Biogeosciences Discussions

Biogeosciences Discuss., referee comment RC4
https://doi.org/10.5194/bg-2021-189-RC4, 2021

Comment on bg-2021-189

Anonymous Referee #4

Referee comment on "Improved prediction of dimethyl sulfide (DMS) distributions in the northeast subarctic Pacific using machine-learning algorithms" by Brandon J. McNabb and Philippe D. Tortell, Biogeosciences Discuss., https://doi.org/10.5194/bg-2021-189-RC4, 2021

McNabb and Tortell present an improved prediction of DMS distributions in the NE Subactic Pacific which they achieve using machine learning algorithms. Understanding air-sea exchange of DMS is important for understanding the marine source of sulfate aerosols to the atmosphere, which act as cloud condensation nuclei. This work demonstrates that both of the machine learning techniques applied to this dataset (RFR and ANN) provide superior fits to the observations than were obtained using previously developed regressions. The paper is well written and the methodology and conclusions are sound. I think it is deserving of publication in Biogeosciences. However, I have some suggestions to help strengthen the manuscript.

I appreciated the analysis of impact of gridding resolution on the results. However, I wonder about the impact of binning the DMS data monthly regardless of year. Looking at the data in Figure 4, there is significant patchiness which I can only imagine is temporally and spatially variable. Given the power of the machine learning algorithms, why not use the full complexity of the dataset and pair the DMS observations with the closest (spatially and temporally) measurement of the predictor data sources?

Two machine learning algorithms were used in this study but there wasn't a robust analysis of which one was better and why. Should future studies use one over the other? Does one need to try multiple methods? Such a discussion would be a valuable addition.

Minor comments:

- The methods are very sparse. More information on the machine learning algorithms should be included (e.g. was this done with a package? If so which one?) This is in the

'code availability' statement to some extent but should be included in the methods along with a brief description of the algorithms and differences between the two.

- Only 20% of the data was held back for testing. It seems that it would be better to have a 50/50 split to provide a sufficiently large dataset for testing to confirm the robustness of the results.

- Are there any issue with correlations between the predictor variables? For example, many are derived from MODIS and so should have inherent correlations (ie not independent measurements).

- Figure 1: It seems a bit surprising that the R2 value decreases so dramatically with resolution but the DMS flux barely changes. Is this just due to the large spatial variability in the flux?

Line 37 missing an 'a' —> by a suite of environmental …

Line 152: typo? Should it be modified from?

Eq 4: are the coefficients provided anywhere?

Figure 1: caption refers to green lines/symbols instead of black

Line 261: it would be helpful to provide the fractional area represented by the study region. For example, if it accounts for only <1% but accounts for 4-8% that is more impactful.

Line 461: it should be "approach for modeling"