Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-23-RC1, 2021

Comment on essd-2021-23

Anonymous Referee #1

Referee comment on "Modelling seabed sediment physical properties and organic matter content in the Firth of Clyde" by Matthew C. Pace et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-23-RC1, 2021

Pace et al. provide quantitative data on seabed properties in the Firth of Clyde; these include sediment composition (mud, sand, and gravel content), whole-sediment median grain size, presence of rock, porosity, permeability, content of particulate organic carbon and nitrogen, areal stocks of organic carbon and nitrogen and mean and maximum bed shear stress. The maps were created based on legacy data from various sources using statistical and machine learning (random forest) methods. In contrast to previous efforts in the literature, the seabed properties were mapped on an unstructured grid, with higher resolution provided near to the coast and lower resolution further offshore.

There has been an increase in the number of studies that attempted to spatially predict seafloor properties quantitatively in recent years. This study adds to the growing body of literature. The interest in quantitative maps of seabed properties is increasing, as it has been realised that data on sediment composition (e.g., mud, sand, and gravel content) are much more flexible than categorical data (e.g., Folk textural classes). Additionally, there is a growing interest in the role the seafloor plays in the marine carbon cycle and in its ability to store organic carbon, but studies estimating organic carbon inventories are still relatively few. This study is therefore a welcome contribution, and the provided data will most likely be of great use for scientists and managers in the context of nature conservation, marine spatial planning, and ecosystem service mapping, among others.

Overall, the study provides a very useful set of spatially predicted and partly derived parameters, some of which are very difficult to measure (e.g., permeability) and all of which are costly to obtain as ship time is expensive. Making best use of existing datasets, as has been done here, is therefore a suitable strategy. The datasets provide full coverage over the Firth of Clyde, covering 3,600 km^2 of seabed. In most instances, the models appear to have acceptable to good predictive performance, apart from gravel content and potentially rock presence. The data ranges and spatial patterns produced appear reasonable to me, when for example judged against the offshore 1:250,000 scale seabed sediments map of the British Geological Survey. The manuscript is generally clearly written and well structured. There are, however, a few open questions and issues that need to be addressed prior to acceptance for publication. These will be detailed in the

following:

Explanatory environmental variables were used to predict rock presence, sediment composition, particulate organic carbon, and particulate organic nitrogen. While these have been submitted to a formal variable selection process as recommended in the literature, it would be good to know why certain predictor variables were chosen in the first place. Usually, such selections are based on a general understanding of the modelled system, experience from previous studies, and data availability. It would be beneficial to briefly outline, which predictor variables were initially chosen and why.

To minimise the impact of spatial autocorrelation on performance estimates, a spatial cross-validation was run with data binned in blocks of 0.125° latitude by 0.25° longitude. It is encouraging to see that spatial autocorrelation is increasingly accounted for in marine modelling studies; however, it would be necessary to explain why the above-mentioned block size was chosen. Usually, the block size should be determined experimentally, e.g., by estimating the spatial autocorrelation range from an empirical variogram. The r package blockCV (Valavi et al., 2019) provides a tool to determine block sizes and might be used here to establish a suitable separation distance.

There is very limited information on the random forest models that were built apart from general information. It would be desirable to include basic information such as the (hyper-)parameters that were chosen and how/why. If the authors feel this would unnecessarily increase the length of the manuscript, the information could be provided as a supplement. Even better would be the provision of the R code, which will increase openness and transparency. Additionally, in the case of the rock model, it would be necessary to explain which threshold value was chosen to convert probabilities into presence/absence predictions. As no information was given, I suspect the "default" of 0.5 was selected. It has, however, been demonstrated that this threshold value is not always optimal and a range of alternative threshold criteria exist (Freeman and Moisen, 2008a, 2008b). I suggest exploring alternative thresholds, assuming the default has been used.

Model performance is reported as area under the curve for rock, r-squared for sediment composition, particulate organic carbon, and particulate organic nitrogen and the explained variance in the case of the median grain-size and permeability. Could the authors explain whether r-squared and explained variance have different definitions here? Additionally, it would be necessary to provide more detail on why certain performance indicators were chosen and others not. For example, in the case of the rock model, which is a binary, presence-absence model, different performance metrics could be used to estimate model calibration and discrimination for continuous and binary outputs (Lawson et al., 2014). The area under the curve measures discrimination of the continuous predictions, but when looking at binary predictions, sensitivity and specificity might be chosen. Calibration might be estimated with the root mean squared error in the case of continuous predictions and mean accuracy in the case of binary predictions.

While sediment composition data were transformed prior to modelling, it is not clear to me whether a transformation was applied to the content of particulate organic carbon and

nitrogen. As these parameters are reported as proportions, it would be advisable to apply an arcsine transformation (Sokal and Rohlf, 1981).

Permeability was predicted with median grain size and mud content in two different models; however, porosity was predicted with median grain size only. This is a little surprising, as previous studies found close relationships between porosity and mud content (e.g. Jenkins, 2005; Silburn et al., 2017) and such a prediction would be "one step closer to measured data", as the authors put it. Was there a specific reason why this relationship was not considered? If not, it might be advisable to investigate whether a similar relationship could be found based on the authors' dataset.

I would also suggest using a different colour palette then the red-white-blue palette used in most figures. Such a palette would suggest a pattern diverging from a central value rather than a continuous increase. See for example Crameri et al. (2020) for advice on choosing suitable colour palettes.

Finally, I noticed that the data doi is not working (yet). This will have to be fixed. I was, however, able to obtain the datasets from https://pureportal.strath.ac.uk/. Datasets are provided in csv and netCDF format. I assume this is because the predictions were made on an unstructured grid. If possible, I would suggest to additionally supply outputs as georeferenced tiff files, as this is a common format in the science community working with geographic information systems.

I am also providing some minor comments in an annotated version of the manuscript.

Overall, I congratulate the authors to an interesting and well executed study that is summarised succinctly and provides important datasets. I recommend accepting the manuscript subject to **minor revisions**.

Please also note the supplement to this comment:
https://essd.copernicus.org/preprints/essd-2021-23/essd-2021-23-RC1-supplement.pdf