

## ***Interactive comment on “Large-scale regionalization of water table depth in peatlands optimized for greenhouse gas emission upscaling” by M. Bechtold et al.***

**M. Bechtold et al.**

michel.bechtold@ti.bund.de

Received and published: 1 July 2014

We very much appreciate the review of anonymous referee 2. Below, we provide replies (standard) to each of the comments (italic). At several points, we thoroughly revised the manuscript, in particular the paragraph about the weighting scheme. We feel that this considerably improved this part of the manuscript. A revised version of the manuscript with tracked changes has been uploaded as supplement to the reply to referee 1.

*General Comments This manuscript contains some interesting new approaches to*

C2145

*generate exhaustive spatial predictions in a data-scarce environment. The structure of the manuscript could be improved by starting with the methodology followed by the data and data requirements. Specific comments The structure of the manuscript would benefit from starting chapter 2 with paragraph 2.3 first describing the general methodology followed by the data used. The current structure requires referring to paragraphs later on in the manuscript.*

We also see that at some points we refer to the methodology during presentation of our data on target and predictor variables. To resolve this problem, we considered shifting the general methodology to the beginning of chapter 2 as suggested by the reviewer. However, several aspects of the model building and weighting scheme are the consequence of the detailed characteristics of the target and predictor variables and thus require the data background for discussion. Regarding the more complex discussion during the presentation of the model building, we preferred to present first the characteristics of the data. Given the comment of the reviewer, we did however shift the more detailed crosslink about data gaps from the data presentation section to the model building section. Further, we now provide first the data weighting scheme before we present the performance criteria in which the weights are used. We hope that this helped to minimize the crosslinks to an acceptable level. (see lines 338 to 345)

*On page 7 line 19 you state that boosted regression trees allow for data gaps, but I can't find an explanation on how predictions are made when predictors are missing.*

This is a good question. The detailed handling of missing values (i.e. NA values) is not well documented in the gbm and dismo package documentations. Given this lack of documentation, there are several online discussion about this question. It is clear that the gbm package handles missing values in predictor variables by introducing surrogate splits. When applying the function `pretty.gbm.tree` to the gbm object, it can be seen that each split consists of 'LeftNode', 'RightNode' and a 'MissingNote'. By own tests, we found out that the mean target value belonging to the predictor NA values

C2146

is attributed to these surrogate splits during model building. In case there is no NA value in the predictor variables, the mean of all target values is stored in the Missing-Note, which is then applied during prediction when NA values in the predictor variables appear. (see lines 338 to 345)

*The weighting scheme discussed in paragraph 2.3.3. is presented as common practice yet there are no references in the entire paragraph. The mathematical formulation for  $W_i$  gives three different formulas (5,6,7) all using the same symbols, this indicates that there are in fact three different  $W_i$ -types, please make this clear from the mathematical notation or give only the formula used. Formulas 8 and 9 require a reference or some mathematical backup.*

We thank the reviewer for pointing out inaccuracies in the way we presented the weighting scheme. Our weighting scheme is not common practice. We are not aware of a 'common practice' for heterogeneous target and predictor data situations similar to our study. Our feeling is that despite the obvious importance of data weighting, there are no clear guidelines how to deal with it in complex modeling problems and that there is always a degree of subjectivity involved (see also Francis et al. 2011). This is emphasized in the new manuscript. We revised the weighting paragraph thoroughly. We clarify on which statistical principles we rely our weighting scheme (missing references and mathematical background is now provided) and how we combine them to a new weighting scheme for our modeling problem. The inaccuracies of the mathematical formulations (former equations 5,6,7) have been corrected. (see lines 433 to 562)

*Ad a formula number for the BIAS.*

Thank you. Done.

*Notation in formulas (3, 4, 4b) is not concise:  $i$  is used both as subscript and superscript, it should be subscript.*

Correct. Our mistake.

C2147

*Simulated or estimated values are commonly indicated with the hat-symbol, individual observed values can be denoted with the subscript  $i$ .*

Here we prefer to keep our notation which is also frequently used, e.g. Gupta et al. 2009

*In paragraph 2.3.2 the weights  $W_i$  are used, but the paragraph on weighting comes later on, start with paragraph 2.3.3.*

Done.

*The reference (Gaast., 2009) is missing in the literature section*

It is "van der Gaast et al. 2009", and it is already in the references.

Kind regards, Michel Bechtold and co-authors

References: Gupta, H. V., Kling, H., Yilmaz, K. K., Martinez, G. F, Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J Hydrol, 377, 80-91, 2009. Francis, R. I. C. C.: Data weighting in statistical fisheries stock assessment models, Can J Fish Aquat Sci, 68, 6, 1124-1138, doi:10.1139/F2011-025, 2011.

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 11, 3857, 2014.

C2148