Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2021-426-RC2, 2022

Comment on gmd-2021-426

Anonymous Referee #2

Referee comment on "A map of global peatland extent created using machine learning (Peat-ML)" by Joe R. Melton et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-426-RC2, 2022

In this manuscript, Melton et al. employed a machine learning approach, i.e., Gradient Boosting Decision Tree (LightGBM) to predict the coverage of peatland over the whole globe by using climate, soil, terrain and vegetation data. The training data for LightBGM were from high-resolution regional dataset of peatland distribution and coverage. It is found that their approach can map peatland coverage reasonably well compared with existing global products using traditional approaches. As stated by the authors, the results of the study are valuable for prescribing peatland cover fraction in the Earth System models to better explore the impact of climate changes on global or regional carbon and water budgets.

In general, the manuscript is well-written. The use of approach and data are appropriate. The results seem to be robust. There are, however, several weaknesses of the analyses, which require further improvement before being accepted by GMD.

Major comments:

1. Issues with the machine learning approach:

1.1. Lack of a baseline model to illustrate the added value of using LightGBM.

The authors have had quite detailed explanations of their machine learning workflow, but the performance of a baseline model is missing, which makes it difficult to judge if the choice of LightGBM has helped to improve the prediction or not.

1.2. The training data shown in Fig. 2 seem not to be well-balanced for producing meaningful machine learning outcomes. To me, there seem to be too many grids with either high peatland cover fraction (>60%) or zero peatland cover, while the grids with medium peat cover fraction (e.g., 1-25%) are rare. I strongly suggest the authors plot a figure showing the probability of peatland cover fraction in the whole training dataset to see if the training dataset is balanced or not. If not, the authors need to address the problem with an appropriate method. For instance, the use of non-peatland grids needs to be constrained to maintain a balanced training dataset.

1.3. The training data used in the study (Table 2) are from existing regional peatland mapping data generated from various approaches. The training data itself has various biases and is not ground truth. I am wondering why the authors did not choose to use the available ground-truthed site-level observations data for their machine learning model? Is such data not sufficient enough to make global predictions?

2. The data used to train and evaluate the machine learning model are not comprehensive and state-of-the-art. The authors have made a good effort in bringing available peatland cover data all over the world to train the model. But I was a bit disappointed that they have not used any recent dataset from China (e.g., Mao et al. 2020) or the US, or at least explain why they have not done so. Given the lack of training and evaluation dataset as also pointed out by the authors in the manuscript, I suggest they take advantage of the available dataset as much as possible.

3. The choice of the spatial resolution of the input dataset (5x5 arc minute grid) for machine learning is a bit arbitrary to me. What is the rationale behind the choice? Why not use coarser spatial resolution or higher spatial resolution? Will the choice of the spatial resolution affect the machine learning results?

4. The authors have done a systematic job in selecting the best predictors for the machine learning model (Figure 3 and Table 1). However, the selected predictors are a mixture of climate, soils, vegetation and terrain variables. Some are relatively stable through time (e.g., terrain and soil), some are not (e.g., climate and vegetation). I understand that such a combination gives the best global prediction. But I am also curious how the machine learning model performs with only terrain and soil variables, or with only climate and vegetation variables. This may provide some useful insights on the relative importance of long-lived vs short-lived conditions in forming peatland.

Other comments:

Line 5-6: please be more specific about which “machine learning technique” used in this study.

Line 79-81: have the authors tested if using different definitions of peatland instead of "30% dead organic material by dry mass" and "minimum thickness of 30 cm of peat", will affect the results?

Line 86-87: Why use 5x5 arc minute grid? Please explain.

Line 97-98: please explain here why use "calculated length of the longest day of the year (hours) for each cell" as a predictor?

Line 103: please specify the years covered by JRA-55.

Line 115: it is ambiguous to say "we use 30 cm depth estimate for all soil variables". Do you use 0-30 cm averaged value or the value only at 30 cm depth?

Line 133-134: please be clearer if you use the multi-year average of "2000-2015 period" or the data from each year?

Line 143: what are the years of S-NPP dataset used in this study?

Line 191-192, 201-202: The transformation of peatland map into cover fraction seems to be arbitrary. Please explain why?

Line 235-244: The creation of zero peatland cover grids is also a bit arbitrary to me. Generating too many zero grids in the training dataset, might give rise to an imbalanced dataset that is detrimental to the performance of the machine learning model.

Line 284: Please specify how Rj is derived.

Line 337-343 and Figure 3: why the importance (%) of wind speed and runoff from the full model simulation do not match with that from the BLOO CV, which is different from the other variable? The explanation provided in the text is not clear to me.

Line 353: Is the lower peatland cover over the northern hemisphere than the other available estimate related to the imbalance of the training dataset (too many zero peatland cover grids in the training dataset)?

Table 1 and 2: please add original spatial resolution of each dataset in the table. Also, it would be great to include the time coverage of each dataset, and how they were aggregated timely.

Figure 2: It would be nice to add a plot showing the probability of different peat cover fraction in the training dataset.

Figure 6: We need more quantitative comparison among different dataset. In addition, I am not quite sure how "Peat-ML from the BLOO CV" is derived in this figure. Did you average the results from all the 14 BLOO CV, or do you use only the results from which the blocks covering Canada were left out. In the figure caption, "three other peatland extent products (d-f)" should be "four other peatland extent products (d-g)".

References:

Mao, D.H., Wang, Z.M.*, Du, B.J., Li, L., Tian, Y.L., Zeng, Y., Song, K.S., Jiang, M., Wang, Y.Q. 2020. National wetland mapping in China: A new product resulting from object based and hierarchical classification of Landsat 8 OLI images. ISPRS Journal of Photogrammetry and Remote Sensing, 164: 11-25