General notes

Comments by referees in bold.

Response in normal.

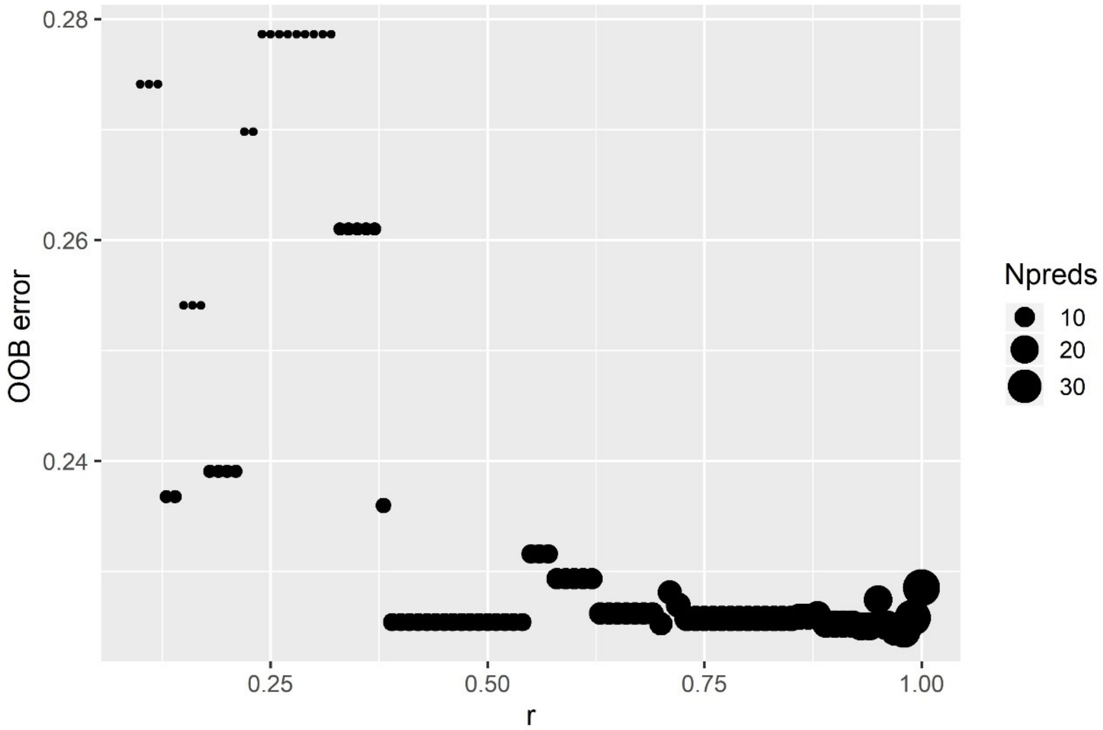
Changes made in italics.

All chapter, line, table, and figure numbers refer to the marked-up manuscript.

Referee 1 was generally positive about the submitted manuscript, but pointed out the following issues that should be addressed:

1. The importance of eliminating correlated input variables was not discussed. Considering that the Random Forests algorithm allows for non linear input relationships, a justification for this does not seem obvious, particularly since its results contradict those of the Boruta algorithm, which deemed all 38 predictors as important.

The selection of predictor variables/features is a two-step process. The first step (Boruta algorithm) narrows the selection to those variables that are relevant. In this case, all variables were identified as relevant/important. It is important to note that the Boruta algorithm is an “all-relevant” feature selection method (Nilsson et al., 2007), which identifies all predictors that might be relevant for classification (Kursa and Rudnicki, 2010). It does not address the question of redundancy in the predictor variable data. To limit redundancy, the second step seeks to identify predictor variables that are correlated with other predictors of higher importance. To do so, it is necessary to define a critical value of the correlation coefficient r. This is arguably subjective, and the choice of r will influence the number of predictor variables that are selected. To investigate the influence of r on the OOB error of a Random Forest model (using default parameter values), several values of r between 0.1 and 1 (step 0.01) were now trialled. When the OOB error is plotted over r (Figure below), it is apparent that initially (with small values of r) the OOB error drops off quickly but stabilises at values of r ≈ 0.4 – 0.5. As a large number of predictors potentially leads to overfitting, increases processing time and decreases interpretability of the model, it was decided to select a small value of 0.5, leading to the selection of eight predictor variables (see section 4.1 Variable selection).



Influence of r-value on the out of bag error of a random forest model with default parameters. The size of the circles indicates the number of selected predictor variables (Npreds).

Referee 1 argues that the Random Forest algorithm allows for non-linear input relationships and hence the elimination of correlated input variables was unnecessary. This claim has indeed frequently been made; however, some studies recommend reducing high-dimension datasets to uncorrelated important variables (Millard and Richardson, 2015). In this case, the number of predictor variables has limited influence on prediction performance with r-values above ≈ 0.4 but marked influence on processing time and interpretability of the results as the number of predictors increases from eight up to 38. Therefore, a low number of predictors was selected through the choice of r = 0.5. This selection of eight variables is also not in contradiction to the results of the Boruta analysis, as the Boruta algorithm identifies relevant features (all-relevant problem, Nilsson et al., 2007), while for model building a low number of variables that achieve comparatively high accuracy was the goal (minimal-optimal problem, Nilsson et al., 2007).

Changes were made to sections 3.2 Predictor variable selection and 4.1 Variable selection (l.202-204). Figure 4 was introduced to clarify the choice of the r-value (same as figure above).

2. The selection of 7 lithology classes as response variables instead of 5 or 13 seemed arbitrary, but the repercussions of this choice are big: the great disparity of class contribution to the test dataset (44.8% for the most represented class v.s. 0.9% for the least represented class) render model accuracy less suitable as a performance metric.

Following the advice from reviewer 1, it was decided to use the five classes of Dutkiewicz et al. (2016). These are broadly in line with lithology classes depicted in textbook maps and comprise Calcareous sediment, Clay, Diatom ooze, Radiolarian ooze and Lithogenous sediment.

This choice still introduced large imbalances in class frequencies, as pointed out by both reviewers. To deal with the issue, a balanced version of Random Forest has now been utilised. This is achieved by using the strata and sampsize arguments of the randomForest function. In this case, we stratify by lithology class. The sampsize is then set to the same number for every class. This means that the class with the lowest frequency (Radiolarian ooze) determines the number of observations used to fit individual trees. However, each sample is still drawn from all available observations and hence this approach is likely more effective then downsampling the whole dataset prior to model building.

Relevant changes were made to sections 2.2 Response variable (l.65-66), 3.1 Data pre-processing (l.81-82), 3.4 Random Forest classification model (l.144-149) and Table 2.

3. Replacing accuracy with another performance metrics such as Intersection over Union (averaging the IoU for all individual classes) would account for a more fair result interpretation and allow for better performance comparisons in future works.

I believe that overall accuracy as a metric for the performance of the model still has some meaning. Basically, if a map has an accuracy of say 60%, then 60 out of 100 randomly placed points in the map will likely be classified correctly. That information is still of importance. However, I agree that this metric is less well suited to give information on rare classes. Such information was conveyed by the class-specific metrics error of commission and error of omission. To further address this shortcoming, I am now also providing the Balanced Error Rate (BER). The BER treats all mapped classes equally, as it is the mean error of all mapped classes, regardless of how often a class is contained in the test set. Together, both global metrics give a more comprehensive picture to assess the accuracy of the map.

Relevant changes were made to sections 3.6 Accuracy assessment and 4.3 Model accuracy.

Referee 2 attested overall good quality, but identified several issues that could be further addressed:

1. Move the section 3.5 Environmental Space before section 3.3 Random Forest classification model.

Agreed, the section was moved accordingly.

2. Similar, the results of this analysis can be presented before the model results.

Agreed, the section was moved accordingly.

3. It is not clear which correlation coefficient has been used, and the significance of these correlations. The reasons for excluding high-correlated predictor variables could be analysed further, mainly focused on the RF performance and potential bias in cases of high-correlated predictors.

The general strategy for predictor variable selection was to initially find all predictors that are potentially relevant, then reduce redundancy by limiting the set of predictors to those that are uncorrelated. Such an approach has also been advocated by Millard and Richardson (2015). Finding important predictors is achieved with the Boruta algorithm, an “all-relevant” feature selection method (Nilsson et al., 2007), which identifies all predictors that might be relevant for classification (Kursa and Rudnicki, 2010). Limiting redundancy is subsequently achieved by a correlation analysis. To do so, it is necessary to define a critical value of the correlation coefficient r. This is arguably subjective, and the choice of r will influence the number of predictor variables that are selected. To investigate the influence of r on the OOB error of a Random Forest model (using default parameter values), several values of r between 0.1 and 1 (step 0.01) were now trialled. When the OOB error is plotted over r, it is apparent that initially (with small values of r) the OOB error drops off quickly but stabilises at values of r ≈ 0.4 – 0.5. As many predictors potentially lead to overfitting, increase processing time, and decrease interpretability of the model, it was decided to select a small value of 0.5. This led to eight predictors being selected.

For clarity, section 3.2 Predictor variable selection was updated.

4. Dutkiewicz et al., 2015 and Dutkiewicz et al., 2016, referred to the absence of iron concentration as potential predictor variable form their model and correlation analysis, respectively. In this study, the author initially includes iron concentration as a potential predictor variable. However, in the final model, the iron concentration is not used.

Dutkiewicz et al. (2016) highlighted that iron concentration in surface waters might be an important predictor, as phytoplankton blooms (e.g. diatoms) are enhanced by iron fertilisation. However, they also point out that the Southern Ocean where diatom oozes are abundant, receives very little iron. In fact, productivity is not the only factor determining the composition of deep-sea sediments, dissolution during sinking and dilution with other materials also have to be considered.

Iron concentration was initially included in this study. However, iron concentration was not included in the final model because it was correlated (above the selected threshold) with sea-floor minimum temperature. Box plots of iron concentration versus seafloor lithology did not reveal high discriminatory power of this predictor. Because of that and because it was intended to keep the variable selection process as “automated” as possible, it was decided not to intervene and force the inclusion of iron concentration as a finally selected predictor.

No changes were made to the manuscript.

5. It would be interesting to be discussed the presence of available global CCD models (and its limitations) that could be used or not as a predictor.

The Carbonate Compensation Depth (CCD) is potentially another important predictor that was not included in the model. This was due to the absence of relevant datasets in the utilised databases. A literature search did not yield any publications with associated datasets that could be used. However, it is assumed that this missing information is (at least partly) provided by other predictors, such as water depth and bottom water temperature.

No changes were made to the manuscript.

6. The author uses the unscaled mean decrease in accuracy as a variable importance measure. Although this is the recommended approach (e.g. Strobl et al., 2007; Strobl et al., 2008), it would be better if the reasons behind this choice are stated inside the text.

Agreed.

The text of section 3.4 Random Forest classification model has been updated accordingly and now reads:

RF also provides a relative estimate of predictor variable importance. The importance() function of the randomForest package allows to assess variable importance as the mean decrease in either accuracy or node purity. However, the latter approach might be biased when predictor variables vary in their scale of measurement or their number of categories (Strobl et al., 2007) *and was not used here. Variable importance is therefore measured as the mean decrease in accuracy associated with each variable when it is assigned random but realistic values and the rest of the variables are left unchanged. The worse a model performs when a predictor is randomised, the more important that predictor is in predicting the response variable. The mean decrease in accuracy was left unscaled as recommended by Strobl and Zeileis* (2008)*, and is reported as a fraction ranging from 0 to 1.*

7. In the manuscript, it is not stated if the RF sub-sampling of predictor variables is performed with or without replacement. In the provided script seems to be with replacement (as the default option in the used package). However, studies have shown that this approach can be biased when predictor variables vary in their scale and/or in their number of categories (e.g. Strobl et al., 2007). Also, it is not also clear if any type of feature scaling has been applied in the predictor variables before modelling. In case that the author would like to continue without feature scaling for the predictor variables, packages like party seem to provide more unbiased results. In any case, it would be good to be provided with a more comprehensive explanation, as the model interpretability is one of the main targets in this work.

Agreed. The predictor variables have now been scaled and the subsampling of predictor variables is performed without replacement.

The manuscript was updated accordingly. Relevant changes were made to sections 3.1 Data pre-processing (l.80) and 3.4 Random Forest classification model (l.157-159).

8. Studies have shown that the traditional cross-validation can result in overoptimistic errors when applied in spatial data (e.g. Roberts et al., 2016). Consequently, a spatial model should also include a spatial cross-validation analysis. Moreover, in cases of spatially unbalanced class distribution, stratified cross-validation can be applied (e.g. Lawson et al., 2017). Here, train & test samples were split with stratification, but the training was conducted only with cross-validation. Recently published spatial versions of RF could alternatively help to this direction (e.g. Hengl et al, 2018,

Georganos et al, 2019). The existing dataset (despite the tremendous effort of Dutkiewicz et al., 2015) is spatially imbalanced, with areas have experienced heavier sampling efforts than other, making even more important the concept of spatial cv or/and spatial RF. The author addresses this issue by setting a minimum distance among the training points and by removing duplicates.

Agreed. I have now implemented a spatial leave one out cross validation scheme to test the accuracy of the model in a more robust way. The spatial autocorrelation distance is determined with the spatialAutoRange function of the blockCV package. This value is utilised to determine the buffer size around observations which serve as a test point. Details can be found in a new section "3.5 Spatial cross-validation".

A new section 3.5 Spatial cross-validation was introduced.

However, this meant that model tuning would have become very complex and even more time-consuming. As model tuning gave a very limited gain in performance, it was therefore decided to run the Random Forest with default parameter values.

The former sections on model tuning were deleted.

As a result of this more robust estimation of map accuracy, the accuracy of the model is now lower (or rather less inflated), but still significantly larger than the no information rate.

The section 4.3 Model accuracy was updated accordingly.

9. Despite the selected class simplification, the two main classes still count for the 77.6% of the training and test sample, resulting in relatively high overall accuracy, but with limited accuracy in the rare classes. Considering the availability of methods and algorithms that try to overcome these class imbalances, (e.g. weights, penalty costs, over/under-sampling. SMOTE, Isolation Forests) it would be interesting to see how further can be improved the performance compared to the presented (baseline) model. The overall accuracy as a performance metric is not the best option in such situations (it is also mentioned from the first Reviewer) However, the no information rate is provided, showing that there is still gain.

The problem of class imbalances is now addressed by utilising a balanced version of Random Forest. This is achieved by using the strata and sampsize arguments of the randomForest function. Here, we stratify by lithology class. The sampsize is then set to the same number for every class. This means that the class with the lowest frequency (Radiolarian ooze) determines the number of observations used to fit individual trees. However, each sample is still drawn from all available observations and hence this approach is likely more effective then downsampling the whole dataset prior to model building.

Relevant changes were made to section 3.4 Random Forest classification model (l.144-149).

In addition to the overall accuracy, I have now also included the balanced error rate, which is more appropriate for datasets with unbalanced class frequencies.

Relevant changes were made to sections 3.6 Accuracy assessment and 4.3 Model accuracy.

As a result, the final map has a different appearance in some areas of the global ocean. It now approximates hand-drawn maps of the distribution of deep-sea sediments in much more detail. For example, equatorial patches of radiolarian ooze in the Indian Ocean are visible now. A near-equatorial band of radiolarian ooze in the eastern Pacific is now visible, too.

Changes were made to section 4.4 Spatial distribution of deep-sea sediments, Figures 6, 7 and A2 and Table 4.

10. The results show good overall agreement with the above mentioned previous mapping efforts. However, the comparison demands parallel examinations of the maps from the two papers. Given the availability of the results from Dutkiewicz et al., 2015, it would be interesting to include a direct categorical map comparison between the two approaches (after the proper modifications due to the different number of lithological classes) showing clearer the areas with the highest agreement and disagreement.

It was not the intention of this contribution, and might go beyond the scope of a data description paper, to compare the final map with that of Dutkiewicz et al. (2015). I would prefer to leave such an analysis to whoever is interested in it. The map products of both publications are readily available.

No changes were made to the manuscript.

11. In Figures 1 & 6a, the use of purple colour for the Mixed Ooze is not ideal, as it has limited contrast with the background map and its surrounded classes. An essential part of this study and its results is related to map creation and interpretation. Consequently, the use of colourblind-friendly palette is recommended, making the manuscript more comfortable on a broader audience.

I agree that the choice of the colour scheme should be inclusive, but when consulting colorbrewer2.org, I could not find a colour-blind safe option for qualitative data with 5 classes. (NB: The classification has been reduced to the five classes used by Dutkiewicz et al. (2016), as suggested by reviewer 1.) The purple colour has, nevertheless, now been removed, as the respective class is no longer included. I also removed the hillshade bathymetry to make the map clearer.

Changes were made to Figures 1 and 6.

References

Dutkiewicz, A., Müller, R. D., O'Callaghan, S. and Jónasson, H.: Census of seafloor sediments in the world's ocean, Geology, 43(9), 795–798, doi:10.1130/G36883.1, 2015.

Dutkiewicz, A., O'Callaghan, S. and Müller, R. D.: Controls on the distribution of deep-sea sediments, Geochemistry, Geophys. Geosystems, 17(8), 3075–3098, doi:10.1002/2016GC006428, 2016.

Kursa, M. and Rudnicki, W.: Feature selection with the Boruta Package, J. Stat. Softw., 36(11), 1–11 [online] Available from: http://www.jstatsoft.org/v36/i11/paper/, 2010.

Millard, K. and Richardson, M.: On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping, Remote Sens., 7(7), 8489–8515, doi:10.3390/rs70708489, 2015.

Nilsson, R., Peña, J. M., Björkegren, J. and Tegnér, J.: Consistent feature selection for pattern recognition in polynomial time, J. Mach. Learn. Res., 8, 589–612, 2007.

Strobl, C. and Zeileis, A.: Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance, Munich. [online] Available from: https://epub.ub.uni-muenchen.de/2111/1/techreport.pdf, 2008.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution, BMC Bioinformatics, 8(1), 25, doi:10.1186/1471-2105-8-25, 2007.

Deep-sea sediments of the global ocean

Markus Diesing[1]

[1]Geological Survey of Norway (NGU), Postal Box 6315 Torgarden, 7491 Trondheim, Norway

Correspondence to: Markus Diesing (markus.diesing@ngu.no)

Abstract. Although the deep-sea floor accounts for more than 70 % of the Earth's surface, there has been little progress in relation to deriving maps of seafloor sediment distribution based on transparent, repeatable, and automated methods such as machine learning. A new digital map of the spatial distribution of seafloor lithologies in the deep sea below 500 m water depth is presented to address this shortcoming. The lithology map is accompanied by estimates of the probability of the most probable class, which may be interpreted as a spatially-explicit measure of confidence in the predictions, and probabilities for the occurrence of ~~seven~~ five lithology classes (Calcareous sediment, Clay, Diatom ooze, Lithogenous sediment, and ~~Mixed calcareous-siliceous ooze,~~ Radiolarian ooze ~~and Siliceous mud~~). These map products were derived by the application of the Random Forest machine learning algorithm to a homogenised dataset of seafloor lithology samples and global environmental predictor variables that were selected based on the current understanding of the controls on the spatial distribution of deep-sea sediments. ~~The overall accuracy of the lithology map is 69.5 %, with 95 % confidence limits of 67.9 % and 71.1 %.~~ It is expected that the map products are useful for various purposes including, but not limited to, teaching, management, spatial planning, design of marine protected areas and as input for global spatial predictions of marine species distributions and seafloor sediment properties. The map products are available at https://doi.pangaea.de/10.1594/PANGAEA.911692 (Diesing, 2020).

Commented [DM1]: Requires updating

1 Introduction

The deep-sea floor accounts for >90 % of seafloor area (Harris et al., 2014) and >70 % of the Earth's surface. It acts as a receptor of the particle flux from the surface layers of the global ocean, is a place of biogeochemical cycling (Snelgrove et al., 2018), records environmental and climate conditions through time and provides habitat for benthic organisms (Danovaro et al., 2014). Being able to map the spatial patterns of deep-sea sediments is therefore a major prerequisite for many studies addressing aspects of marine biogeochemistry, deep-sea ecology, and palaeo-environmental reconstructions.

Until recently, maps of global deep-sea sediments were essentially variants of a hand-drawn map presented by Berger (1974) and typically depicted five to six sediment types, namely calcareous ooze, siliceous ooze (sometimes split into diatom ooze and radiolarian ooze), deep-sea (abyssal) clay, terrigenous sediment and glacial sediment. Since then, Dutkiewicz et al. (2015) collated and homogenised approximately 14,500 samples from original cruise reports and interpolated them using a support vector machine algorithm (Cortes and Vapnik, 1995). Their map displayed the spatial distribution of 13 lithologies across the world ocean and exhibited some marked differences from earlier maps.

The controls on the distribution of deep-sea sediments have long been discussed (e.g. Seibold and Berger, 1996): Biogenous oozes (>30 % microscopic skeletal material by weight) dominate on the deep-sea floor and their composition is controlled by productivity in overlying surface ocean waters, dissolution during sinking and sedimentation and dilution with other materials. The ocean is undersaturated with silica. Preservation of siliceous shells is therefore a function of shell thickness, sinking time (water depth) and water temperature, as siliceous shells dissolve slower in colder water. The dissolution of calcareous shells is increased with increasing pressure (water depth) and CO_2 content of the water (decreasing temperature). The water depth at which the rate of supply with calcium carbonate to the sea floor equals the rate of dissolution (calcite compensation depth; CCD) varies across ocean basins. Deep-sea clays dominate in the deepest parts of ocean basins below the CCD. Deposition of terrigenous material is thought to be a function of proximity to land (distance to shore).

Dutkiewicz et al. (2016) investigated the bathymetric and oceanographic controls on the distribution of deep-sea sediments with a quantitative machine-learning approach. The influence of temperature, salinity, dissolved oxygen, productivity, nitrate, phosphate, silicate at the sea surface and bathymetry on lithogenous sediment, clay, calcareous sediment, radiolarian ooze and diatom ooze was quantified. They found that bathymetry, sea surface temperature and sea surface salinity had the largest control on the distribution of deep-sea sediments. Calcareous and siliceous oozes were not linked to high surface productivity according to their analysis. Diatom and radiolarian oozes were associated with low sea surface salinities and discrete sea surface temperature ranges.

The aim of this study is to derive a map of deep-sea sediments of the global ocean by utilising environmental predictor variables for the development and application of a machine-learning spatial prediction model. Besides a categorical map giving the spatial representation of seafloor types in the deep sea, probability surfaces for individual sediment classes and a map displaying the probability of the most probable class in the final prediction will also be provided.

2 Data

2.1 Predictor variables

The initial choice of the predictor variables was informed by the current understanding of the controls on the distribution of deep-sea sediments and the availability of data with full coverage of the deep sea at a reasonable resolution. We chose predictor variables mentioned above, but also included sea-surface iron concentration~~s~~, which was~~ere~~ not available to Dutkiewicz et al. (2016), but is an important nutrient for phytoplankton (Table 1). The predictor variable raster layers from Bio-ORACLE (Assis et al., 2018; Tyberghein et al., 2012) and MARSPEC (Sbrocco and Barber, 2013) were utilised. Whenever available, statistics of the variable other than mean were downloaded. These included the minimum, maximum and the range (maximum – minimum).

2.2 Response variable

The response variable is seafloor lithology, a qualitative multinomial variable. The seafloor sediment sample data (seafloor_data.npz) from Dutkiewicz et al. (2015) were downloaded from ftp://ftp.earthbyte.org/papers/Dutkiewicz_etal_seafloor_lithology/iPython_notebook_and_input_data/. The original dataset consisted of 13 seafloor lithology classes, while Dutkiewicz et al. (2016) simplified these to five major classes. The latter scheme was chosen here (Table 2), as the five major classes agree well with lithologies typically depicted in hand-drawn maps.~~As a compromise, seven lithology classes (Table 2) were defined: These include those five classes defined in Dutkiewicz et al. (2016), supplemented with mixed lithologies (calcareous-siliceous ooze and siliceous mud).~~ For a detailed description of the original lithology classes refer to GSA Data Repository 2015271 (https://www.geosociety.org/datarepository/2015/2015271.pdf).

3 Methods

The general workflow for building a predictive spatial model was outlined by Guisan and Zimmermann (2000). This involves five main steps: (1) Development of a conceptual model, (2) statistical formulation of the predictive model, (3) calibration (training) of the model, (4) model predictions and (5) evaluation of the model results (accuracy assessment). The conceptual model was already presented in the introduction. The remaining steps are described in the following sections. The analysis was performed in R 3.6.1 (R Core Team, 2018) and RStudio 1.2.1335 and is documented as an Executable Research Compendium (ERC), which can be accessed at https://o2r.uni-muenster.de/#/erc/GWME2voTDb5oeaQFuTWMCEMveKS1MiXm.

Commented [DM2]: Requires updating.

3.1 Data pre-processing

The raster layers (predictor variables) were stacked, limited to water depths below 500 m, ~~and~~ projected to Wagner IV global equal-area projection with a pixel resolution of 10 km by 10 km and scaled.

The sample data (response variable) were pre-processed in the following way: Only samples of the five major lithologies (Table 2) deeper than 500 m were used and ~~a minimum distance between sample locations of 14.5 km (the diagonal of a 10 km pixel) was enforced to limit the number of samples per pixel to one. This had also the effect of removing~~ duplicates were removed from the original sample dataset. ~~and reduced the~~ The number of records was therefore reduced from 14,400 to ~~10,190~~10,438. The data were projected to Wagner IV. Locations of the sample locations and their respective lithology class are shown in Fig. 1.

Predictor variable values were extracted for every sample location. ~~The resulting data frame was subsequently split into a training and a test set with a ratio of 2:1 using a stratified random approach. The stratification was based on the lithology class. This ensured that the test set contained approximately the same~~The class frequencies ~~as the training set (~~are shown in Table 2~~)~~.

3.2 Predictor variable selection

Variable selection reduces the number of predictor variables to a subset that is relevant to the problem. The aims of variable selection are three-fold: (1) to improve the prediction performance, (2) to enable faster predictions and (3) to increase the interpretability of the model (Guyon and Elisseeff, 2003). It is generally advisable to reduce high-dimension datasets to uncorrelated important variables (Millard and Richardson, 2015). Here, a two-step approach was utilised to achieve this goal. The first step identifies those variables that are relevant to the problem. The second step minimises redundancy in the remaining predictor variables.

~~Predictor variables were selected in a two-step approach:~~ Initially, the Boruta variable selection wrapper algorithm (Kursa and Rudnicki, 2010) was employed to identify all potentially important predictor variables. Wrapper algorithms identify relevant features by performing multiple runs of predictive models, testing the performance of different subsets (Guyon and Elisseeff, 2003). The Boruta algorithm creates so-called shadow variables by copying and randomising predictor variables. Variable importance scores for predictor and shadow variables are subsequently computed with the Random Forest algorithm (see below). The maximum importance score among the shadow variables (MZSA) is determined and for every predictor variable, a two-sided test of equality is performed with the MZSA. Predictor variables that have a variable importance score significantly higher than the MZSA are deemed important, while those with a variable importance score significantly lower than the MZSA are deemed unimportant. Tentative variables have a variable importance score that is not significantly different from the MZSA. Increasing the maximum number of iterations (maxRuns) might resolve tentative variables (Kursa and Rudnicki, 2010). Only important variables were retained for further analysis.

The Boruta algorithm is an "all-relevant" feature selection method (Nilsson et al., 2007), which identifies all predictors that might be relevant for classification (Kursa and Rudnicki, 2010). It does not address the question of redundancy in the predictor variable data, which would be required for "minimal optimal" feature selection (Nilsson et al., 2007) usually preferred for model building. –To limit redundancy, a second step seeks to identify predictor variables that are correlated with other predictors of higher importance. ~~In a subsequent step~~To achieve this, the Boruta importance score was used to rank the remaining predictor variables. Beginning with the most important variable, correlated variables ~~(|r| > 0.5)~~ with lower importance were subsequently removed. Values of the correlation coefficient r were trialled between 0.1 and 1 with a step size of 0.01 to find an appropriate r value that strikes a balance between prediction performance and model interpretability. ~~As a result, only important and uncorrelated predictor variables are retained.~~

3.3 Environmental space

It is generally preferable to apply a suitable sampling design for model calibration and evaluation. This would ensure that the environmental variable space is sampled in a representative way. Various methods have been proposed to optimise sampling effort, including stratified random, generalised random tessellation stratified (Stevens Jr and Olsen, 2003) and conditioned Latin hypercube sampling (Minasny and McBratney, 2006) among others. However, such approaches are not feasible here due to time and financial constraints. Instead, we utilised available (legacy) sampling data. It might nevertheless be prudent to assess to what extent the selected samples cover the environmental space of the predictor variables. This was achieved by creating a random subsample (n = 10,000) of the selected environmental predictor variables and displaying the density distribution of the random subsample together with the density distribution of environmental variables based on the observations. This allows for a qualitative check to what degree the environmental space is sampled in a representative way.

3.4~~3~~ Random Forest classification model

The Random Forest (RF) prediction algorithm (Breiman, 2001) was chosen for the analysis due to its high predictive performance in a number of domains (Che Hasan et al., 2012; Cutler et al., 2007; Diesing et al., 2017; Diesing and Thorsnes, 2018; Huang et al., 2014; Prasad et al., 2006). The RF is an ensemble technique based on classification trees (Breiman, 1984). Randomness is introduced in two ways: by constructing each tree from a bootstrapped sample of the training data, and by using a random subset of the predictor variables at each split in the tree growing process. As a result, every tree in the forest is unique. By aggregating the predictions over a large number of uncorrelated trees, prediction variance is reduced and accuracy improved (James et al., 2013, p. 316). The 'votes' for a specific class can be interpreted as a measure of probability for that class occurring in a specific location. The final prediction is determined by the class with the highest probability (vote count) to occur in a specific location. The randomForest package (Liaw and Wiener, 2002) was used to perform the analysis.

RF generally performs well with default settings, i.e. without the tuning of parameters. ~~However, it might be advisable to tune those parameters that have the largest impact on predictive accuracy. These are~~Initial tuning of the number of trees in the forest (n_{tree}) and the number of variables to consider at any given split (m_{try}) showed a very limited impact on model performance, while at the same time the tuning process was very time-consuming.~~. Those two parameters were tuned in a grid search with the package e1071 (Meyer et al., 2019) using a 10-fold cross-validation scheme with three repeats on the training dataset.~~ It was therefore decided to use the default parameter values.

The response variable is highly imbalanced (Table 2). This was accounted for by utilising a balanced version of RF (Chen et al., 2004). This is achieved by specifying the strata and sampsize arguments of the randomForest() function. The strata are the lithology classes and the sample size is determined by $c \cdot nmin$, where c is the number of lithology classes (seven) and $nmin$ is the number of samples in the least frequent class. Hence, downsampling is applied when growing individual trees. However, each sample is drawn from all available observations as many trees are grown, making this scheme likely more effective than downsampling the dataset prior to model building.

RF also provides a relative estimate of predictor variable importance. The importance() function of the randomForest package allows to assess variable importance as the mean decrease in either accuracy or node purity. However, the latter approach might be biased when predictor variables vary in their scale of measurement or their number of categories (Strobl et al., 2007) and was not used here. ~~This~~ Variable importance is therefore measured as the mean decrease in accuracy associated with each variable when it is assigned random but realistic values and the rest of the variables are left unchanged. The worse a model performs when a predictor is randomised, the more important that predictor is in predicting the response variable. The mean decrease in accuracy was left unscaled as recommended by Strobl and Zeileis (2008), and is reported as a fraction ranging from 0 to 1. Per default, individual trees of the forest are built using sampling with replacement (replace=TRUE). However, it has been shown that this choice might lead to bias in predictor variable importance measures (Strobl et al., 2007). It was therefore opted to use sampling without replacement.

3.5 Spatial cross-validation

Detailed guidelines for optimising sampling design for accuracy assessment have been developed (Olofsson et al., 2014; Stehman and Foody, 2019). However, this would require collecting new samples after modelling, which was not feasible given the geographic scope. Cross-validation schemes are frequently used to deal with such situations. It can be assumed that the response variable is spatially structured to some extent and cross-validation therefore requires accounting for the spatial structure (Roberts et al., 2017). Here, a spatial leave-one-out cross-validation (S-LOO CV) scheme was applied. In a conventional LOO CV, a single observation is removed from the dataset and all other observations (n – 1) are used to train the model. The class of the withheld observation is then predicted using the n – 1 model. This is repeated for every observation in the dataset, producing observed and predicted classes at every location. In a S-LOO CV scheme, a buffer is placed around the withheld observation and training data from within this buffer are omitted from both model training and testing so that there are no training data proximal to the test. The S-LOO CV scheme used here was adapted from Misiuk et al. (2019). The buffer size was estimated with the spatialAutoRange() function of the blockCV package (Valavi et al., 2018).

3.6~~4~~ Accuracy assessment

The accuracy of the model was assessed based on a confusion matrix that was derived by ~~predicting the model on the test set~~the S-LOO CV. Overall accuracy and the balanced error rate (BER) were~~as~~ used to evaluate the global accuracy of the model, while error of omission and error of commission were selected as class-specific metrics of accuracy. The overall accuracy gives the percentage of cases correctly allocated and is calculated by dividing the total number of correct allocations by the total number of samples (Congalton, 1991). The BER is the average of the error rate for each class (Luts et al., 2010). The error of omission is the number of incorrectly classified samples of one class divided by the total number of reference samples of that class. The error of commission is the number of incorrectly classified samples of one class divided by the total number of samples that were classified as that class (Story and Congalton, 1986). The overall accuracy, its 95% confidence intervals and a one-sided test to evaluate whether the overall accuracy was significantly higher than the no information rate (NIR) were calculated by applying the confusionMatrix() function of the caret package (Kuhn, 2008). The confidence interval is estimated using a binomial test. The NIR is taken to be the proportion of the most frequent class. Errors of omission and commission are not provided by the function but can be calculated from the confusion matrix. The BER was calculated with the BER() function of the package measures (Probst, 2018).

~~3.5 Environmental space~~

~~It is generally preferable to apply a suitable sampling design for model calibration and evaluation. This would ensure that the environmental variable space is sampled in a representative way. Various methods have been proposed to optimise sampling effort, including stratified random, generalised random tessellation stratified (Stevens Jr and Olsen, 2003) and conditioned Latin hypercube sampling (Minasny and McBratney, 2006) among others. Guidelines for optimising sampling design for accuracy assessment have also been developed (Olofsson et al., 2014; Stehman and Foody, 2019). However, such approaches are not feasible here due to time and financial constraints. Instead, we utilised available (legacy) sampling data, which was split into training and accuracy testing data sets. It might nevertheless be prudent to assess to what extent the selected samples cover the environmental space of the predictor variables. This was achieved by creating a random subsample (n = 10,000) of the selected environmental predictor variables and displaying the density distribution of the random subsample together with the density distribution of environmental variables based on the training data. This allows for a qualitative check to what degree the environmental space is sampled in a representative way.~~

4 Results

4.1 Variable selection

The Boruta algorithm was run with maxRuns = 500 iterations and a *p*-value of 0.05, leaving no variables unresolved (i.e. tentative). All 38 predictor variables initially included in the model were deemed important according to the Boruta analysis (Fig. 2). Based on a plot of the RF out of bag (OOB) error estimates over the correlation coefficient r, a value of 0.5 was selected (Fig. 3) This selection ensured high model performance while at the same time minimising the number of predictor variables. Subsequent ~~C~~correlation analysis reduced the number of retained predictor variables to eight. These were bathymetry (MS_bathy_5m), distance to shore (MS_biogeo5_dist_shore_5m), ~~sea-floor maximum temperature (BO2_tempmax_bdmean),~~ sea-surface temperature range (BO2_temprange_ss), sea-surface maximum primary productivity (BO2_ppmax_ss), sea-floor minimum temperature (BO2_tempmin_bdmean), sea-surface maximum salinity (BO2_salinitymax_ss), sea-surface salinity range (BO2_salinityrange_ss) and sea-surface minimum silicate (BO2_silicatemin_ss). The strongest correlation between remaining predictor variables (Fig. ~~3~~4) was found between bathymetry and sea-floor m~~ax~~inimum temperature (r = 0.38), sea-surface maximum salinity ~~range~~ and sea-surface minimum silicate (r = -0.36~~5~~) and bathymetry and distance to shore (r = -0.33~~4~~). Maps of the selected predictor variables are shown in Fig. A1.

4.2~~6~~ Environmental space

The environmental space (Fig. 5~~8~~) is generally sampled adequately, although there is a tendency for an over-representation of shallower water depths ~~(above 3,000 m)~~ and areas closer to land ~~(less than 500 km)~~. ~~Sea-floor maximum temperature, s~~Sea-surface temperature range, sea-surface maximum primary productivity, sea-floor minimum temperature, –and sea-surface maximum salinity are all slightly biased towards higher values. Sea-surface salinity range and sea-surface minimum silicate are the environmental variables that are most closely represented by the samples.

~~4.2 Model tuning~~

~~The RF model was tuned with m_{try} ranging between 2 and 9 (the maximum number of predictors) and n_{tree} assumed values between 100 and 2000 with steps of 100. The covered range was based on previous experience. The optimum parameter combination, based on the classification error, was $m_{try} = 2$ and $n_{tree} = 600$. The differences in classification error were however small, with a maximum difference of 0.95 % between the best and the worst performing combination (Fig. 4).~~

4.3 Model accuracy

The confusion matrix based on the ~~test set~~S-LOO CV is shown in Table 3. The overall accuracy of the model is ~~69.5~~59.4 %, with 95 % confidence limits of ~~67.9~~58.4 % and ~~71.1~~60.3 %. This is significantly higher (p < 2.2e-16) than the NIR (~~44.8~~50.3 ~~-~~%). The BER is 0.56. The two dominant classes, Calcareous sediment, and Clay~~, which make up approximately 78 % of the observations~~ have the lowest error of commission with ~~13.0~~18.0 % and ~~28.8~~32.7 %, respectively. Calcareous sediment is most frequently mis-classified as Clay and vice versa. ~~Diatom ooze has the third lowest error of omission (41.1 %), despite being only the fifth frequent class.~~ All other classes ~~are rare and~~ have high errors of commission (>~~65~~ 70 %). Errors of ~~e~~co~~m~~mission are slightly higher than those of commission for the frequently occurring lithologies ~~(~~Calcareous sediment, and Clay~~ and Diatom mud)~~, while ~~considerably~~ lower for the rare classes ~~(~~Diatom ooze, Lithogenous sediment~~, Mixed calcareous-siliceous sediment~~, and Radiolarian ooze~~ and Siliceous mud)~~.

4.4 Spatial distribution of deep-sea sediments

Probability surfaces of individual sediment classes with verbal descriptions of likelihood (Mastrandrea et al., 2011) based on the estimated probabilities are displayed in Fig. ~~5~~6. For any given pixel in the map, the final lithology class is that one with the highest probability. The probability of the most probable class might be interpreted as a spatially explicit measure of map confidence. The resulting maps of the spatial distribution of deep-sea sediments and their associated confidence are shown in Fig. ~~6~~7. Calcareous sediment and Clay dominate throughout the Pacific, Atlantic and Indian Oceans, whereby Clay occupies the deep basins and Calcareous sediment is found in shallower parts of the ocean basins. In the Southern Ocean, seafloor sediments are arranged in a banded pattern around Antarctica, with ~~Siliceous mud closest to land, followed by Clay and~~ Lithogenous sediment forming an inner ring closest to land (Fig. A2). An outer ring of siliceous oozes~~, mainly~~ (Diatom ooze, and Radiolarian ooze) dominates in the Southern Ocean. The width of this “opal belt” (Lisitzin, 1971) varies and in places, most notably south of South America, it is discontinuous.~~ The Arctic Ocean appears to be dominated by Clay; however, confidence is generally low here, most likely due to the absence of samples north of 75° N (Fig. 1).~~ Overall, map confidence varies between 0.21~~8~~ and 1. It is generally lower in the vicinity of class boundaries and higher in the geographic centre of a class.

The sea-floor lithology map bears a notable resemblance with previously published hand-drawn maps (e.g. Berger, 1974). ~~in that the distribution of Calcareous sediment, Clay and Diatom ooze are very similar. Clear differences were however also noted: Most strikingly, our map does not display a band of Radiolarian ooze in the equatorial Pacific. This was also noted by Dutkiewicz et al. (2015) and is likely related to the sample data, which do not show a predominance of Radiolarian ooze in the equatorial Pacific. Whether the band of Radiolarian ooze is under-represented in the sample data or in fact less prominent than previously depicted remains unresolved.~~The general patterns are very similar, e.g. the distribution of Calcareous sediment, Clay, and Diatom ooze in the major ocean basins. Patterns of Radiolarian ooze in the Indian Ocean resemble those in Thurman (1997: Fig. 5-22). In the Pacific Ocean, Radiolarian ooze is mapped widespread in the vicinity of the equator, although not in the form of a narrow band as frequently depicted in hand-drawn maps (Berger, 1974; Thurman, 1997).

Based on the predicted distribution of lithology classes, Calcareous sediments cover approximately ~~155~~ 121 million km^2 of seabed below 500 m water depth, equivalent to ~~48.1~~36.8 % of the total area (Table 4). Clays are the second most frequent lithology occupying ~~133~~ 102 million km^2 (~~41.3~~31.0 %). Diatom ooze, ~~Siliceous mud and~~ Lithogenous sediment, and Radiolarian ooze account for ~~5~~8.5 %, ~~2.8~~9.5 % and 1~~4~~4.~~5~~2 % of deep-sea floor, respectively~~. Mixed calcareous-siliceous ooze and Radiolarian ooze are the least-frequent lithologies covering just over 1 million km^2 of deep-sea floor.~~.

4.5 Predictor variable importance

The three most important predictor variables were sea-surface maximum salinity, ~~sea-floor maximum temperature and~~ bathymetry, and sea-floor maximum temperature with mean decreases in accuracy ~~between 10 % and 12~~above 5 % (Fig. 8~~7~~). These findings are similar to results from Dutkiewicz et al. (2016), who determined sea-surface salinity, sea-surface temperature and bathymetry as the most important controls on the distribution of deep-sea sediments. Sea-surface minimum silicate was of medium importance (~~8.6~~4.3 % decrease in accuracy), while sea-surface temperature range, sea-surface maximum primary productivity, distance to shore and sea-surface salinity range were of lower importance (<~~5~~3 % decrease in accuracy).

~~4.6 Environmental space~~

~~The environmental space (Fig. 8) is generally sampled adequately, although there is a tendency for an over-representation of shallower water depths (above 3,000 m) and areas closer to land (less than 500 km). Sea-floor maximum temperature, sea-surface temperature range, sea-surface maximum primary productivity and sea-surface maximum salinity are all slightly biased towards higher values. Sea-surface salinity range and sea-surface minimum silicate are the environmental variables that are most closely represented by the samples.~~

5 Limitations of the approach

This study utilised~~s~~ legacy sampling data to make predictions of the spatial distribution of seafloor lithologies in the deep-sea. This is the only viable approach as it is unrealistic to finance and execute a survey programme that samples the global ocean with adequate density within a reasonable timeframe. However, this approach also has some drawbacks:

The presented spatial predictions were based on forming relationships between lithology classes and environmental predictor variables. For such a task, it would be desirable to cover the range of values of each of the predictor variables used in the model (Minasny and McBratney, 2006). Although it was not possible to design a sampling survey, it became nevertheless obvious that the environmental space is reasonably well covered, presumably because of the relatively large number of observations, which was achievable as there was virtually no cost associated with “collecting” the samples. However, it might not always be the case that a large sample dataset leads to adequate coverage of the environmental space. In such a case, it might be desirable to draw a suitable sub-sample that approximates the distribution of the environmental variables.

Data originating from many cruises over long time periods are most likely heterogeneous, which might lead to increased uncertainty in the predictions. Sources of uncertainty might relate to sampling gear type, vintage and timing of sampling, representativeness of subsampling, analytical pre-treatment, inconsistency of classification standards and more (van Heteren and Van Lancker, 2015). However, Dutkiewicz et al. (2015) made efforts to homogenise the data. From a total number of more than 200,000 samples, they selected 14,400 based on strict quality-control criteria. Only surface and near-surface samples that were collected using coring, drilling, or grabbing methods were included. Furthermore, only samples whose descriptions could be verified using original cruise reports, cruise proceedings and core logs were retained. Their classification scheme is deliberately generalised in order to successfully depict the main types of sediments found in the global ocean and to overcome shortcomings of inconsistent, poorly defined and obsolete classification schemes and terminologies (Dutkiewicz et al., 2015).

Additional uncertainty might be introduced through imprecise positioning of the samples, which might lead to incorrect relations between the response variable and the predictor variables. No metadata exist on the positioning accuracy or even the method of determining the position, which might give some clues on the error associated with the recorded positions. However, the chances that this shortcoming leads to significant problems when making associations between target and predictor variables are relatively low, as the chosen model resolution of 10 km is relatively coarse when compared with positioning accuracy.

The initial choice of predictor variables was informed by the current understanding of the controls on deep-sea sedimentation (Dutkiewicz et al., 2016; Seibold and Berger, 1996). Consequently, all selected predictor variables were deemed important (Fig. 2). The three most important predictor variables (Fig. ~~7~~8) are also in good agreement with Dutkiewicz et al. (2016). However, the large errors of ~~c~~o~~m~~mission and especially commission for the rare lithologies Diatom ooze, Lithogenous sediment, ~~Mixed calcareous-siliceous ooze,~~ and Radiolarian ooze ~~and Siliceous mud~~ might indicate that the environmental controls are less well represented for these sediment types. Lithogenous sediment comprises a wide range of grain-sizes (silt, sand, gravel and coarser) and proximity to land might be an insufficient predictor. In fact, distance to shore had the second lowest variable importance (Fig. 8).

Sedimentation rates in the deep-sea typically range on the order of 1 – 100 mm per 1000 yrs (Seibold, 1975). The sample depths in the dataset used here might have ranged from core top to a few dm. The lithologic signal might therefore be integrated over timescales of approximately 100 yrs to a few 100,000 yrs. The model hindcasts to derive the oceanographic predictors typically cover approximately 25 yrs, while bathymetry and distance to coast might be nearly constant since global sea-level rise ceased approximately 6,700 yrs ago (Lambeck et al., 2014). Hence, there likely exists a mismatch between the time intervals, although oceanographic variables might not have changed dramatically over much longer timescales than a few decades.

6 Potential usage

Despite the good agreement with previously published maps and a reasonable overall map accuracy of ≈~~70~~ 60 %, there is large variation in the class-specific error as well as the spatial distribution of map confidence. It is therefore recommended to always consult the information on map confidence along with the map of seafloor lithologies.

The probability surfaces of the seven lithologies might be used as input for spatial prediction and modelling, e.g. marine species distribution modelling on a global scale, which typically lacks information on seafloor sediments, although substrate type is assumed to be an important environmental predictor. Additionally, the presented data layers might be useful for the spatial prediction of sediment properties (e.g. carbonate and organic carbon content).

The categorical map might serve as a resource for education and teaching, provide context for research pertaining to the global seafloor, support marine planning, management and decision-making and underpin the design of marine protected areas globally. Additionally, the provided lithology map might be useful for survey planning, especially in conjunction with confidence information to target areas where a certain lithology is most likely to occur. Conversely, areas of low confidence could be targeted to further improve the accuracy of and confidence in the global map of deep-sea sediments.

7 Data availability

The presented model results (probability surfaces of the seven lithologies, lithology map and associated confidence map) are archived at https://doi.pangaea.de/10.1594/PANGAEA.911692 (Diesing, 2020).

Commented [DM3]: Requires updating

8 Conclusions

Based on a homogenised dataset of seafloor lithology samples (Dutkiewicz et al., 2015) and global environmental predictor variables from Bio-ORACLE (Assis et al., 2018; Tyberghein et al., 2012) and MARSPEC (Sbrocco and Barber, 2013) it was possible to spatially predict the distribution of deep-sea sediments globally. The general understanding about the controls on deep-sea sedimentation helped building a spatial model that gives a good representation of the main lithologies Calcareous sediment, Clay, ~~and~~ Diatom ooze, Lithogenous sediment, and Radiolarian ooze ~~which collectively cover nearly 95% of the mapped seafloor~~. Further improvements should be directed towards the controls on the distribution of rarer lithologies (Diatom ooze, Lithogenous sediment, and ~~Mixed calcareous-siliceous ooze,~~ Radiolarian ooze ~~and Siliceous mud~~).

Author contribution

MD designed the study, developed the model code, executed the analysis, and wrote the manuscript.

Competing interests

The author declares no competing interests.

Acknowledgements

Thanks to Karl Fabian (NGU) for providing valuable feedback. The reviews from Everardo González Ávalos and an anonymous reviewer were extremely helpful and led to significant improvements in the methodology and the manuscript.

References

Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E. A. and De Clerck, O.: Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling, Glob. Ecol. Biogeogr., 27(3), 277–284, doi:10.1111/geb.12693, 2018.

Berger, W. H.: Deep-Sea Sedimentation, in The Geology of Continental Margins, edited by C. A. Burk and C. L. Drake, pp. 213–241, Springer Berlin Heidelberg, Berlin, Heidelberg., 1974.

Breiman, L.: Classification And Regression Trees, Routledge, New York., 1984.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, 2001.

Che Hasan, R., Ierodiaconou, D. and Monk, J.: Evaluation of Four Supervised Learning Methods for Benthic Habitat Mapping Using Backscatter from Multi-Beam Sonar, Remote Sens., 4(11), 3427–3443 [online] Available from: http://www.mdpi.com/2072-4292/4/11/3427, 2012.

Chen, C., Liaw, A. and Breiman, L.: Using Random Forest to Learn Imbalanced Data. [online] Available from: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf, 2004.

Congalton, R. G.: A review of assessing the accuracy of classifications of remotely sensed data, Remote Sens. Environ., 37(1), 35–46 [online] Available from: http://www.sciencedirect.com/science/article/pii/003442579190048B, 1991.

Cortes, C. and Vapnik, V.: Support-vector networks, Mach. Learn., 20(3), 273–297, doi:10.1007/BF00994018, 1995.

Cutler, D., Edwards, T., Beards, K., Cutler, A., Hess, K., Gibson, J. and Lawler, J.: Random Forests for classification in Ecology, Ecology, 88(11), 2783–2792, 2007.

Danovaro, R., Snelgrove, P. V. R. and Tyler, P.: Challenging the paradigms of deep-sea ecology, Trends Ecol. Evol., 29(8), 465–475, doi:10.1016/J.TREE.2014.06.002, 2014.

Diesing, M.: Deep-sea sediments of the global ocean mapped with Random Forest machine learning algorithm, PANGAEA, doi:10.1594/PANGAEA.911692, 2020.

Diesing, M. and Thorsnes, T.: Mapping of Cold-Water Coral Carbonate Mounds Based on Geomorphometric Features: An Object-Based Approach, Geosciences, 8(2), 34, doi:10.3390/geosciences8020034, 2018.

Diesing, M., Kröger, S., Parker, R., Jenkins, C., Mason, C. and Weston, K.: Predicting the standing stock of organic carbon in surface sediments of the North–West European continental shelf, Biogeochemistry, 135(1–2), 183–200, doi:10.1007/s10533-017-0310-4, 2017.

Dutkiewicz, A., Müller, R. D., O'Callaghan, S. and Jónasson, H.: Census of seafloor sediments in the world's ocean, Geology, 43(9), 795–798, doi:10.1130/G36883.1, 2015.

Dutkiewicz, A., O'Callaghan, S. and Müller, R. D.: Controls on the distribution of deep-sea sediments, Geochemistry, Geophys. Geosystems, 17(8), 3075–3098, doi:10.1002/2016GC006428, 2016.

ESRI: World Continents, [online] Available from: https://www.arcgis.com/home/item.html?id=a3cb207855b348a297ab85261743351d (Accessed 24 August 2017), 2010.

GEBCO: The GEBCO_2014 Grid, version 20150318, [online] Available from: http://www.gebco.net (Accessed 24 January 2019), 2015.

Guisan, A. and Zimmermann, N. E.: Predictive habitat distribution models in ecology, Ecol. Modell., 135(2–3), 147–186, doi:10.1016/S0304-3800(00)00354-9, 2000.

Guyon, I. and Elisseeff, A.: An Introduction to Variable and Feature Selection, J. Mach. Learn. Res., 3, 1157–1182, 2003.

Harris, P. T., Macmillan-Lawler, M., Rupp, J. and Baker, E. K.: Geomorphology of the oceans, Mar. Geol., 352, 4–24, doi:10.1016/j.margeo.2014.01.011, 2014.

van Heteren, S. and Van Lancker, V.: Collaborative Seabed-Habitat Mapping: Uncertainty in Sediment Data as an Obstacle in Harmonization, in Collaborative Knowledge in Scientific Research Networks, edited by P. Diviacco, P. Fox, C. Pshenichy, and A. Leadbetter, pp. 154–176, Information Science Reference, Hershey PA, USA., 2015.

Huang, Z., Siwabessy, J., Nichol, S. L. and Brooke, B. P.: Predictive mapping of seabed substrata using high-resolution multibeam sonar data: A case study from a shelf with complex geomorphology, Mar. Geol., 357, 37–52 [online] Available from: http://www.sciencedirect.com/science/article/pii/S0025322714002205, 2014.

James, G., Witten, D., Hastie, T. and Tibshirani, R.: Tree-Based Methods, in An Introduction to Statistical Learning, pp. 303–335, Springer, New York., 2013.

Kuhn, M.: Building Predictive Models in R Using the caret Package, J. Stat. Software; Vol 1, Issue 5 [online] Available from: https://www.jstatsoft.org/v028/i05, 2008.

Kursa, M. and Rudnicki, W.: Feature selection with the Boruta Package, J. Stat. Softw., 36(11), 1–11 [online] Available from: http://www.jstatsoft.org/v36/i11/paper/, 2010.

Lambeck, K., Rouby, H., Purcell, A., Sun, Y. and Sambridge, M.: Sea level and global ice volumes from the Last Glacial Maximum to the Holocene, Proc. Natl. Acad. Sci. , 111(43), 15296–15303, doi:10.1073/pnas.1411762111, 2014.

Liaw, A. and Wiener, M.: Classification and regression by randomForest, R News, 2(3), 18–22, doi:10.1159/000323281, 2002.

Lisitzin, A. P.: Distribution of siliceous microfossils in suspension and in bottom sediments, in The Micropaleontology of Oceans, edited by B. M. Funnell and W. R. Reidel, pp. 173–195, Cambridge University Press, Cambridge., 1971.

Luts, J., Ojeda, F., Plas, R., Van De Moor, B., De Huffel, S. and Van Suykens, J. A. K.: A tutorial on support vector machine-based methods for classification problems in chemometrics, Anal. Chim. Acta, 665(2), 129–145, 2010.

Mastrandrea, M. D., Mach, K. J., Plattner, G. K., Edenhofer, O., Stocker, T. F., Field, C. B., Ebi, K. L. and Matschoss, P. R.: The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups, Clim. Change, 108, 675, doi:10.1007/s10584-011-0178-6, 2011.

Millard, K. and Richardson, M.: On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping, Remote Sens., 7(7), 8489–8515, doi:10.3390/rs70708489, 2015.

Minasny, B. and McBratney, A. B.: A conditioned Latin hypercube method for sampling in the presence of ancillary information, Comput. Geosci., 32(9), 1378–1388, doi:10.1016/J.CAGEO.2005.12.009, 2006.

Misiuk, B., Diesing, M., Aitken, A., Brown, C. J., Edinger, E. N. and Bell, T.: A spatially explicit comparison of quantitative and categorical modelling approaches for mapping seabed sediments using random forest, Geosci., 9(6), 254, doi:10.3390/geosciences9060254, 2019.

Nilsson, R., Peña, J. M., Björkegren, J. and Tegnér, J.: Consistent feature selection for pattern recognition in polynomial time, J. Mach. Learn. Res., 8, 589–612, 2007.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V, Woodcock, C. E. and Wulder, M. a.: Good practices for estimating area and assessing accuracy of land change, Remote Sens. Environ., 148, 42–57, doi:10.1016/j.rse.2014.02.015, 2014.

Prasad, A. M., Iverson, L. R. and Liaw, A.: Newer classification and regression tree techniques: Bagging and random forests for ecological prediction, Ecosystems, 9(2), 181–199, doi:10.1007/s10021-005-0054-1, 2006.

Probst, P.: Performance Measures for Statistical Learning, , 36 [online] Available from: https://cran.r-project.org/web/packages/measures/measures.pdf, 2018.

R Core Team: R: A Language and Environment for Statistical Computing, 2018.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F. and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography (Cop.)., 40(8), 913–929, doi:10.1111/ecog.02881, 2017.

Sbrocco, E. J. and Barber, P. H.: MARSPEC: ocean climate layers for marine spatial ecology, Ecology, 94(4), 979, doi:10.1890/12-1358.1, 2013.

Seibold, E.: Der Meeresboden Forschungsstand und Zukunftsaufgaben, Naturwissenschaften, 62(7), 321–330,

Formatted: Norwegian (Bokmål)

doi:10.1007/BF00608892, 1975.

Seibold, E. and Berger, W. H.: The sea floor. An introduction to marine geology, 3rd editio., Springer, Berlin., 1996.

Snelgrove, P. V. R., Soetaert, K., Solan, M., Thrush, S., Wei, C.-L., Danovaro, R., Fulweiler, R. W., Kitazato, H., Ingole, B., Norkko, A., Parkes, R. J. and Volkenborn, N.: Global Carbon Cycling on a Heterogeneous Seafloor, Trends Ecol. Evol., 33(2), 96–105, doi:10.1016/J.TREE.2017.11.004, 2018.

Stehman, S. V. and Foody, G. M.: Key issues in rigorous accuracy assessment of land cover products, Remote Sens. Environ., 231, 111199, doi:10.1016/J.RSE.2019.05.018, 2019.

Stevens Jr, D. L. and Olsen, A. R.: Variance estimation for spatially balanced samples of environmental resources, Environmetrics, 14(6), 593–610, doi:10.1002/env.606, 2003.

Story, M. and Congalton, R. G.: Accuracy Assessment: A User's Perspective, Photogramm. Eng. Remote Sensing, 52, 397–399, 1986.

Strobl, C. and Zeileis, A.: Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance, Munich. [online] Available from: https://epub.ub.uni-muenchen.de/2111/1/techreport.pdf, 2008.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution, BMC Bioinformatics, 8(1), 25, doi:10.1186/1471-2105-8-25, 2007.

Thurman, H. V.: Introductory Oceanography, 8th ed., Prentice-Hall, Upper Saddle River, New Jersey., 1997.

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F. and De Clerck, O.: Bio-ORACLE: a global environmental dataset for marine species distribution modelling, Glob. Ecol. Biogeogr., 21(2), 272–281, doi:10.1111/j.1466-8238.2011.00656.x, 2012.

Valavi, R., Elith, J., Lahoz-Monfort, J. J. and Guillera-Arroita, G.: BLOCKCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models, Methods Ecol. Evol., doi:https://doi.org/10.1111/2041-210X.13107, 2018.

Table 1: Environmental predictor variables tested in this study

Environmental variable	Statistics	Unit	Source
Bathymetry	mean	m	Sbrocco and Barber (2013)
Distance to shore	mean	km	Sbrocco and Barber (2013)
Sea-surface temperature	mean, min, max, range	°C	Assis et al. (2018)
Sea-surface salinity	mean, min, max, range	PSS	Assis et al. (2018)
Sea-surface dissolved oxygen	mean, min, max, range	mol m^{-3}	Assis et al. (2018)
Sea-surface primary productivity	mean, min, max, range	g m^{-3} day^{-1}	Assis et al. (2018)
Sea-surface iron concentration	mean, min, max, range	µmol m^{-3}	Assis et al. (2018)
Sea-surface nitrate concentration	mean, min, max, range	mol m^{-3}	Assis et al. (2018)
Sea-surface phosphate concentration	mean, min, max, range	mol m^{-3}	Assis et al. (2018)
Sea-surface silicate concentration	mean, min, max, range	mol m^{-3}	Assis et al. (2018)
Sea-floor temperature	mean, min, max, range	°C	Assis et al. (2018)

455

Table 2: Seafloor lithology classes used in this study, their abbreviations, their relationships to classes in Dutkiewicz et al. (2015) and the number and percentage of samples~~ in the training and test datasets~~. Not included are Ash and volcanic sand/gravel, Mixed calcareous-siliceous ooze, Siliceous mud, Sponge spicules and Shells and coral fragments of the original classification.

~~**Lithology class**~~	~~**Abbreviation**~~	~~**Relation to Dutkiewicz et al. (2015)**~~	~~**Training dataset**~~
~~**Calcareous sediment**~~	~~Calc.Sed~~	~~Calcareous ooze~~ ~~Fine-grained calcareous sediment~~	~~3029 (44.8 %)~~
~~**Clay**~~	~~Clay~~	~~Clay~~	~~2219 (32.8 %)~~
~~**Diatom ooze**~~	~~Dia.Ooze~~	~~Diatom ooze~~	~~361 (5.3 %)~~
~~**Lithogenous sediment**~~	~~Lith.Sed~~	~~Gravel and coarser~~ ~~Sand~~ ~~Silt~~	~~438 (6.5 %)~~
~~**Mixed calcareous-siliceous ooze**~~	~~Mxd.Ooze~~	~~Mixed calcareous-siliceous ooze~~	~~123 (1.8 %)~~
~~**Radiolarian ooze**~~	~~Rad.Ooze~~	~~Radiolarian ooze~~	~~60 (0.9 %)~~
~~**Siliceous mud**~~	~~Sil.Mud~~	~~Siliceous mud~~	~~539 (8.0 %)~~
Lithology class	Abbreviation	Relation to Dutkiewicz et al. (2015)	No. of observation
Calcareous sediment	Calc.Sed	Calcareous ooze Fine-grained calcareous sediment	5251 (50.3 %)
Clay	Clay	Clay	3714 (35.6 %)
Diatom ooze	Dia.Ooze	Diatom ooze	623 (6.0 %)
Lithogenous sediment	Lith.Sed	Gravel and coarser Sand Silt	751 (7.2 %)
Radiolarian ooze	Rad.Ooze	Radiolarian ooze	99 (0.9 %)

Formatted Table

Formatted: Left

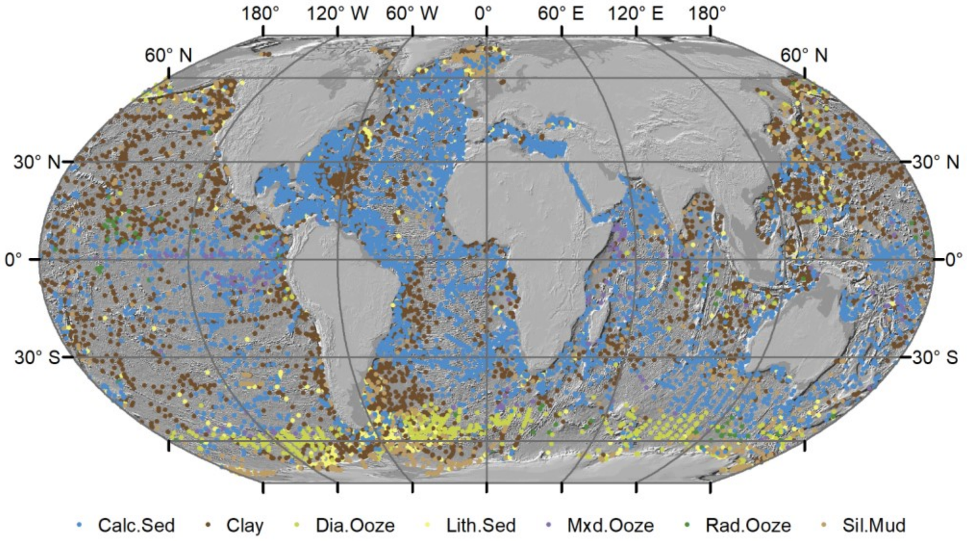
Table 3: Confusion matrix. Observed (reference) classes are in columns, predicted classes in rows.

	~~**Calc.Sed**~~	~~**Clay**~~	~~**Dia.Ooze**~~	~~**Lith.Sed**~~	~~**Mxd.Ooze**~~	~~**Rad.Ooze**~~	~~**Sil.Mud**~~	~~**Row total**~~	~~**Error of commission**~~
~~**Calc.Sed**~~	~~1316~~	~~229~~	~~11~~	~~64~~	~~47~~	~~5~~	~~52~~	~~1724~~	~~0.237~~
~~**Clay**~~	~~136~~	~~789~~	~~41~~	~~71~~	~~5~~	~~11~~	~~87~~	~~1140~~	~~0.308~~
~~**Dia.Ooze**~~	~~17~~	~~27~~	~~106~~	~~19~~	~~3~~	~~7~~	~~17~~	~~196~~	~~0.459~~
~~**Lith.Sed**~~	~~12~~	~~27~~	~~10~~	~~36~~	~~0~~	~~0~~	~~18~~	~~103~~	~~0.650~~
~~**Mxd.Ooze**~~	~~10~~	~~3~~	~~4~~	~~0~~	~~6~~	~~0~~	~~1~~	~~24~~	~~0.750~~
~~**Rad.Ooze**~~	~~0~~	~~1~~	~~0~~	~~0~~	~~1~~	~~3~~	~~1~~	~~6~~	~~0.500~~
~~**Sil.Mud**~~	~~21~~	~~32~~	~~8~~	~~28~~	~~0~~	~~4~~	~~93~~	~~186~~	~~0.500~~
~~**Column total**~~	~~1512~~	~~1108~~	~~180~~	~~218~~	~~62~~	~~30~~	~~269~~		
~~**Error of omission**~~	~~0.130~~	~~0.288~~	~~0.411~~	~~0.835~~	~~0.903~~	~~0.900~~	~~0.654~~		

	Calc.Sed	**Clay**	**Dia.Ooze**	**Lith.Sed**	**Rad.Ooze**	**Row total**	**Error of commission**
Calc.Sed	3735	660	10	139	10	4554	0.180
Clay	760	2001	44	151	17	2973	0.327
Dia.Ooze	234	215	299	294	37	1079	0.723
Lith.Sed	277	456	111	128	3	975	0.869
Rad.Ooze	245	382	159	39	32	857	0.963
Column total	5251	3714	623	751	99		
Error of omission	0.289	0.461	0.520	0.830	0.677		

Table 4: Breakdown of areal coverage by lithology types in the global ocean below 500 m water depth.

~~Lithology~~	~~Number of pixels~~	~~Area (10^6 km^2)~~	~~Area (%)~~
~~**Calcareous sediment**~~	~~1,546,723~~	~~154.672~~	~~48.1~~
~~**Clay**~~	~~1,328,012~~	~~132.801~~	~~41.3~~
~~**Diatom ooze**~~	~~177,545~~	~~17.755~~	~~5.5~~
~~**Lithogenous sediment**~~	~~47,469~~	~~4.747~~	~~1.5~~
~~**Mixed calcareous-siliceous ooze**~~	~~12,014~~	~~1.201~~	~~0.4~~
~~**Radiolarian ooze**~~	~~10,342~~	~~1.034~~	~~0.3~~
~~**Siliceous mud**~~	~~91,028~~	~~9.103~~	~~2.8~~
~~**Sum**~~	~~3,213,133~~	~~321.313~~	~~100~~
Lithology	Number of pixels	Area (10^6 km^2)	Area (%)
Calcareous sediment	1,211,063	121.106	36.80
Clay	1,019,160	101.916	30.97
Diatom ooze	279,955	27.996	8.51
Lithogenous sediment	312,668	31.267	9.50
Radiolarian ooze	467,775	46.778	14.22
Sum	3,290,621	329.062	100


180°
120° W
60° W
0°
60° E
120° E
180°
60° N
60° N
30° N
30° N
0°
0°
30° S
30° S
Calc.Sed
Clay
Dia.Ooze
Lith.Sed
Mxd.Ooze
Rad.Ooze
Sil.Mud

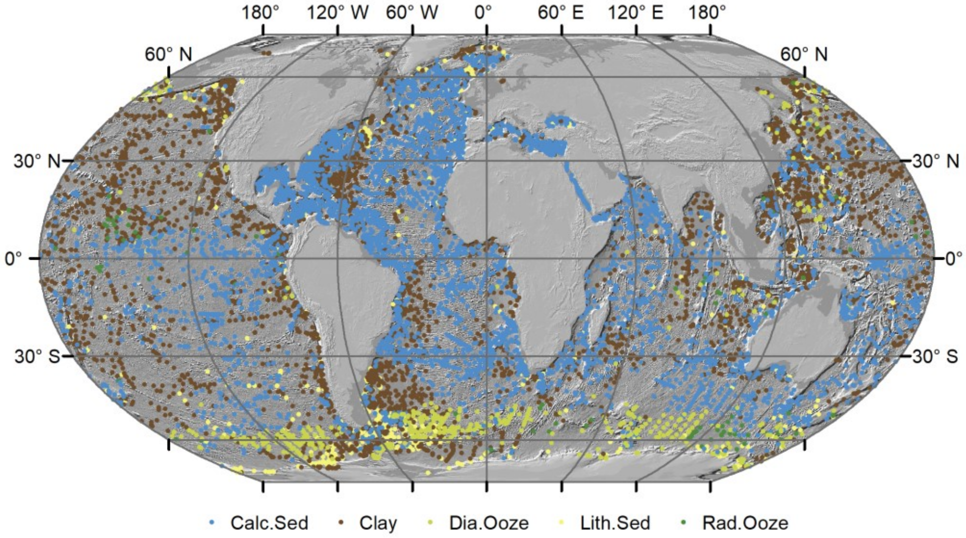
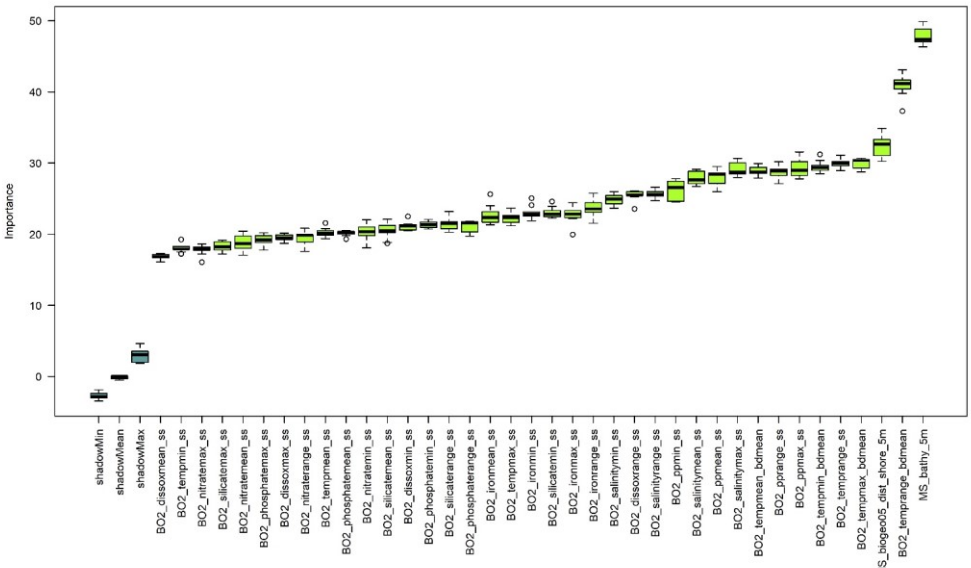


Figure 1: Locations of samples used in this study based on data from Dutkiewicz et al. (2015). Land masses are derived from ESRI (2010)**. Hillshade topography is derived from GEBCO** (2015)**.**


Importance
50
40
30
20
10
0
shadowMin
shadowMean
shadowMax
BO2_dissoxmean_ss
BO2_tempmin_ss
BO2_nitratemax_ss
BO2_silicatemax_ss
BO2_nitratemean_ss
BO2_phosphatemax_ss
BO2_dissoxmax_ss
BO2_nitraterange_ss
BO2_tempmean_ss
BO2_phosphatemean_ss
BO2_nitratemin_ss
BO2_silicatemean_ss
BO2_dissoxmin_ss
BO2_phosphatemin_ss
BO2_silicaterange_ss
BO2_phosphaterange_ss
BO2_ironmean_ss
BO2_tempmax_ss
BO2_ironmin_ss
BO2_silicatemin_ss
BO2_ironmax_ss
BO2_ironrange_ss
BO2_salinitymin_ss
BO2_dissoxrange_ss
BO2_salinityrange_ss
BO2_ppmin_ss
BO2_salinitymean_ss
BO2_ppmean_ss
BO2_salinitymax_ss
BO2_tempmean_bdmean
BO2_pprange_ss
BO2_ppmax_ss
BO2_tempmin_bdmean
BO2_temprange_ss
BO2_tempmax_bdmean
MS_biogeo05_dist_shore_5m
BO2_temprange_bdmean
MS_bathy_5m

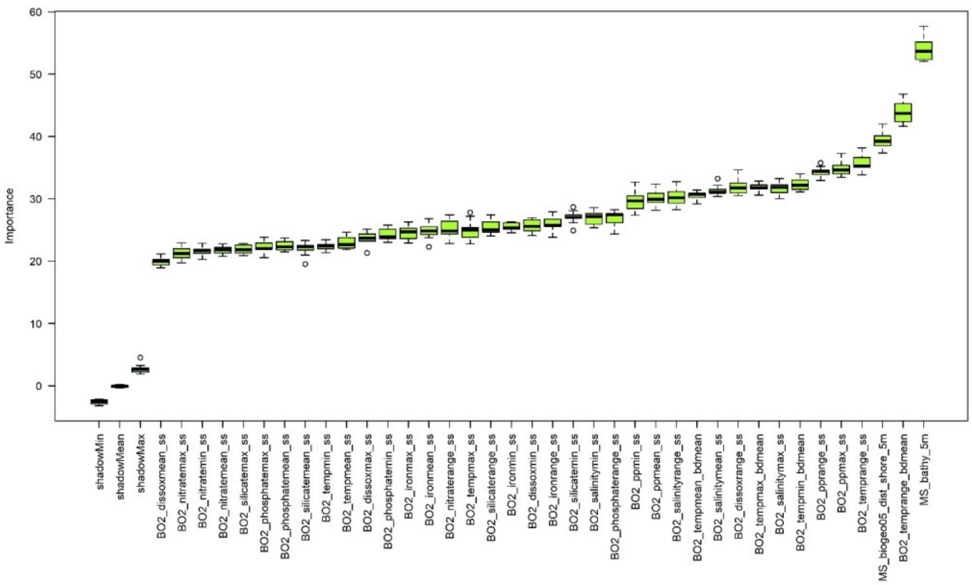


Figure 2: Results of the Boruta variable selection process. All environmental predictor variables had an importance significantly higher than the shadow variables (shadowMin, ShadowMean and shadowMax).

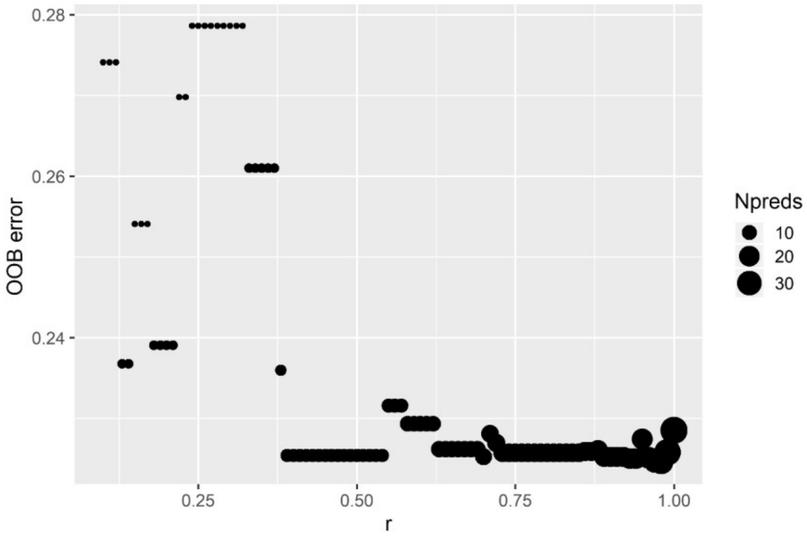
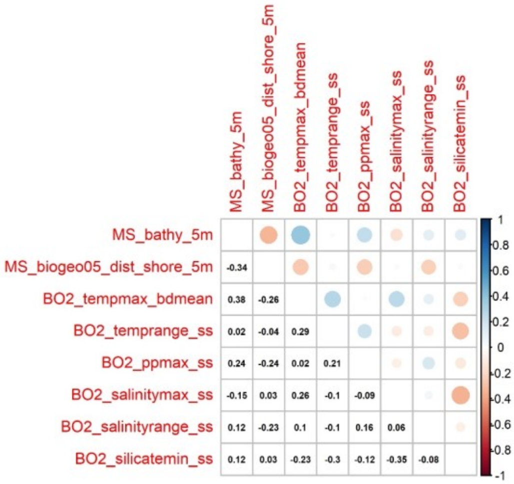


Figure 3: Influence of r-value on the out of bag error of a random forest model with default parameters. The size of the circles indicates the number of selected predictor variables (Npreds).


MS_bathy_5m
MS_biogeo05_dist_shore_5m
BO2_tempmax_bdmean
BO2_temprange_ss
BO2_ppmax_ss
BO2_salinitymax_ss
BO2_salinityrange_ss
BO2_silicatemin_ss
MS_bathy_5m
MS_biogeo05_dist_shore_5m -0.34
BO2_tempmax_bdmean 0.38 -0.26
BO2_temprange_ss 0.02 -0.04 0.29
BO2_ppmax_ss 0.24 -0.24 0.02 0.21
BO2_salinitymax_ss -0.15 0.03 0.26 -0.1 -0.09
BO2_salinityrange_ss 0.12 -0.23 0.1 -0.1 0.16 0.06
BO2_silicatemin_ss 0.12 0.03 -0.23 -0.3 -0.12 -0.35 -0.08
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

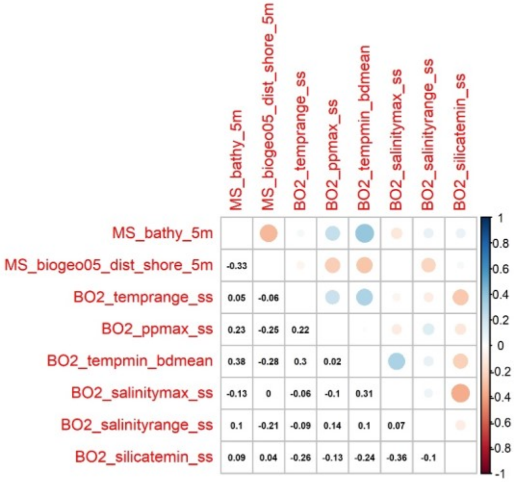
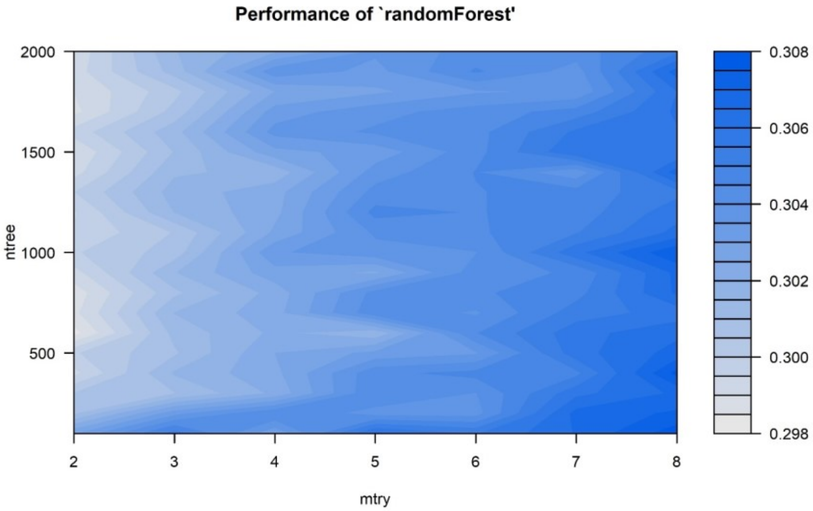

MS_bathy_5m
MS_biogeo05_dist_shore_5m
BO2_temprange_ss
BO2_ppmax_ss
BO2_tempmin_bdmean
BO2_salinitymax_ss
BO2_salinityrange_ss
BO2_silicatemin_ss
MS_bathy_5m
MS_biogeo05_dist_shore_5m -0.33
BO2_temprange_ss 0.05 -0.06
BO2_ppmax_ss 0.23 -0.25 0.22
BO2_tempmin_bdmean 0.38 -0.28 0.3 0.02
BO2_salinitymax_ss -0.13 0 -0.06 -0.1 0.31
BO2_salinityrange_ss 0.1 -0.21 -0.09 0.14 0.1 0.07
BO2_silicatemin_ss 0.09 0.04 -0.26 -0.13 -0.24 -0.36 -0.1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure ~~3~~4: Correlation plot showing the correlation coefficients of the selected predictor variables.



~~**Figure 4: Results of the model parameter tuning. The best performance (lowest error) is achieved with n_{tree} = 600 and m_{try} = 2. Differences in error are however less than 1 % between the best and worst-performing parameter combination.**~~

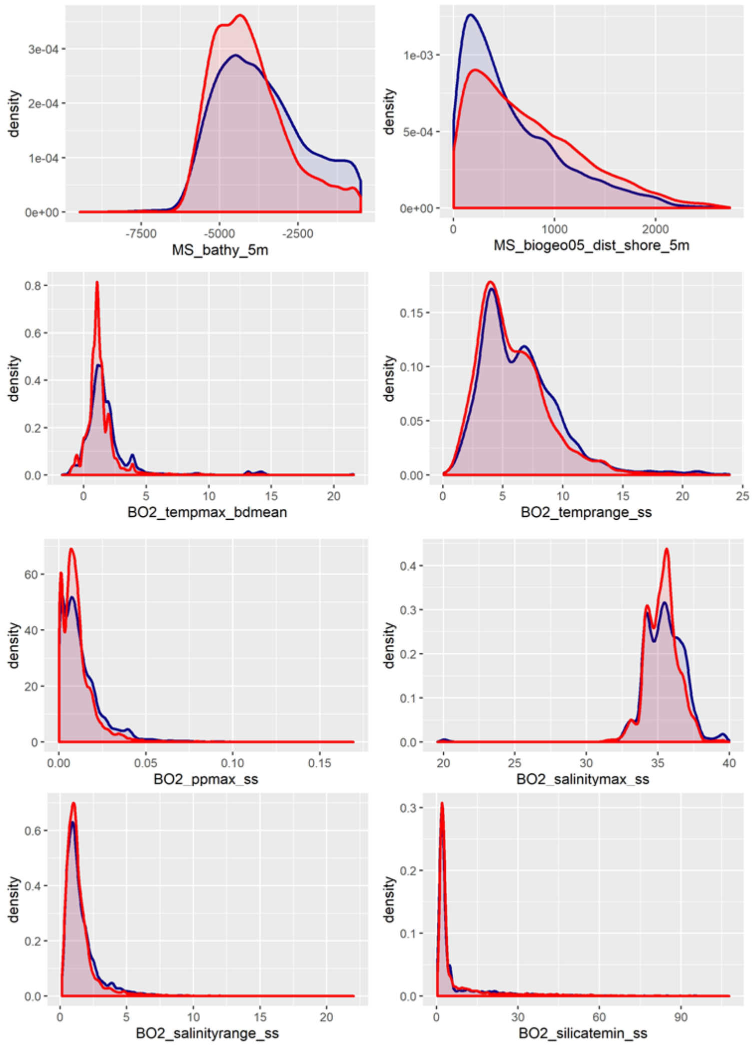
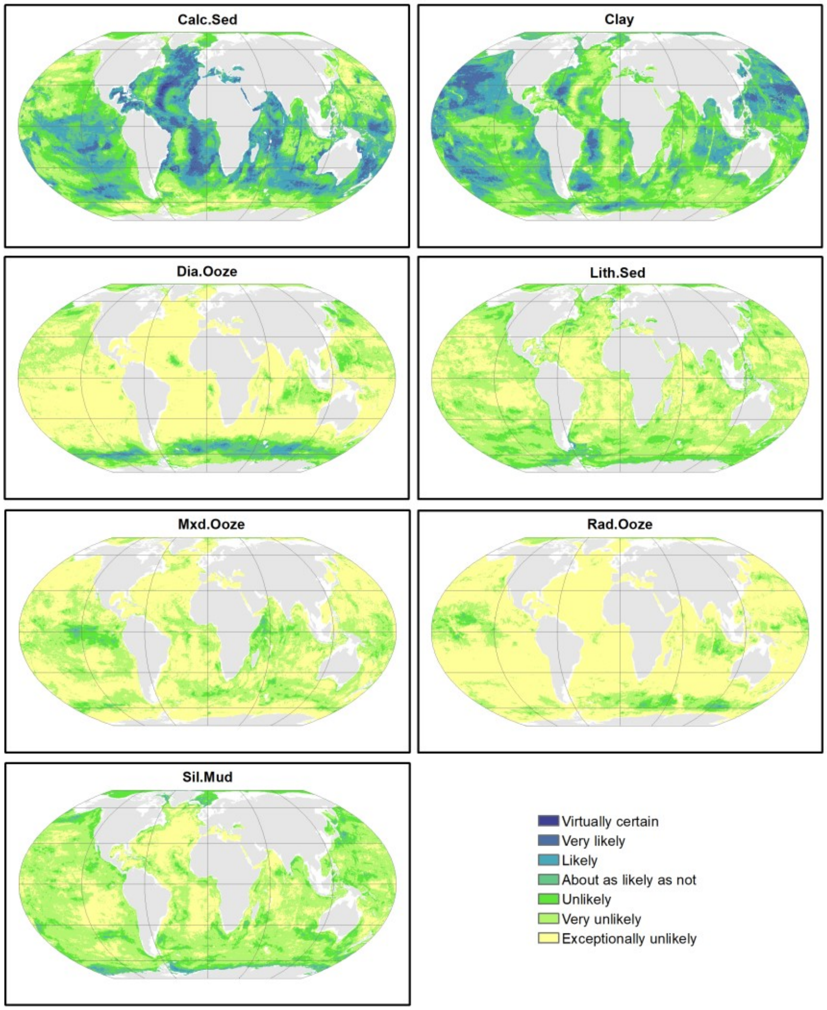


Figure ~~8~~5: A visual check to what extent the samples cover the environmental space. Blue: Samples; Red: Environmental data.


Calc.Sed
Clay
Dia.Ooze
Lith.Sed
Mxd.Ooze
Rad.Ooze
Sil.Mud
Virtually certain
Very likely
Likely
About as likely as not
Unlikely
Very unlikely
Exceptionally unlikely

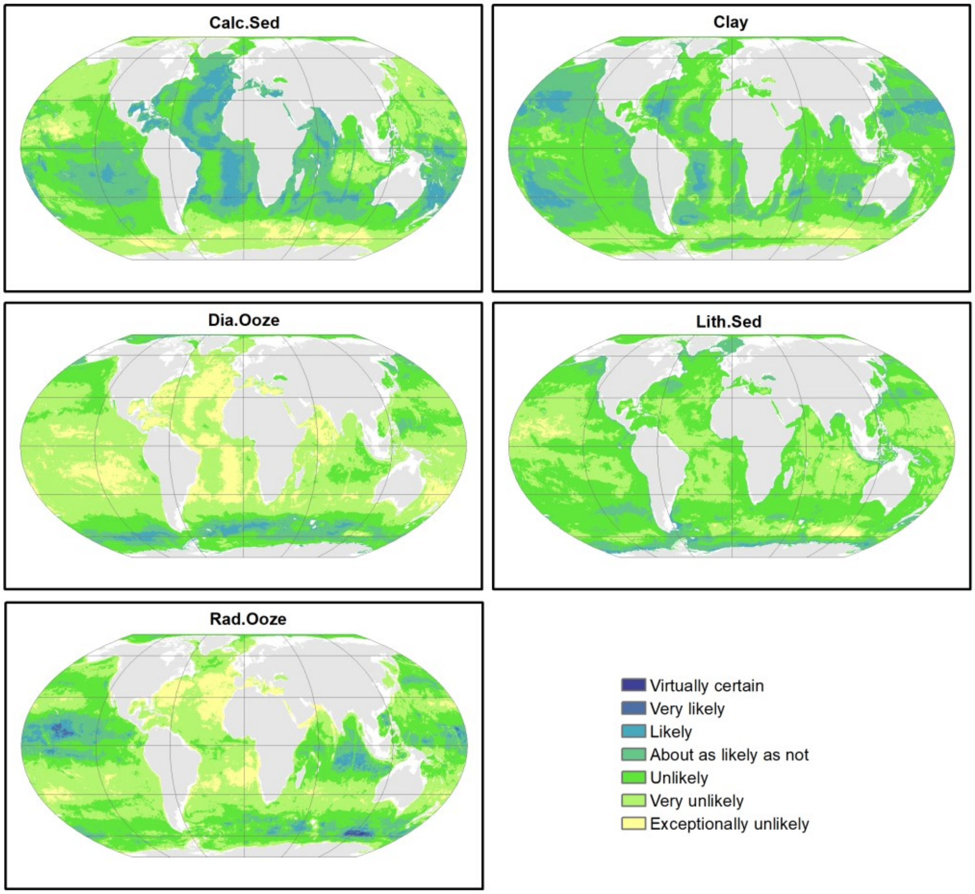
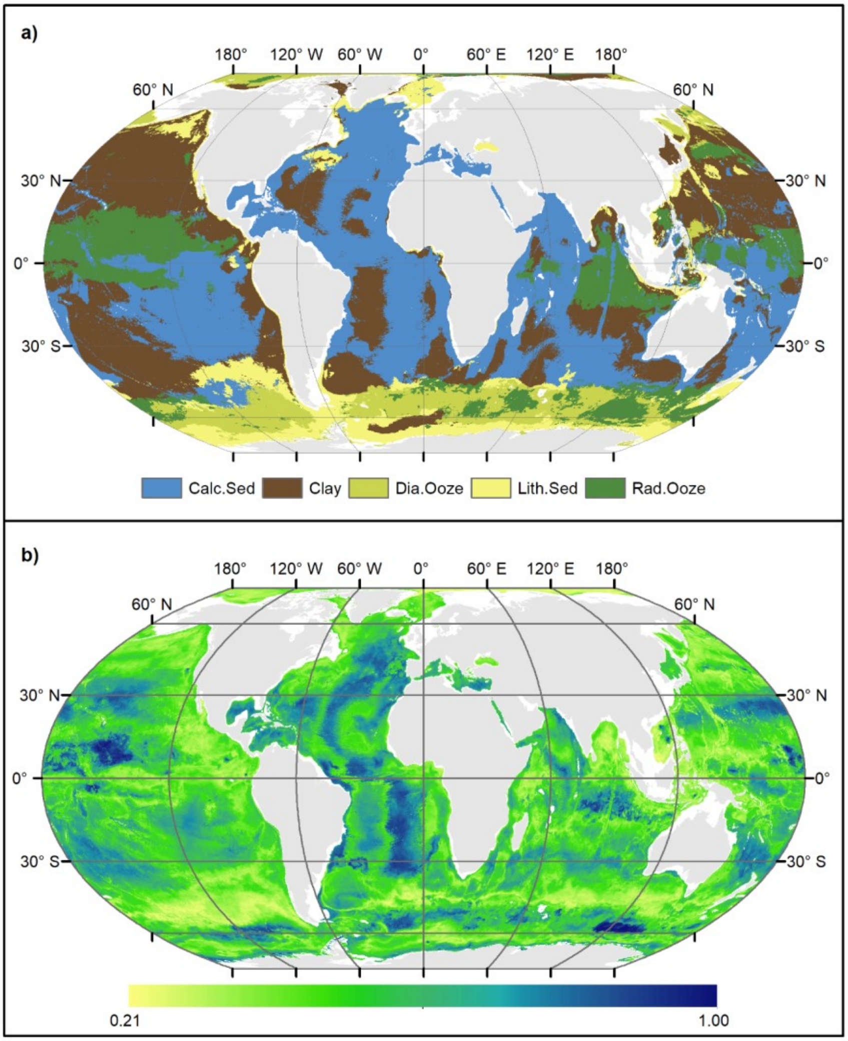


Figure ~~5~~6: Probability surfaces of the ~~seven~~ five predicted lithologies. The verbal likelihood scale is based on Mastrandrea et al. (2011). Land masses are derived from ESRI (2010).


a)
180°
120° W
60° W
0°
60° E
120° E
180°
60° N
60° N
30° N
30° N
0°
0°
30° S
30° S
Calc.Sed
Clay
Dia.Ooze
Lith.Sed
Rad.Ooze
b)
180°
120° W
60° W
0°
60° E
120° E
180°
60° N
60° N
30° N
30° N
0°
0°
30° S
30° S
0.21
1.00

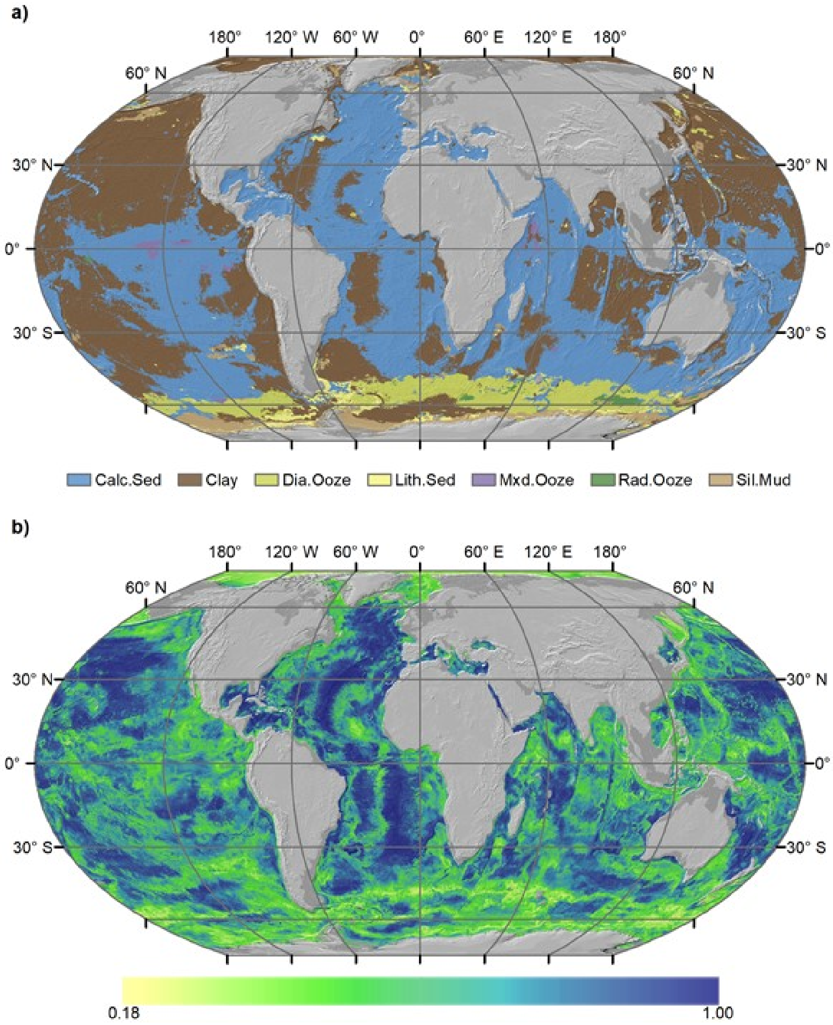
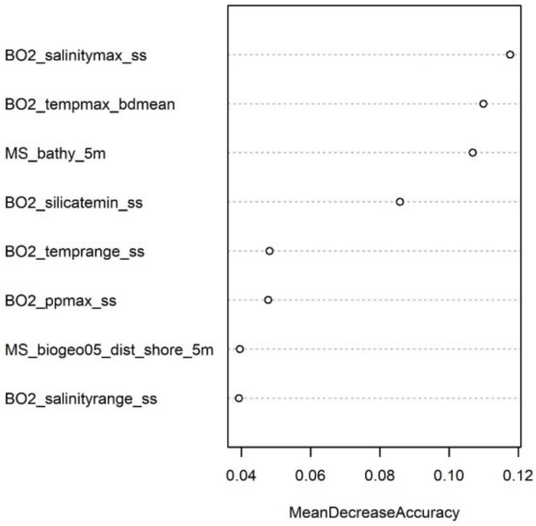


Figure ~~6~~7: a) Predicted lithology classes and b) associated confidence in the predictions. Land masses are derived from ESRI (2010). ~~**Hillshade topography is derived from (GEBCO, 2015).**~~


BO2_salinitymax_ss
BO2_tempmax_bdmean
MS_bathy_5m
BO2_silicatemin_ss
BO2_temprange_ss
BO2_ppmax_ss
MS_biogeo05_dist_shore_5m
BO2_salinityrange_ss
0.04
0.06
0.08
0.10
0.12
MeanDecreaseAccuracy

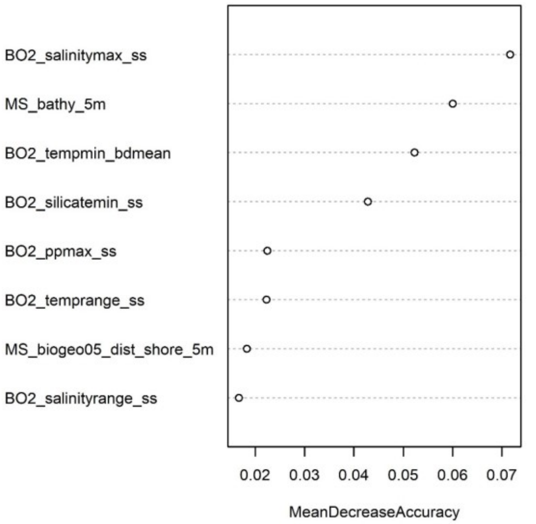
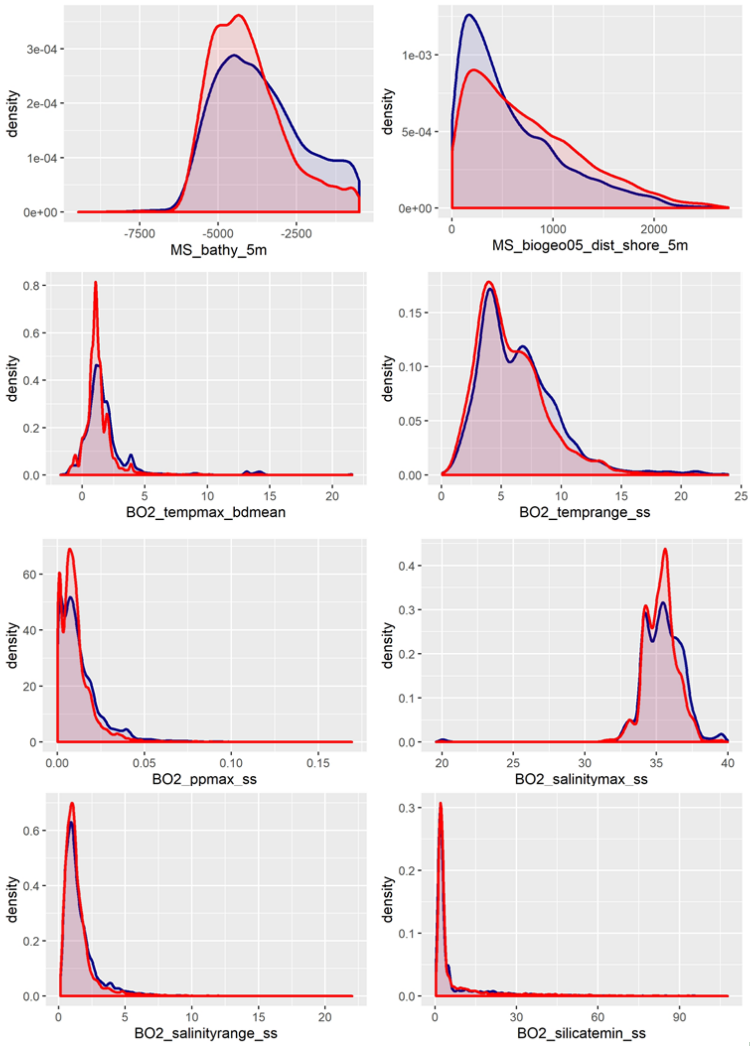
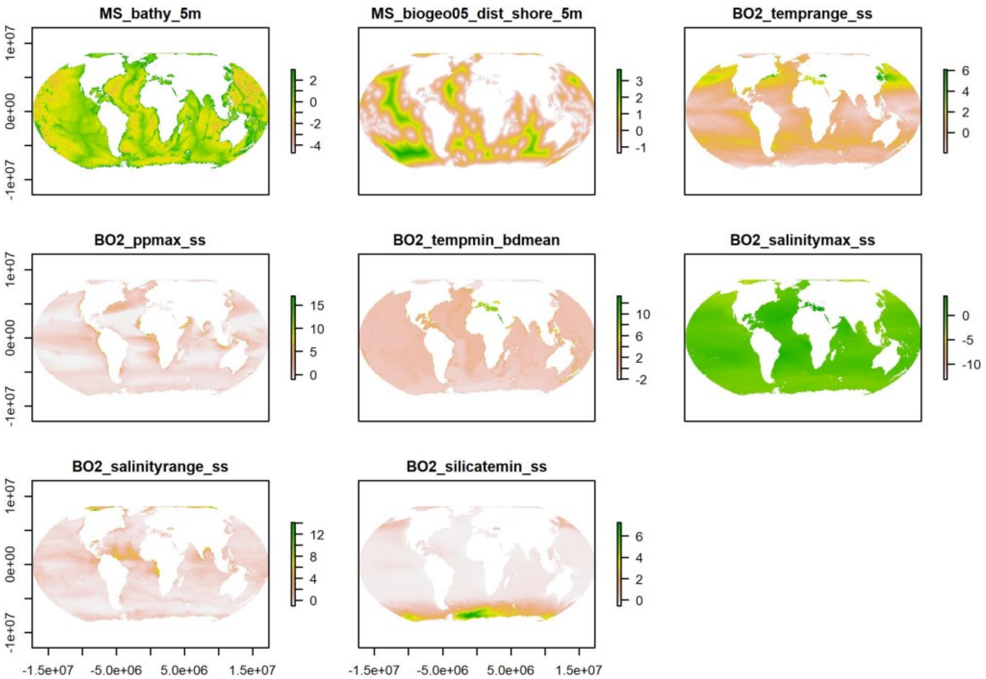

BO2_salinitymax_ss
MS_bathy_5m
BO2_tempmin_bdmean
BO2_silicatemin_ss
BO2_ppmax_ss
BO2_temprange_ss
MS_biogeo05_dist_shore_5m
BO2_salinityrange_ss
0.02
0.03
0.04
0.05
0.06
0.07
MeanDecreaseAccuracy

Figure ~~7~~8: Random Forest variable importance.



~~Figure 8: A visual check to what extent the samples cover the environmental space. Blue: Samples; Red: Environmental data.~~


MS_bathy_5m
2
0
-2
-4
MS_biogeo05_dist_shore_5m
3
2
1
0
-1
BO2_temprange_ss
6
4
2
0
BO2_ppmax_ss
15
10
5
0
BO2_tempmin_bdmean
10
6
2
-2
BO2_salinitymax_ss
0
-5
-10
BO2_salinityrange_ss
12
8
4
0
BO2_silicatemin_ss
6
4
2
0
1e+07
0e+00
-1e+07
-1.5e+07
-5.0e+06
5.0e+06
1.5e+07

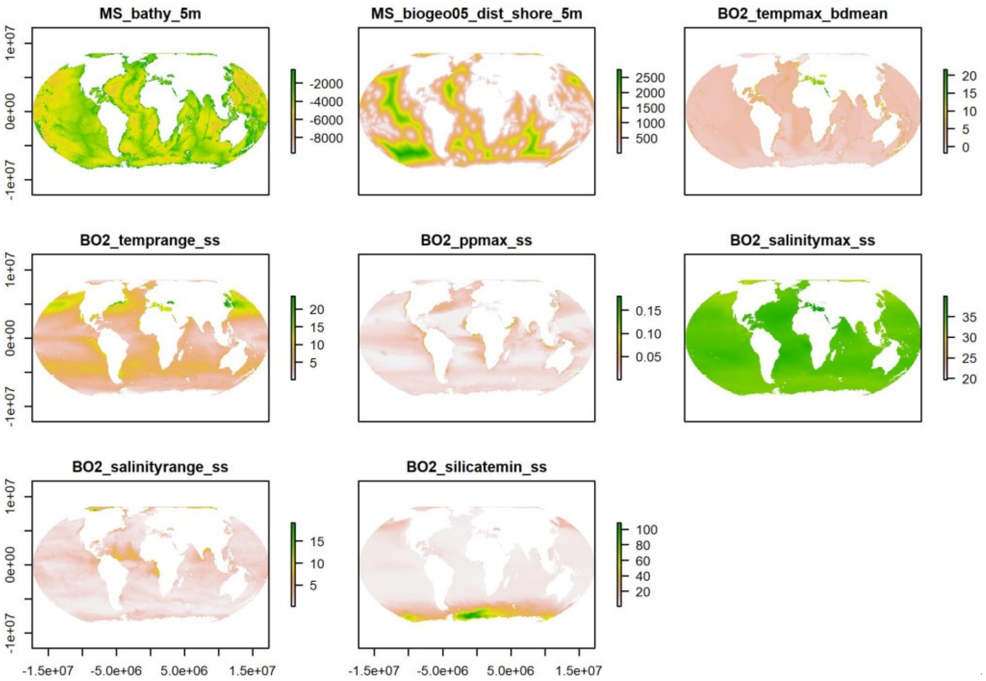
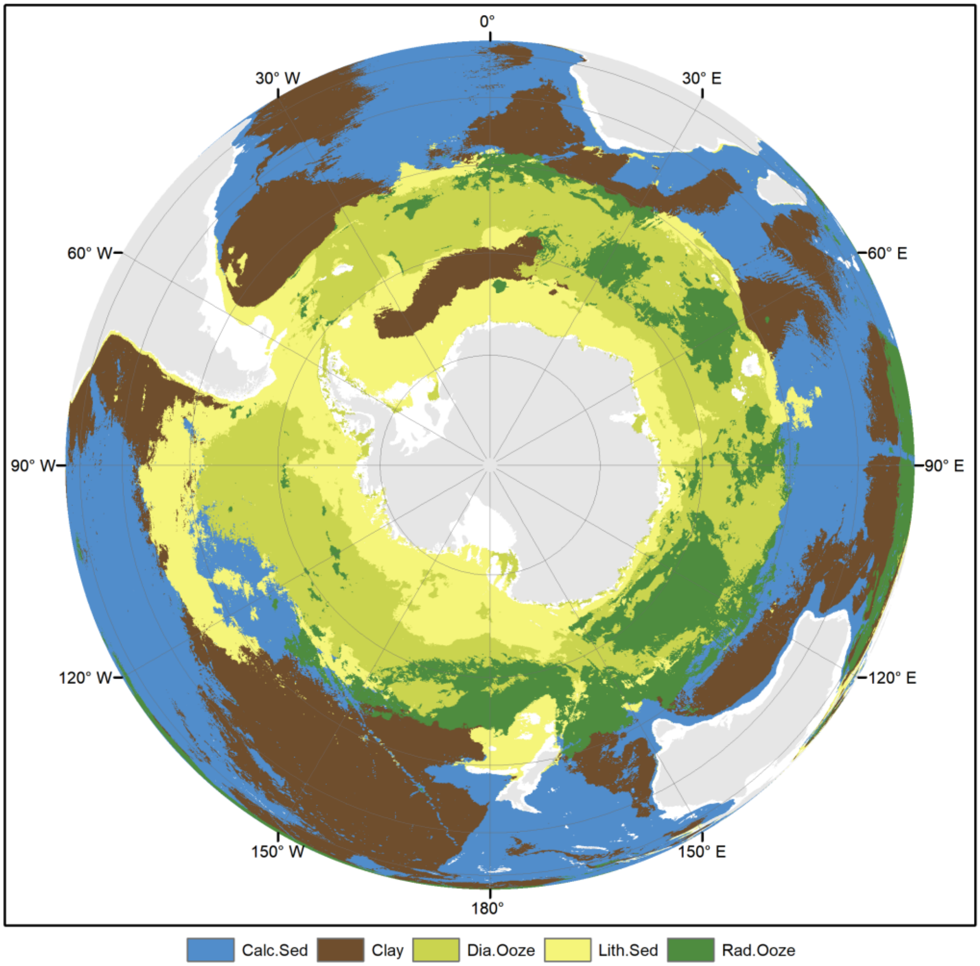


Figure A1: Plots of the selected scaled predictor variables. ~~Units as in Table 1.~~


0°
30° W
30° E
60° W
60° E
90° W
90° E
120° W
120° E
150° W
150° E
180°
Calc.Sed
Clay
Dia.Ooze
Lith.Sed
Rad.Ooze

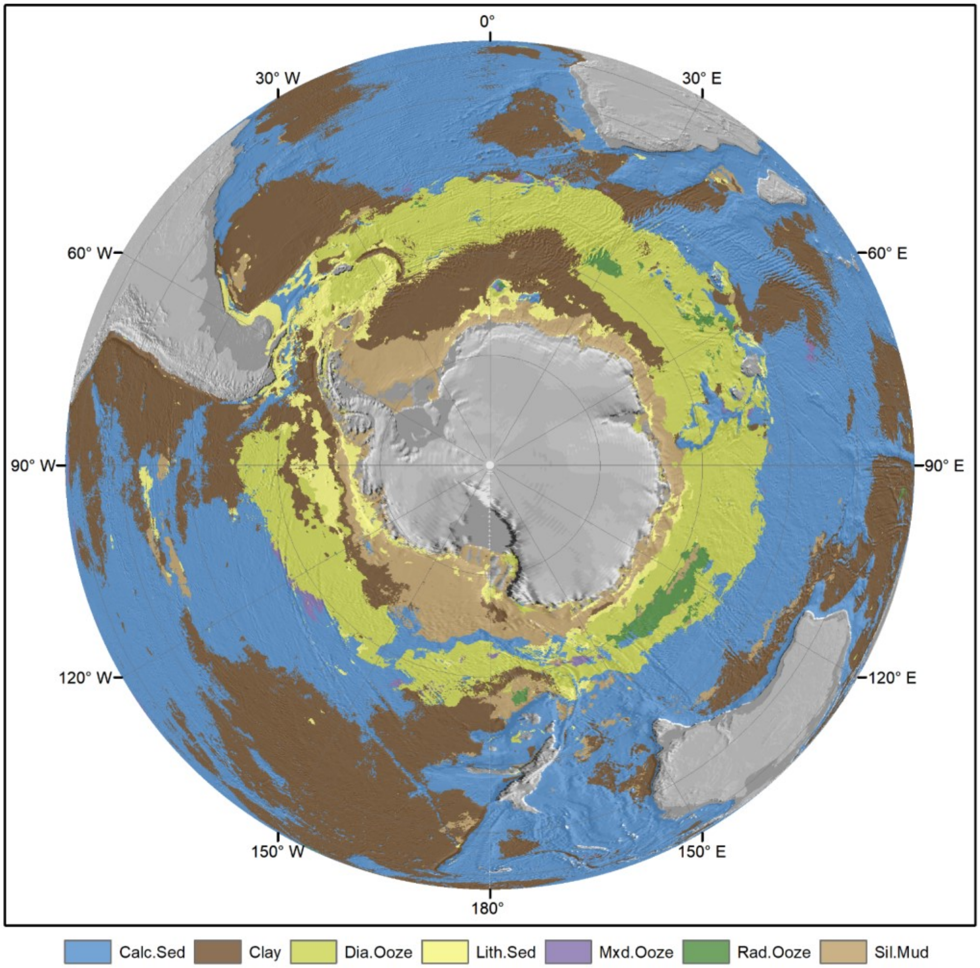


Figure A2: Predicted lithology classes in the Southern Ocean. Land masses are derived from ESRI (2010). ~~**Hillshade topography is derived from** (2015).~~