Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-152-RC1, 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.

Anonymous Referee #1

Received and published: 9 September 2017

The authors use available soil data, climate data, and topographic data, combined with a machine learning method to produce a spatially continuous global peatland fraction map. Such a map is relevant and useful for modelling communities. They start by including all available variables and the algorithm selects the variables that have the strongest predictive power for mapping peatlands. In the end, the only variables with a lot of predictive power are soil carbon quantities. This is of course expected, given that a peatland generally has very high soil carbon - more than any other soil type - making it a good proxy.

An advantage of this method is that it does not make any assumptions about the statis-

tical distribution of the data or any of their relationships. It is useful to show that using additional variables (e.g. meteorological data) cannot give a much better distribution of peatlands than just using soil carbon alone. The method could also be very useful for other, similar applications.

The paper is very clearly written and presented, and a potentially valuable contribution, but I would recommend some substantial changes to ensure that its value is fully realised.

Comments

I wouldn't take the result that carbon is the only useful proxy as definitive. The peatland maps themselves (i.e. used for training and validation) also have relatively large uncertainties, and so the fine details of peatland distribution beyond broad classifications are probably not resolved (thus any impacts of hydrology, topography etc, may simply not be well resolved in the training dataset). However, I would also suggest trying the fraction of grid cell with topographic index higher than a certain value as an input variable, rather than basing it only on slope, because this additionally takes into account the amount of water draining into the land (essentially whether it is an upland or a lowland), which has a major impact on hydrology and the potential for peatland formation.

As in the short comment that was posted already, I do not find the evaluation of the new map very convincing, and given the uncertainties in all of the soil datasets, the extremely high accuracy that the authors claim does suggest overfitting. With using a relatively new approach in this field, I think it would be wise to go into a more thorough evaluation, perhaps discussing in more detail the form of the relationships that are produced by the algorithm, and considering whether it would be possible, for example, that using both subsoil and topsoil organic carbon content can give more information than just the fraction of histosols in the grid cell. Currently it does appear that you are using the same dataset that was input to the model, as evaluation (in Section 3.2). In

Section 3.2 there is some discussion about threshold behaviour, but could this also be a problem in the HWSD inputs to the model? And if so, if this do not show up in the modelled results, that might suggest that they have been smoothed away by overfitting to other variables. I would recommend taking the suggestions from the 'short comment' in terms of dividing the data into training and validation points, and also expanding the evaluation/discussion to make it clearer why your approach is really an improvement on simply using the histosol map.

Whether or not the approach represents a major improvement on the HWSD histosol map, it could also be applicable to other problems. The statistical modelling method is not described in much detail in the paper, however. Since this might be a direction that other modelling groups want to take, I suggest that the authors expand the description of the method, highlighting its scientific importance for Earth System Science and why it is in theory better than alternative methods. So, at the end of the introduction I would recommend adding another paragraph highlighting the basic principles of machine learning and its potential as a method for Earth System Science. Then in the methods, go into more detail in Section 2.2 (for example, what is a sparse coefficient?). References are provided, but in my opinion we should be as clear as possible in papers, particularly when these approaches are not yet well known in Earth system science (and for this journal in particular).

Comparing to Alexandrov et al. (2016), the authors say that it's surprising that climate variables don't play more of a role in their model: this is not a valid comparison. Alexandrov et al. were using climate variables alone to determine suitable conditions for peatlands, whereas the model in the present manuscript has soil carbon content as an input, which is basically a proxy for observed peatlands. So, using something close to 'observed peatlands' will of course give a closer match to observed peatlands than climate variables alone. (However, in the last glacial maximum as considered by Alexandrov et al., there is no direct observation of peatlands or soil carbon, hence the need for a predictive model based on climate and topography.)

Technical comment

On Figure 5 it's a bit difficult to see which part of the world is being shown. Could you add some latitude/longitude markers?

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