We have taken all of the review comments into consideration. We would like to thank the reviewer for these comments which we believe have substantially improved the m/s. The methodological parts are now better described and the formulas are now given in a more correct fashion. Discussion about uncertainties have been expanded. Below we reply in detail on how and if we taken the individual comments into consideration.

Reviewer nr 1

The article is interesting to read, but seems to be lacking some information that is necessary for better understanding by the reader. I suggest that the authors describe certain parts more clearly and consider setting more subtitles to keep different parts of model description and results apart. A comparison or discussion on pros and cons of the used model for retention in comparison to other possible models would be interesting.

ANSWERS: We have added more sub-section headings in the Results and Discussion section to increase the readability. It is now divided into
3.1 Parametrisation results
3.2 Major retention estimate results
3.3 Uncertainty aspects
We have expanded the discussion about pros and cons of the used model. A quantitative comparison with other models is outside the scope of this paper but we have included a qualitative discussion in the revised m/s (section 3.3). More precisely, the model used is an advanced regression model that goes beyond normal multiple regression analysis and can be seen as comparable with the SPARROW model (Schwarz, G.E., Smith, R.A., Alexander, R.B., and Gray, J.R., 2001).

Some questions to be clarified:
a) It is stated clearly which inputs are used, but the model description is confusing. Which parameters are estimated? Are all parameters areas specific, and if so do they vary a lot between areas? How is expert knowledge used in the fitting of the model. E.g. for equation 1 are there parameters estimated in all parts of this formuls (S, P, D and R) or are some of them observed or considered known. This information is given in the text later, should however be given right after formula 1 (e.g. page 10836 line 14 states what is assumed to be known, move this ahead).

ANSWER: The Model description has been substantially improved. All the formulas are now clearly given. Initially we described the general model given in Grimvall&Stålnacke (1996) but have in the revised version focused better on the adjustment made and parametric function used in this particular case study. We believe that this have increased the readability. In fact all the 4 given formulas have been changed. They now reads as:

$$L_i = \sum_{i=1}^{n} (1-R)S_i + (1-R)P_i + (1-R)D_i + \varepsilon_i \qquad (1)$$

where L_i is the load at outlet of basin i;

S_i is total losses from soil to water in basin *i*;

P_i is the point source discharges (WWTP and industry) to waters in basin *i*;

D_i is the atmospheric deposition on surface waters in sub-basin *i*;

R denote the retention for the source emissions *S, P* and *D*, respectively;

n is the number of basins, and

ε_i is the statistical error term.

The total diffuse loss of N from soil to water, S_i, in the i^{th} sub-basin was assumed to be a function of the land cover (Eq. (2)):

$$S_i = (\theta_1 a_{1i} + \theta_2 a_{2i} + \theta_3 a_{3i}) \quad (2a)$$

where a_{1i}, a_{2i} and a_{3i} in our study refer to the areas of three land cover classes, i.e. cultivated land, wetlands and other land (mainly forests), respectively. θ_1, θ_2 and θ_3 are unknown emission coefficients for the three land use categories that are statistically estimated in MESAW jointly with the retention (see Eq. (3) below). The point source emissions, P_i, and atmospheric deposition on surface waters, D_i, were assumed to be known (see Section 2.1).

Throughout the exploratory analysis we found that certain basins deviated from the relationship and in most cases also where geographically located near to each other. Thus we introduced a 'grouping variable' according to the following:

$$S_i = (\theta_1 a_{1i} + \theta_2 a_{2i} + \theta_3 a_{3i}) * \omega_j \quad (2b)$$

where each group j consisted of 2 or more basins depending on the model run (see Table 1) and where ω is the unknown coefficient(s). The model was run with different combinations of basin sub-groups in order to obtain reasonable model coefficients and load estimates (i.e. little deviation between predicted and observed loads). The grouping of basins was based on prior knowledge of similarities between basins as well as geographic location. For example, the 10 smaller Danish sub-basins formed one group, as a residual analysis showed that these sub-basins deviated from the general relationships. In its practical meaning, we simply adjusted the 'global' diffuse emission coefficients to the local conditions (despite we don't know the underlying causes). This can be justified since applying the same coefficient to such a large drainage basin (1 745 000 km^2) seems less logic.

…… Irrespective of the exact retention mechanism, the parameterisation of the retention in the different basins was after several exploratory runs with alternative models done with the following empirical function (Eqs. (3) and (4)):

$$R_i = 1 - \frac{1}{1 + \lambda_1 \sqrt{drainagearea_i}} * \frac{1}{1 + \lambda_2 \frac{lakearea_i}{drainagearea_i}} \quad i = 1,2,...,n \quad (3)$$

where λ_1 and λ_2 denotes a non-negative parameter and R_i denote the retention in the i^{th} basin. The empirical function were in our case derived from the conception that the removal of N takes place primarily in the surface waters (both instream and in lakes). The first part of the function reflects the instream retention whereas the second part reflects the retention in lakes and reservoirs.

Regarding the question if all the parameters are areas specific, and if so do they vary a lot between areas?
The final model include 9 estimated parameters (Model run #4 in table 1) and they don't vary between the drainage basins besides the case with the grouping-variables (see answer under comment f) below). The diffuse emission parameters give the area-specific loads (i.e., source emissions). For example, Model run 4 for cultivated land gives a point estimate of 1073 kg km^{-2}. Interestingly this is a value that normally could be monitored in small agricultural catchments in the Nordic/Baltic region (Stålnacke et al. 2014). We have included a better clarification of this in the revised m/s.

b) The total loss (S) is modelled from 3 land cover classes (cultivated, wetlands and other land). Do these 3 land cover classes add up to 100% of land cover? If so this should influence the estimation of the 3 parameters, since the variables will be linearly correlated. How is this handled? If there are land cover classes not in the model, this should be stated clearly.

ANSWER: Yes the 3 land cover classes adds up to 100% and are for sure inter-correlated. This will have less influence on the method applied although there is always a risk of multicollinearity of these kind of regression-type of models. It should be noted that the model inputs are areas of the land cover and not the percentages which will decrease the risk of multicollineariety. Experiences with the MESAW models as also given in the earlier quoted papers in different geographical areas (Liden et al; Vassiljev&Stålnacke, Vassilijev et al and Povilaitis et al) have not indicated any problem with possible interrelated explanatory variables..
In addition, parameter estimates displayed reasonable stability; little change occurred in the values of the most statistically significant model coefficients when additional variables were added in exploratory regressions.

c) Two formulas are given to compute/estimate retention. I the difference between them that one is used if there are lakes in the area, whereas the other one is used if there are no lakes? Or how do you choose between these for the different basins? Is lambda the same in these two models, i.e. if lambda a common estimate for both equations? State in the article. Hesse et al. ECOLOGICAL MODELLING Volume: 269 Pages: 70-85 made comparisons for different retention models. This might be interesting for you to comment in the article.

ANSWER: Both formulas for retention (Eq 3 and 4) is used in the simultaneous estimation of the source emission coffecients and retention coefficients. There are in fact 2 lamdas that is estimated. Formula 3 and 4 have been corrected accordingly

Given the confusion we have modified formulas 3 and 4 and replaced it with:

$$R_i = 1 - \frac{1}{1+\lambda_1\sqrt{drainagearea_i}} * \frac{1}{1+\lambda_2\frac{lakearea_i}{drainagearea_i}} \quad i = 1,2,...,n \quad (3)$$

A sentence that better explains this is included. The reference to Hesse et al have been included. Thanks for that reference.

d) The risk of overfitting/overparametrisation is mentioned and given as reason that retention parameters are the same for all source categories. Is this reasonable and can be motivated? How? How do you control for overfitting in this model, is it by only allowing a few parameters to vary or do you control it? Would any kind of cross-validation help to avoid overfitting?

ANSWER: We have the removed the sentence on ovefitting/overparametrisation. In total, 9 parameters were fitted on the 88 observations. Parameter estimates displayed reasonable stability; little change occurred in the values of the most statistically significant model coefficients when additional variables were added in exploratory regressions. Moreover, the diffuse source coeffcients (thetas) where all realistic in its value which is further explained in the revised m/s. We thus regard the issue with overfitting/overparametrisation as less likely.

e) In page 10837 line 9 you talk about the total N retention that is estimated. Does this regard fitting R*Si+R*Pi+ R*Di, related to equation 1? When you do fitting on different groups, are parameter

estimated individually for a group? If 10 danish subbasins form one group, how many parameters do you estimated from those, is it 4 (3 theta and 1 lambda) or more? Are estimates for thetas and lambda very different for the groups of basins? Parameter estimates should be given, at least as example.

ANSWER: We have now better explained how the total retention is estimated and how this is related to Eq1. The question on the grouping parameter/variable is explained under answer f) below. The parameter estimates is given in Table 1 and we have in addition included the thetas and lamda into the table heading for clarification and better references to the formulas given in Material and Methods

f) If groupings of basins is made due to geographical location or similarities, would not that suggest dependence/correlation between the basins and influence p-values (with the concept of statistical inference based on independent observations). The error term in (1) does not indicate that dependencies are taken into account. Can p-values be trusted?

ANSWER: This is a misunderstanding. The basins are not merged. Instead we during the modelling found that some basins deviated from the general relationship and most of these basins were in fact located geographically in the same geographical region. To the end, we identified 3 such 'groups' of basins (lower part of table 1). This will not by any means affect the independency criteria in this kind of statistical modelling. Instead we were with this 'grouping' able to differentiate eg the diffuse emission coefficients. For example, it is known that basins in Denmark and southern Sweden (due to more intensive agriculture) differ from the ones on northern Finland and Sweden. The procedure applied can be seen as introducing a dummy variable in normal multiple regression.

g) In the results unit-area specific loads are discussed. As the model is designed to predict N load rather than unit-area loads: was this expected? Could the model be adjusted if unit-area loads are interesting? Could this be a result of overfitting in the original model?

ANSWER: The model was fitted to river loads given in kg. We wanted to show-case the model results also as unit-area loads since this give higher credibility to the results and analysis. Principally, the model is generic and can also be applied with any dependent variable.

h) In figure 4 the relationship between estimated retention and total drainage area are given. In these figures it seems that drainage area has no influence on retention in %, whereas lake area (%) has a clear nonlinear relationship. How do these curves related to equations 3 and 4? Probably the equations and estimated parameter lambda are used to compute the estimated retention, i.e. the curves should reflect the relation in 3 and 4. Is this true? The line shown in the plot 'retention and lake area', why is it plotted there? How is it related to the model? Since this line does not fit well, does this indicate that the model does not fit well?

ANSWER: We agree that figure 4 can be confusing for the reader. The intention was to illustrate how the estimated retention (in %) is pair-wise correlated to the 2 main variables (lake are and drainage area) included in the retention expression. Apparently there is a strong curvelinear relationship between retention and the lake-share in a drainage basin and that there is a much weaker relationship between the retention and size of drainage basin. A further discussion about the interpretation of this is given in the revised m/s. We have also removed the fitted line in Figure 4 (left panel) since it is not connected to the parameter estimation at all.

i) Also the function fitted to specific load and lake area (%) is strange, why do you use this fitted line instead of an exponential/logarithmic relationsship or a square-root relationship. Where does the function come from? How is it motivated?

ANSWER:The figure 5 on area-specific N-loads vs lake area (%) is just given as an illustration on the relationships in the input data and just a support to the retention formula applied. It is given to the reader as an example. We have removed the fitted lines and the regression equations from the figure to avoid confusion.

Smaller notes
Relative differences are used to give equal weights to small and large basins. A motivation why this is a good choice in this context would be appreciated.

ANSWER: The model as given in formula 1 is based on loads at river mouths. In order to avoid that large basins (large basins will for most cases have more loads than small catchments) will have more effect in the parameter estimates we used the relative differences between observed and fitted loads. This is a standard procedure in many statistical analysis of this kind.