and the model will never have a perfect fit. The fit achieved reflects the degree to which the imposed process-based model structure, parameters, and forcing can represent the global patterns in river nutrient concentrations and loads.

The reviewer's concern about the calibration was:

"This is exactly why this calibration method is not good enough. The load could be correct for t=0, but the trend is not."

We responded by adding a comparison of time trends against observed ones from select rivers where climate-scale time series were available. We found that the trends in the model, which was not calibrated but emerged naturally from the imposed forcing and process-based model dynamics, are largely consistent with those observed (Table 7, page 36, lines 756-758). This was the most direct way that we could possibly respond to the reviewer's concern and we are confused by this reviewer's response. Perhaps it is linked to miscommunication of the differences between our model data comparison and a standard regression? We have attempted to clarify this above. We cannot simply "fit the slope" in the process-based model FANSY and every change we make requires thousands of years of integration. Even if we could, the limited climate scale time trend data would provide little in the way of new constraints. We do, however, view the fact that the simulated model trends are largely consistent with those observed as a positive, as it builds support for the model's capacity to simulate future changes.

Conclusion: I don't think that the new version of the manuscript has improved.

We feel that we responded directly and convincingly to the two primary requests that the reviewer made in the reviewer's second round of comments. We reconciled our results with those of Maranger et al. (2018) by showing that our results are fully within the bounds of current published uncertainties (and have now further clarified this in our response to the reviewer's above comment). We added the additional analysis of time trends in response to the reviewer's

specific request to do so. In this revision, we again have done our best to respond to the reviewer's new comments below.

A very simple test to look at the results/validation is to take the average over all observations and the model results (coupled to the observations). I made an overview of Figure SI6. The expectation of this exercise is that the average of both is almost equal. I divided the observation by the model result to get a fraction. When the fraction is below 1, then the model gives an underestimation. When the fraction is above one, the model gives an underestimation.

My conclusions from this table:
SS: Model gives an underestimation of more than a factor of 2. Not good.
NO3: Yield and Loads reasonable, but overestimation of concentration.
NH4: Good.
DON: Yield fine, Loads to low, concentration to low.
TKN: Overestimation of results.
PO4: Yield and Loads underestimation, concentration small difference.
DOP: Yield fine, Loads underestimation, concentration overestimation.
TP: underestimation of results.

So suspended solids needs some attention. The loads of all phosphorus forms are all underestimated. Which is a huge problem, because there is almost no loss in the rivers, so it is not
easy to increase the phosphorus loads! The concentration of all nitrogen forms are overestimated and for the loads nitrate is compensating DON.
Based on the analyses above, I really think the model should be improved! The model description is OK now, but the results and implementation is still too poor to meet the standards of GMD.

We disagree with these comments and will clearly and objectively rebut them in 3 steps:

1. We will demonstrate that the arithmetic mean which the reviewer suggests as a way to compare the model results and observations results in a misfit that largely reflects the model's fidelity to the largest few rivers, rather than its capacity to represent globally distributed rivers of different magnitudes and characteristics. The latter is our goal and the log-transform is a community standard tool for quantifying the extent to which this has been achieved.
2. We will clearly show that the widely-used IMAGE-GNM and Global NEWS models, the former of which is published in GMD, would have been rejected if the reviewer's criteria were applied. This, combined with the facts that our model has achieved a skill

comparable to these models using standard approaches while including far more comparisons objectively rebuts the reviewer's contention that the skill does not meet GMD and broader community standards.

3. We have prominently recognized (in the Abstract) the potential of global models to have substantial misfits for individual rivers. Hopefully this will raise awareness of this shared limitation of all global river solid and nutrient loading products. Addressing this, however, is a community-wide challenge requiring improvements in observations, model forcing and dynamics that will take years. We have enhanced discussion of pathways to model improvement.

Each of these steps is described in more detail below.

Step 1: The problem with using the cross-site arithmetic mean as a global skill metric

The reviewer's contention is that the model should reproduce the sum/average of loads across the rivers that happen to have been sampled. When river solid and nutrient concentrations and loads vary over many orders of magnitude, however, this choice emphasizes only one or a few large rivers in the sample. Our goal, as is the case with other global models, is to capture global contrasts between rivers of vastly different magnitudes and nutrient concentrations. The log-transformed data is the standard tool for this (e.g., Mayorga et al., 2010; Tian et al., 2023; Pelletier et al., 2012) because it weights fits equally across different magnitudes (a factor of 2 misfit is penalized the same whether the baseline value is 1 or 1000). The "Bad" classifications the reviewer has given to some quantities using the reviewer's arithmetic mean approach does not reflect a systematic bias across rivers, but is primarily because the largest river happened to be either over- or under-estimated.

To illustrate the implications of the reviewer's alternative evaluation method, we consider the fit to suspended solids (SS). As the reviewer points out, the ratio of the arithmetic mean of the observed to modeled SS loads is 2.35. The observed loads, however, vary over 4 orders of magnitude (e.g., 118 to 7,846,601 kt/yr for SS loads) across globally distributed rivers. Removing the single largest river (i.e., Huang He/Haiho) of the 65 sampled rivers, reduces the ratio from 2.35 to 1.03 (which becomes "Good" according to the reviewer's judgment criteria). ***This example shows how the arithmetic mean is more a measure of the model fit to the largest river, rather than an assessment of the model fit across globally distributed rivers varying in size and other climate/land use conditions. Like other global modeling studies, our primary goal is the latter.***

Step 2: Is the model fit sufficient for publication?

To further investigate the issue of whether the reviewer's ratio provides a valid measure of whether our model's skill is sufficient to merit publication, we repeated the suggested evaluation for the most widely known models Global NEWS and IMAGE-GNM (the latter is also published in GMD). As shown in the below Table R1, our model achieves a similar level of skill with the reviewer's metric as these two alternative models despite being less directly fit to the data. ***That is, the arithmetic mean ratio-based evaluation and criteria the reviewer suggests would have led to the rejection of these prominent and highly successful models. It thus clearly does not reflect GMD or broader community standards***.

	LM3-FANSY			Global NEWS			IMAGE-GNM
	Yields	Loads	Concentrations	Yields	Loads	Concentrations	Concentrations
SS	2.40	2.35	2.23	1.05	0.83	2.40	
NO_3^-	0.79	0.87	0.60				
NH_4^+	1.03	1.15	0.92				
DIN	0.76	0.96	0.68	0.98	0.76	1.13	
DON	1.01	1.66	0.51	0.68	0.57	0.77	
TKN	0.53	0.56	0.41				
PO_4^{3-}	1.54	1.82	1.22	1.31	1.21	1.30	
DOP	0.86	1.42	0.37	0.93	1.02	0.88	
TP	1.71	2.45	1.58	2.00	1.92	3.39	
TN							0.68

Table R1: Ratio of the arithmetic mean of the observed to modeled yields, loads, and concentrations. The ratio for Global NEWS was calculated by using the data reported in Fig. SI2 and the supplementary excel file. The ratio for IMAGE-GNM was calculated by using the data reported in Figure 10 and the supplementary excel file of Beusen et al. (2015).

The log transform that we and others (e.g., Table R2, Mayorga et al., 2010; Tian et al., 2023; Pelletier et al., 2012) have used for model validation addresses the issue demonstrated above by ensuring that the model fit is not disproportionately influenced by a few very large rivers. That is, following the log transform, a factor of 2 misfit for a base SS load of 118 kt/yr is given the same weight as a factor of 2 misfit around a value of 7,846,601 kt/yr. Without the log-transform, higher values are given much greater weight than lower ones. For this reason, no matter whether a model is dynamic/process-based (e.g., Tian et al., 2023) or statistical (e.g., Global NEWS), log transformed data analysis based on the Pearson correlation coefficients (r), Determination coefficient (R^2), and/or Nash–Sutcliffe model efficiency coefficients (NSE) has been a standard means of cross-watershed validation of global river models over the last several decades. Throughout the manuscript, we have shown that our overall model skill based on this analysis metric is at least as good as the skill of previous models (e.g., Fig. SI2 and refer to the explicit discussion of relative skill throughout our Results section). We feel this is particularly notable

given that our model is simulating coupled algae, solid, and nutrient cycles at the global scale for the first time.

Furthermore, thanks in part to the reviews, the extent of our evaluation now far surpasses that of any other global river models in the published literature. We have provided a Table R2 below that contrasts the extensive comparisons in our paper with those included in Global NEWS, IMAGE-GNM, DLEM-TAC, and the model of Pelletier et al. (2012). ***The FANSY model development description and validation thus objectively meet and exceed those of the past model documentation papers of GMD and others.***

	LM3-FANSY	DLEM-TAC	IMAGE-GNM	Global NEWS	Pelletier et al. (2012)
Instream biogeochemical, physical, and/or hydrological processes					
Process-based	√	√			
Dynamic	√	√	√		
Cross-watershed validation using one time point results (i.e., temporal mean condition validation)					
Log transformed	√	√		√	√
Statistics	r, NSE, prediction error	R^2	No statistics	R^2, NSE, prediction error	r
Time series validation using yearly and/or monthly results (i.e., temporal variability and/or trends validation)					
Statistics	r, prediction error, t-statistic of linear trend	R^2, NSE	RMSE	No comparison	No comparison
Analyzed unit					
Yield	√			√	√
Load	√	√		√	
Concentration	√		√		
Validated chemical species					
SS	√			√	√
NO_3^-	√				
NH_4^+	√				
DIN	√			√	
DON	√			√	
TKN	√				
PO_4^{3-}	√			√	
DOP	√			√	
TP	√		√		
TN	√		√		
DIC		√			
DOC		√			
POC		√			
TOC		√			

Table R2: Synthesis of the model validation and approaches used in model documentation papers for widely enlisted models. Note the prevalence of the log-transformed approach, and the much more extensive evaluation included in this study relative to the others. In the table, "process-based" refers to a model that resolves freshwater biogeochemical and physical processes in mechanistic ways. "Dynamic" refers to a model that provides time series results.
Pearson correlation coefficients (r), Determination coefficient (R^2), Nash–Sutcliffe model efficiency coefficients (NSE), and Root mean squared error (RMSE).

Step 3: Communicating current skill limitations and a pathway for further improvement

Finally, we agree with the reviewer that the misfits of global models like ours can be quite large for individual rivers and that this impacts the uncertainty in global total/mean loads. This is, however, a community-wide challenge common across global river models that must value

global robustness over regional optimality. This fundamental challenge still exists for even the simplest of variables (e.g., surface air temperatures in global climate models still have prominent regional biases). Regional fidelity must be addressed via concerted improvements in observations, models, and forcings (e.g., precipitation, fertilizer estimates) and thousands of years of simulations over the course of many years. Addressing these fundamental limitations is a "years to decades" problem. It is not something we can do within the timescale of a response to a reviewer. We have extensively noted the prominent uncertainties that will be prioritized in future work towards achieving this goal in the Discussion (pages 42-44, lines 829-900). In addition to this, we now call attention to the challenge of regional fidelity in the Abstract, explicitly noting that ***"While the simulations are able to capture significant cross-watershed contrasts at a global scale, disagreement for individual rivers can be substantial. This limitation is shared by other global river models and could be ameliorated through further refinements in nutrient sources, freshwater model dynamics, and observations."***

References

Tian, H., Yao, Y., Li, Y., Shi, H., Pan, S., Najjar, R. G., et al.: Increased terrestrial carbon export and CO2 evasion from global inland waters since the preindustrial era. Global Biogeochemical Cycles, 37, e2023GB007776. https://doi.org/10.1029/2023GB007776, 2023.

Di Toro, D. M.: Sediment Flux Modeling. Wiley-Interscience, New York, NY, USA, 2001.