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:

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.

2. The validation method. 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. 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).

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

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.

	Yield	Yield	Yield	Loads	Loads	Loads	Concentra	Concentra	Concentra
	Obs	LM3-FANSY	Obs/LM3-FANS	Obs	LM3-FANSY	Obs/LM3-FANSY	Obs	LM3-FANS	Obs/LM3-
SS	247346	102938	2.40	209177	88943	2.35	1613	723	2.23
NO3	129	163	0.79	81	94	0.87	0.56	0.93	0.60
NH4	43	42	1.03	38	33	1.15	0.18	0.20	0.92
DON	63	62	1.01	118	71	1.66	0.23	0.44	0.51
TKN	152	286	0.53	97	173	0.56	0.73	1.78	0.41
PO4	20	13	1.54	16	9	1.82	0.07	0.06	1.22
DOP	5	6	0.86	15	10	1.42	0.01	0.04	0.37
TP	133	78	1.71	461	188	2.45	0.45	0.28	1.58

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.