Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-152-SC2, 2017

Geoscientific Model Development Discussions

Interactive comment on "A map of global peatland distribution created using machine learning for use in terrestrial ecosystem and earth system models" *by* Yuanqiao Wu et al.

M Bechtold

michel.bechtold@kuleuven.be

Received and published: 1 September 2017

This is a very needed and relevant work. Good globally harmonized peatland maps do not exist, and machine learning as a tool to derive these maps with better resolution and quantiative information is a track that should be gone (next to other efforts). At best, these different efforts give us soon good global peatland maps that even include information about the peatland type.

I want to limit my non-referee comment to two major points that need to be considered before this work should be accepted.

1) Splitting the dataset randomly (!) into training and validation must not be done when data points are correlated. In this application, there is a strong spatial correlation between the data points, i.e. pixels. The random selection of points for training means that validation is done with a highly correlated set of data points. This has two consequences: a) performance of the trained model is highly overestimated when looking at the "validation" metrics (R, RMSE, etc.) b) the training person tends to overtune/overfit the machine learning model (encouraged by the good validation) and includes more and more degress of freedom. Normally, a validation (that should be done with independent data) indicates the model developer where to stop. Recommendation: Divide your data set (left side of Fig.1) into e.g. 10 regional pieces for which you can say that they are spatially not (or only very limitted) correlated and then either use 40 % of these regions as pure validation data or perform a cross validation always leaving out only one region for which you make a prediction with a model trained on the other 9 regions. Recent publications that point out the importance of the independance of the validation data in machine learning applications: Jorda, H.; Bechtold, M.; Jarvis, N. and Koestel, J. (2015): Using boosted regression trees to explore key factors controlling saturated and near-saturated hydraulic conductivity, European Journal of Soil Science, 66(4), 744-756. Bechtold, M.; Tiemeyer, B.; Laggner, A.; Leppelt, T.; Frahm, E. and Belting, S. (2014): Large-scale regionalization of water table depth in peatlands optimized for greenhouse gas emission upscaling, Hydrology and Earth System Sciences, 18, 3319-3339.

2) Subsection 3.2: Comparison against using the HWSD soil database for peatland distribution This is not a proof that the trained model is superior than HWSD. You train a model on Tarnocai and Peregon maps while using among others HWSD as input. Then of course the trained model performs better in predicting Tarnocai and Peregon maps than HWSD. The comparison is not fair and misleading.

I hope this comment helps to improve manuscript and reliability of the map, and I hope it is taken in the spirit intended: scientific openness and fair reviewing. Supporting a

policy of open reviews and comments, I wish my name to be revealed to the authors.

Michel Bechtold KU Leuven, Belgium, 1 Sep 2017

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-152, 2017.