**Biogeosciences
Discussions**

Interactive
Comment

Interactive comment on "The use of machine learning algorithms to design a generalized simplified denitrification model" *by* F. Oehler et al.

F. Oehler et al.

f.oehler@niwa.co.nz

Final response

Thank you for your comments. We hope our answers will clarify some points.

detailed Response

(Referee comment is in italic)

The presented manuscript "The use of machine learning algorithms to design a generalized simplified denitrification model" describes the use of two machine learning algorithms to design a simplified denitrification model, which is a new and interesting approach in this field. However, to use the ANN model not only as a black box but as a

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

generalized simplified denitrification model and interpret its response shapes, requires further work in proving the generality of the obtained model. Therefore, I would like to ask the authors to address the following three comments:

1)Inconsistent choice of methods: Artificial neural networks (ANNs) and boosted regression trees (BRT) were used to determine the configuration of the final ANN model. These two machine learning algorithms present very different mathematical approaches. An influencing variable in a BRT is one that minimizes the variance of the output and fulfils a splitting task. This variable might not have a similar importance to present a (non-linear) response in an ANN with a minimized root mean square error. There are several methods to determine the input relevance of variables directly for ANNs (Gevrey, M., et al. (2003). "Review and comparison of methods to study the contribution of variables in artificial neural network models." Ecological Modelling 160(3): 249-264.). Please discuss this inconsistency.

Reading a 5 variable model with interaction is certainly a difficult task. As highlighted in our text (ref (Olden et al., 2004) in 3.1) and by the referee, there are methods to understand the variable relationship captured by the ANN, but opening the model, i.e. to analyse interactions from within, is difficult and in that regard, yes, ANNs may be defined as "black box" models. This is why we used the BRT approach, and not the ANN, to analyse the relationship between variables, as stated in the Abstract, the introduction, in 2.3, 3.1, and in the Discussion. The referee is right when highlighting that the measure of the influence of an input variable will be different in a BRT and an ANN, and also that this measure of the influence will change when the metric used to evaluate the residuals (and guide the training) changes (e.g. RMSE, Laplacian, Variance...). But here we do not oppose nor compare ANN and BRT results regarding variable influences. Hence there is no inconsistency. Moreover, the choice of the input variables was not solely done using BRT, but also considering prior knowledge on the denitrification processus.

We added the reference provided by the referee in 3.1: "Potentially, different techniques could be used to "open" the ANN and try to understand the variable relationships (Gevrey, M., et al., 2003)."

2)Missing evidence of the generality of univariate response functions: The univariate response shapes presented in Figure 8 are only ONE very special case of the hyperspace ANN(T,WFPS,NO3,OM,pH), with the values of the other variables fixed. For example, if the value of WFPS would have been fixed not to 100% but to 60%, the response of OM in Figure 8 D would have shown a very flat response instead of a steep increase according to Figure 9 D. Therefore these results should be discussed as a very specific cases of the fixed values, or evidence should be provided that the found univariate responses bear any generality beyond the presented fixed values.

As a general comment, we could argue that if the univariate (or bivariate) response functions were bearing "generality", then the whole exercise would have been futile, as the non-linear variable interactions would have been negligible, and the NEMIS type modelling approach would have been already â€{œgenericâ€Ì. Yes the presented response shapes (Figure 8 and 9) are special cases, chosen to show some "classic" environmental conditions where denitrification occurs, as stated in 3.5: "For figure 8 the fixed values were chosen to represent the conditions of a classic cultivated silty loam soil with medium OM%. Temperature and WFPS are set to be non-limiting. For figure 9 the fixed values were chosen to represent a silty clay riparian soil with plausible occurrence of a Temperature of 16 C and full water saturation (WFPS=100%). Beside Temperature and WFPS, the fixed values were around the median of the dataset. In particular, the figure 9 presents the Da rate response to variations of different factors, at a fixed temperature of 16 C. At 10 C and down to 5 C, response shapes, and trends are globally similar." In these conditions, and some others (as stated in 3.5, 3.5.1, 3.5.3 and in the discussion section), the shapes and trends are similar to NEMIS representation, which was already capturing the main effects. We provided the model formula and a

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

scatter plot of all variable combinations in the appendix so that anyone can test the model in different conditions, inside (or outside) the data space.

We already proposed a metric for the generality, and put warnings, as stated in 3.5: "Intuitively the level of trust of the model is related to the data density in each part of the data space. Denser data points are needed to represent fast gradient change area. Without external knowledge, we cannot know if the gradients are well represented: indeed, they are built with the data (training). The assessment of the performance (or trust) is made at a global level (the NRMSE). **The prediction performance of the ANN outside the training (calibration) dataset space is not assessed and it can display physically unrealistic behaviour.** Dataset boundaries are shown in Fig. 8 and Fig. 9. As guidelines to evaluate data point distribution, scatter plots of the combinations of Da, NO3-, WFPS, OM, pH and T are available in the Appendix." In extenso, the model should not be used outside the training dataset space because it was not validated outside (so not for extrapolation). This is true for any model, and models are often misused because their domain of validity is overlooked. The degree of generality, or "trust" is evaluated using an independent validation (here a NRMSE).

3)Missing evidence of the generality of the coupled response functions: The same argument is true for the different couples of response in Figure 9. Would these surfaces still bear the same features, if for example the pH would have been fixed at 5.5 or the WFPS at 40%? Which features would hold, which features change? Again, please provide evidence that the response surfaces presented in Figure 9 have relevance beyond the fixed values of the input variables before discussing their implication. Missing evidence of the generality of the simplified model:

Please see our previous comment

The more input variables are used, the more freedom is in the ANN mapping and the

more likely will the ANN model have ecologically implausible mapping.

We agree. The upper limit of the number of input variables is determined by the number of records. When increasing the number of input variables, overfitting tends to occur. The early stopping technique was used to control this effect in ANN, and cross-validation was used in BRT. The "implausible" is controlled by an independent validation, the metric being a NRMSE (Please see our previous comment and what is written in section 3.5). Consequently, the model should not be used outside the training dataset space because it was not validated outside.

Does an ANN with only the three major variables as inputs ANN(T,WFPS,NO3) exhibit the same univariate and coupled responses?

Globally, it exhibits the same trends, like what was captured by NEMIS, but it is less plausible/generic in details (NRMSE of 1.63).

Does an ANN with a different set of five input variables (e.g. ANN(T,WFPS,NO3,DEA,pH)) still shows the same features of the Da rates as in Figure 9? How general and plausible is this simplified model?

Overall, replacing OM% with DEA was leading to similar response shapes and trends. The generality and plausibility has been evaluated by an independent validation, and its validity is limited by the validation data space, as any model.

Furthermore, the writing style of the manuscript could be improved for clarity and consistency. For example, rather than jumping back and forth between the three methods, the calibration section could follow the description of each technique, and the headers

in the results section could be more meaningful, e.g. almost the whole chapter describes "ANN results", not only 3.2. The abstract is difficult to read, again jumping back and forth between ML, BRT, ANN, and Nemis. The first two sentences of the abstract have a repetitive start of "we designed", rather than the overall focus and goal of this paper. Please quote the list of abbreviations.

For more clarity we have reworded the titles of our sections as following:

3.1 ANN results => ANN training results 3.2 BRT results => BRT training results 3.4 The ANN5(OM,pH)G model =>The ANN5(OM,pH)G model performance 3.5 ANN model response shapes =>The ANN5(OM,pH)G model response shapes

We have also reworked the abstract.

Overall, the presented paper has the potential to provide a new alternative to design a simplified denitrification model. However, the generality of the presented simplified model remains to be proven and I am looking forward to the revised manuscript.

Yes, the first intent of the paper was to present a way of **designing** a generalized simplified denitrification model. As for any model, the evaluation of the "generality" of the model, or "performance", or "level of trust", is only valid in the validation data space. Interestingly, by using ML approaches, the generality of the model will expand when adding more and more data spanning more gradients. The results also provided some insights about the influence of pH, and we could expect more refinement to our understanding with more data. The generality of the model was evaluated by an independent validation (i.e. data which was never used in any stage of the calibration), and the metric used was a NRMSE. Using this metric, the ANN and BRT approaches were more generic than the NEMIS type approach.

Interactive comment on Biogeosciences Discuss., 7, 2313, 2010.