Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2018-353-RC2, 2018

Biogeosciences
Discussions
Open Access

Interactive comment on "Quantitative mapping and predictive modelling of Mn-nodules' distribution from hydroacoustic and optical AUV data linked by Random Forests machine learning" *by* Iason-Zois Gazis et al.

Anonymous Referee #2

Received and published: 24 August 2018

This paper was very interesting to read and clearly demonstrates how combining several state of the art scientific tools can achieve results that were, until recently, difficult to produce. The main idea in the manuscript; using Machine Learning to derive abundance of nodules from predictor variables remotely sensed with an AUV, has been applied by these authors and others datasets and this paper combines data and methodologies that have both been featured in other publications (cited in the paper). However, it presents a thorough protocol to make use of these tools, combine and optimize them, account for known caveats in the procedure and demonstrates the applicability of this

protocol in a practical situation. The scientific approach is complex but transparently detailed throughout the method, results and appendixes. Thus, this paper is a useful case study and a method that should be applicable to other similar datasets and, as such, is a valuable contribution to the exploration of the Manganese nodules fields in the CCZ. It is well written but could be streamlined and made easier to read. The important findings could be further highlighted in the results section by moving some of the subsections in the appendix (as highlighted by reviewer 1). In addition, I found that several sentences or groups of sentences in the discussion either were confused in their formulation or didn't make a clear point. Furthermore, the discussion could be structured into several paragraphs to help readers perceive the different points made by the authors.

I also have a couple of specific remarks and suggestion to add to those of reviewer 1:

R400: If RF is not good at predicting outside the ranges of the training set, could it affect the projected map of nodule abundance? Other studies projecting RF models (of species distribution) in space (or time) have used multivariate environmental similarity surfaces (MESS) maps (Elith et al. 2010). This procedure is mapping how dissimilar to known data points the predictors are across the projection area. This could potentially highlight that predictions in deeper and shallower areas than where nodule abundance samples are should be considered with care. This could also help target areas for future sampling. See Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. Methods in Ecology and Evolution 1:330-342

R415: The relevance of the depth as the most important predictor could be discussed further. Is there a geological reason why depth is the main driver of nodules distribution (as it looks unintuitive as to why such small changes in depth could drive nodule distribution)? Is it likely to be a proxy for another driver?

R499: Minor point but Judging by figure 12, the relation between MSR and the different tuning parameters, particularly the number of training samples is not linear and thus, could either increase asymptotically towards a maximum or might continue increase logarithmically. Either way, It is unclear if more data would be a major improvement. Thus, collection of new data should focus on better distributed data

R510: Given the rarity of corers data compared to photo data, would it not be better to take all cores where there is photos to strengthen the comparison between the two nodule counting methods? The photos of areas where some of the cores have been taken can still be excluded from the RF model and externally validated afterward in order to make best use of available ship time and data.

And a few technical corrections and suggestions:

R56: "data points"? "Data sets"?

R180: could you specify what the correction would be?

R474: "resulting in biased results where Mn-nodules are bigger"?

R480: This is true, but is it necessary to state that here? Maybe it could be moved to the introduction

R476 - 485: It is hard to follow the authors point in that group of sentences. Do you mean that the observed influence of bathymetric factors on nodules distribution cannot necessarily be explained , but this observation is an interesting fact in itself it may later lead to further understanding of an underlying process?

R490: "as it ignores"?

R490: "To this end, several authors, have included the values of latitude/longitude and even LMI as predictor variables"?

516: "thus, high priority areas (e.g. these with highest commercial interest) can be targeted for sampling based on the results of optic data and RF modelling"?

Hope this is helpful

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2018-353, 2018.