Received: 16 July 2019 | Accepted: 12 February 2020

DOI: 10.1111/1365-2664.13621

Check for updates

RESEARCH ARTICLE

Journal of Applied Ecology BRITISH ECOLOGICAL SOCIETY

Predicting meiofauna abundance to define preservation and impact zones in a deep-sea mining context using random forest modelling

Katja Uhlenkott[1,2] | Annemiek Vink[3] | Thomas Kuhn[3] | Pedro Martínez Arbizu[1,2]

[1]German Centre for Marine Biodiversity Research (DZMB), Senckenberg am Meer, Wilhelmshaven, Germany

[2]Marine Biodiversity Research, Institute for Biology and Environmental Sciences, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

[3]Bundesanstalt für Geowissenschaften & Rohstoffe, Hannover, Germany

Correspondence
Katja Uhlenkott
Email: katja.uhlenkott@senckenberg.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 03F0707E and 03F0812E

Handling Editor: Melinda Coleman

Abstract

1. There is a strong economic interest in commercial deep-sea mining of polymetallic nodules and therefore a need to define suitable preservation zones in the abyssal plain of the Clarion Clipperton Fracture Zone (CCZ). However, besides ship-based multibeam data, only sparse continuous environmental information is available over large geographic scales.
2. We test the potential of modelling meiofauna abundance and diversity on high taxonomic level on large geographic scale using a random forest approach. Ship-based multibeam bathymetry and backscatter signal are the only sources for 11 predictor variables, as well as the modelled abundance of polymetallic nodules on the seafloor. Continuous meiofauna predictions have been combined with all available environmental variables and classified into classes representing abyssal habitats using *k*-means clustering.
3. Results show that ship-based, multibeam-derived predictors can be used to calculate predictive models for meiofauna distribution on a large geographic scale. Predicted distribution varies between the different meiofauna response variables.
4. To evaluate predictions, random forest regressions were additionally computed with 1,000 replicates, integrating varying numbers of sampling positions and parallel samples per site. Higher numbers of parallel samples are especially useful to smoothen the influence of the remarkable variability of meiofauna distribution on a small scale. However, a high number of sampling positions is even more important, integrating a greater amount of natural variability of environmental conditions into the model.
5. *Synthesis and applications*. Polymetallic nodule exploration contractors are required to define potential mining and preservation zones within their licence area. The biodiversity and the environment of preservation zones should be representative of the sites that will be impacted by mining. Our predicted distributions of meiofauna and the derived habitat maps are an essential first step to enable the identification of areas with similar ecological conditions. In this way, it is possible

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2020 The Authors. *Journal of Applied Ecology* published by John Wiley & Sons Ltd on behalf of British Ecological Society

to define preservation zones not only based on expert opinion and environmental proxies but also integrating evidence from the distribution of benthic communities.

KEYWORDS

Clarion Clipperton Fracture Zone, deep sea, distribution model, habitat mapping, meiofauna, nodule mining, preservation zone, random forest

1 | INTRODUCTION

Although the deep sea represents the largest ecosystem on Earth, it remains largely unexplored (Ramirez-Llodra et al., 2010). Nevertheless, increasing demands for minerals and metals have greatly enhanced the interest in potential mining of deep-sea resources (Miller, Thompson, Johnston, & Santillo, 2018). In the Clarion Clipperton Fracture Zone (CCZ) in the north-eastern equatorial Pacific Ocean, such resources are widespread in the form of polymetallic nodules (Wedding et al., 2015), which are a potential source of copper, cobalt, nickel and manganese (Wegorzewski & Kuhn, 2014). However, mining such resources will inherently have a severe impact on environment and fauna, the scale of which is difficult to ascertain despite the considerable number of disturbance experiments that have been carried out so far (Jones et al., 2017). A common method to protect the marine environment from human impacts is through the establishment of marine-protected areas. In the CCZ, such areas have been established by the International Seabed Authority (ISA) based on environmental proxies and expert opinion (Wedding et al., 2013, 2015). However, these so-called Areas of Particular Environmental Interest (APEI) are mainly positioned outside the core of the CCZ (Kaiser, Smith, & Martínez Arbizu, 2017). Therefore, it will be crucial to establish additional preservation zones on smaller spatial scales within the contractor areas in order to protect the environment from serious harm, promote recovery and provide baseline conditions against which the impacts of mining in an adjacent, ecologically similar area can be assessed (Vanreusel, Hilario, Ribeiro, Menot, & Martínez Arbizu, 2016).

To develop a sound mining and environmental management plan, it is essential that contractors define sites that are prospective in terms of mining but also potential preservation zones that shall not be impacted by mining but contain benthic communities representative for the communities of mining sites. To do so, it is important to investigate the continuous spatial distribution of taxa and community characteristics across the contractor areas and combine it with the available environmental data. But how can a continuous mapping of benthic communities be warranted within a license area that is 75,000 km^2 large, has a water depth of over 4,500 m and in which barely any information on benthic communities is available?

One possibility is to compute species distribution models (Guisan & Thuiller, 2005). However, modelling distribution and abundance of deep-sea organisms is often constrained by the availability of spatially continuous predictor variables at appropriate geographic scales (Ostmann, Schnurr, & Martínez Arbizu, 2014). Most of the GIS layers available in open databases are at low (global) resolution, sometimes only including coastal areas or referring to surface water variables such as temperature and salinity derived from satellite remote sensing (Sayre et al., 2017). The most frequent continuous environmental variables measured in deep marine environments are bathymetry, backscatter strength and derived variables that can be measured using ship-based, seafloor-mapping multibeam echosounders (Lamarche, Orpin, Mitchell, & Pallentin, 2016).

Although deep-sea organisms may not be influenced by depth directly, bathymetric parameters can mirror the influence of other environmental co-variables on the benthic fauna (Ostmann & Martínez Arbizu, 2018) acting as proxies for ecological patterns (Elith & Leathwick, 2009). In the deep sea, increasing depth has usually been linked to a decrease in meiofauna abundance, but this relationship is presumably attributable to a decrease in food availability due to a steady reduction of POC-flux with depth (Ostmann & Martínez Arbizu, 2018). Similarly, Stefanoudis, Bett, and Gooday (2016) observed that abyssal hills influence foraminiferal density, but this pattern can possibly be traced back to enhanced bottom-flow currents that, in turn, influence grain size distribution and food availability for suspension feeders (Stefanoudis et al., 2016).

Another variable having a significant influence on the meiofauna of the CCZ is the distribution and size of polymetallic nodules (Miljutina, Miljutin, Mahatma, & Galéron, 2010). Backscatter strength can be used to distinguish nodule fields on the seafloor from areas devoid of nodules (Weydert, 1990). Moreover, the analysis of backscatter data in the German license area (GLA) also allows for the discrimination between seafloor areas with small-sized nodules (long axis <4 cm) and medium-sized to large-sized nodules (>4 cm) (Knobloch et al., 2017; Kuhn, Uhlenkott, Vink, Rühlemann, & Martínez Arbizu, 2020).

In this study, we focus on metazoan meiofauna. Meiofauna is the size class of benthic organisms passing through a 1-mm mesh and being retained on a 32-µm mesh-sized sieve. With increasing oligotrophy, meiofauna often becomes the most abundant metazoan size class and has the highest biomass (Galéron, Sibuet, Mahaut, & Dinet, 2000). In the CCZ, point source data of meiofauna have been published from different areas (e.g. Miljutina et al., 2010; Pape, Bezerra, Hauquier, & Vanreusel, 2017), but no modelling has been conducted so far. However, our modelling approach could also be applied to macrofauna and megafauna.

Different methods can be used to produce species distribution models (Elith & Leathwick, 2009). In the marine environment, the use of general linear models has become popular (Elith & Leathwick, 2009). However, there can also be nonlinear responses to environmental

parameters, especially using multiple variables of unknown influence (Elith & Leathwick, 2009). Tree-based methods like random forest (Breiman, 2001) are more robust under such conditions (Hastie, Tibshirani, & Friedman, 2001) and are hence preferred here.

Assuming that scientific evidence is essential to support sound management plans in the polymetallic nodule areas, we aim at exploring the predictive power of models calculated with ship-based predictors only, as this is the most likely source of continuous large-scale environmental information for most of the contractors in the CCZ. Combining the environmental data with predictions of faunal distribution based on such models in the form of habitat maps, it is possible to target preservation zones with similar environmental conditions and similarly predicted benthic communities.

To achieve this, distribution models have been computed for meiofauna abundance and diversity to produce environmental baseline information for the target area. To evaluate their use and robustness in management planning, three different aspects of the models have been investigated:

- comparison of the model-based predictions with actual observations;
- influence of varying numbers of parallel samples per site on the predictions;
- influence of varying numbers of sampling sites on the predictions.

2 | MATERIALS AND METHODS

The study area is the eastern GLA allocated by the ISA to the Federal Institute for Geosciences and Natural Resources, Germany (BGR) for the exploration of polymetallic nodules (Figure 1). It is located at the easternmost limit of the polymetallic nodule belt of the CCZ having an average water depth of approximately 4,200 m. Meiofauna has been sampled in the GLA during seven cruises between 2010 and 2016. Sampling has been conducted with a multicore equipped with 12 cores. A total of 88 multicore deployments (319 corers) was used in this study (Figure 1).

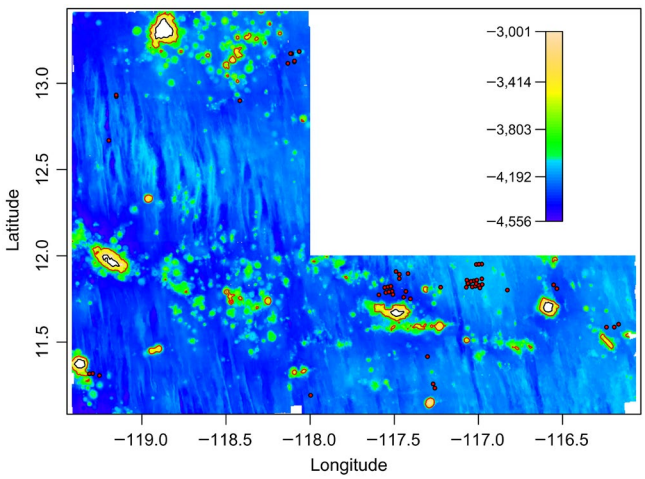

FIGURE 1 Bathymetric map of the German license area for the exploration of polymetallic nodules; lines mark water depths of 3,000 m (black) and 3,700 m (red), points indicate meiofauna sampling positions

On board, the upper 5 cm of sediment was fixed with a 4% formaldehyde/seawater solution. Polymetallic nodules were fixed separately in the same way to enable the investigation of their attached fauna and crevice meiofauna. To extract the meiofauna organisms, the sieved sediment was centrifuged according to the differential flotation method (Heip, Vincx, & Vranken, 1985) with the colloidal gel Levasil®. The polymetallic nodules were pestled gently and centrifuged in the same way as the sediment. All organic materials in the supernatant were stained with Rose Bengal to simplify the recognition of organisms. Finally, metazoan meiofauna was identified and counted on high taxonomic level under a dissecting microscope.

All distribution models were computed with random forest (Breiman, 2001) in R (R Core Team, 2019) using the R-package randomForest (Liaw & Wiener, 2002). Every model is based on 1,000 trees with 1/3 of all variables randomly sampled at each split as suggested by Liaw and Wiener (2002).

Figure 2 shows the basic workflow for deriving environmental predictors from ship-based multibeam data. The ship's multibeam echosounder system with a frequency of 12 kHz produces two basic datasets: water depth based on sound velocity and backscatter based on signal energy (Wiedicke-Hombach & Shipboard Scientific Party, 2009). Backscatter data were standardized to values between 0 and 255. The higher the value, the more likely it is to reflect hard

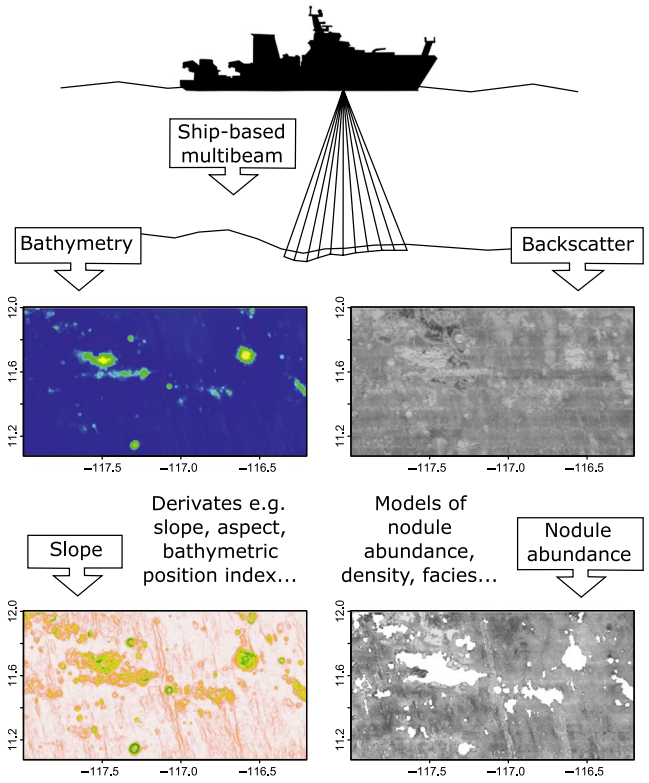

FIGURE 2 Workflow of ship-based multibeam data obtained from the sea surface and their use as predictors for modelling meiofauna in the German license area

rocks on the ocean floor, while lower backscatter values indicate unconsolidated, water-saturated sediments.

From the predicted depth, we produced a contiguous bathymetry map and calculated the derivatives of the maps. Slope, aspect, terrain ruggedness index (TRI) and roughness were computed using the *terrain*-function of the R-package RASTER (Hijmans, 2017). The bathymetric position index (BPI) was computed on two different scales, at the smallest parameter radius appropriate (1 km) and at an intermediate distance (17 km), using the function *focal* also implemented in the RASTER-package (Hijmans, 2017). Additionally, the distance to the next seamount was determined using two different limits. The higher limit defines seamounts as elevations exceeding a water depth of 3,000 m, thus representing an elevation of around 1,000 m above the seafloor in the study area. The lower limit is based on a water depth of 3,700 m, representing approximately 300 m elevation. The distance was calculated from every grid pixel to the nearest contour line of the respective limit.

Both, the backscatter as a proxy for nodule presence and size and the bathymetry data with its derivatives, were used as predictor to model the abundance of polymetallic nodules on the seafloor using artificial neural networks that were trained with a dataset of true nodule abundance obtained from >200 box-corer samples (Knobloch et al., 2017).

Summarizing, a total of 11 predictor variables (Table 1) were used as grid layers with 372,510 pixels having a resolution of 0.0045° (715 m). In the prediction dataset, pixels with depths less than 3,000 m were excluded to avoid prediction on seamounts, resulting in a grid of 206,250 regularly spaced points.

Based on these predictors, distribution models for overall meiofauna abundance and richness (number of higher taxa) were computed. Additionally, the diversity index by Simpson (1949) and the evenness based in this diversity index were modelled. Spatial distributions of the most abundant individual taxon Nematoda are also presented here, as well as of the taxa Tardigrada, Kinoryncha and Ostracoda as their model performances were best of all models (Table 2). To reduce the influence of high small-scale variability of the fauna, the mean of the response variables within all cores from one multicore deployment was used as response variable.

Using the resulting model, meiofauna abundance and diversity were predicted for the whole GLA. Maps were generated based on these predictions using the *rasterize* function (Hijmans, 2017). Prediction accuracy of the random forest models was determined using Pearson's product moment correlation coefficient (r), correlating the predicted values to the observed values at the 88 sampling locations. Additionally, the explained variance computed during model computation and the mean residuals are given.

For habitat mapping of the GLA, only models with positive explained variance and significant Pearson's correlation coefficient were used. The predicted values of the meiofauna were combined with the environmental variables and clustered using the *k*-means clustering algorithm (Hartigan & Wong, 1979) as implemented in the function *cascadeKM* in the R-package VEGAN (Oksanen et al., 2018). The number of clusters used in the habitat maps was set according to the Calinski criterion (Calinski & Harabasz, 1974).

To evaluate the influence of the number of multicore deployments as well as the number of cores used, the random forest models were computed simulating a varying amount of these as training data. To evaluate the number of cores used, subsets of all available cores per multicore were randomly sampled with replacement using 1–5 cores per deployment but integrating all deployments into the models. Likewise, subsets of 21, 42, 63 and 84 deployments were randomly sampled as training data. For both the numbers of cores and the numbers of deployments, numbers were picked that would represent a realistic sampling scope with varying effort. For every sampling scope, 1,000 individual models were computed for meiofauna abundance, diversity and evenness as well as the taxa

TABLE 1 Environmental variables used as predictors for random forest regressions, given R-functions have been used for computation

Predictor variable	Source/citation	R-function
Water depth	Wiedicke-Hombach and Shipboard Scientific Party (2009)	—
Backscatter	Wiedicke-Hombach and Shipboard Scientific Party (2009)	—
Polymetallic nodule abundance	Knobloch et al. (2017)	—
Distance to seamount (>1,000 m)	—	distance()
Distance to seamount (>300 m)	—	distance()
Broad-scale bathymetric position index	Lundblad et al. (2006)	focal()
Small-scale bathymetric position index	Lundblad et al. (2006)	focal()
Slope	Wilson, O'Connell, Brown, Guinan, and Grehan (2007)	terrain()
Aspect	Wilson et al. (2007)	terrain()
Roughness	Wilson et al. (2007)	terrain()
Terrain ruggedness index	Wilson et al. (2007)	terrain()

TABLE 2 Model evaluation for the random forest models applied to different meiofauna response variables; *n* = number of sampled individuals, *r* = pairwise Pearson's correlation coefficient between predicted and observed values; *p* = probability value of *r*

Response variable	*n*	Actual mean per 100 cm^2	Predicted mean per 100 cm^2	Mean absolute residuals per 100 cm^2	% Variance explained	*r*	*p*
Overall abundance	1,114,540	3,494 ± 1,311	3,593 ± 742	433 ± 440	8.56	0.35	0.0007
Richness		10.6 ± 1.9	13.7 ± 1.1	0.8 ± 0.6	−12.34	0.05	0.63
Simpson Index		0.14 ± 0.04	0.13 ± 0.02	0.01 ± 0.01	7.65	0.30	0.004
Evenness		0.51 ± 0.13	0.61 ± 0.08	0.05 ± 0.04	10.85	0.34	0.001
Annelida	8,039	40 ± 21	39 ± 8	6 ± 5	−5.9	0.20	0.06
Copepoda	30,137	149 ± 56	146 ± 24	18 ± 20	−14.47	0.06	0.59
Gastrotricha	1,182	6 ± 8	6 ± 3	2 ± 2	−20.81	0.06	0.58
Kinoryncha	721	3 ± 3	3 ± 1	1 ± 1	5.58	0.29	0.006
Loricifera	1,502	7 ± 10	8 ± 5	4 ± 5	−20.17	−0.002	0.98
Nematoda	665,986	3,256 ± 1,267	3,369 ± 732	415 ± 423	9.31	0.36	0.0005
Ostracoda	2,737	13 ± 9	13 ± 4	2 ± 2	20.05	0.45	~0
Tantulocarida	1,154	6 ± 7	6 ± 2	2 ± 2	−11.19	0.13	0.23
Tardigrada	1,715	8 ± 12	8 ± 7	3 ± 4	32.33	0.58	~0

Nematoda, Kinoryncha, Ostracoda and Tardigrada using identical subsets. Hence, 1,000 habitat maps could be produced for every scope. To evaluate the improvement of the predictions with increasing training data, the predictions across the GLA were compared to the distribution and habitat maps computed with all available data.

3 | RESULTS

3.1 | Meiofauna communities modelled with all available data

The overall mean meiofauna abundance amounts to 3,494 ± 1,311 individuals per 100 cm^2 of sediment (Table 2). Nematoda are largely dominant (92.8 ± 2.5% of all organisms), followed by Copepoda (4.5 ± 1.8%) and Annelida (1.2 ± 0.7%). All other taxa comprise less than 1% of the overall abundance. The richness varies between 3 and 15 higher taxa, with a mean of 11 ± 2 taxa (Table 2). Due to the small number of taxa and the dominance of Nematoda on high taxonomic level, Simpson's diversity is relatively low at 0.14 ± 0.04. The dominance of Nematoda is also the reason for the relatively small evenness of 0.51 ± 0.13.

A highly significant Pearson correlation is obtained for overall meiofauna abundance and diversity as well as for the taxa Kinoryncha, Nematoda, Ostracoda and Tardigrada (Table 2). Regarding these response variables, Pearson's correlation coefficient varies between 0.29 for Kinoryncha and 0.58 for Tardigrada (Table 2), indicating good model fit. Regarding random forest's own indicator 'percent variance explained', the best values can be computed for the taxon Tardigrada (Table 2). The percentage of variance explained is very low for taxon richness and the taxa Annelida, Copepoda, Gastrotricha, Loricifera and Tantulocarida (Table 2); hence, these variables were excluded from further computations.

Although the residuals are large for all taxa, mean predicted abundance and mean observed abundance are almost identical (Table 2). Regarding overall abundance, richness and diversity, the mean absolute residuals are always smaller than the observed standard deviation for all predicted variables. There is also great congruence between observed and predicted mean values (Table 2).

Regarding the spatial distribution predicted with these models, overall abundance is elevated especially within the north-to-south-trending ridges across the GLA, as well as in the north-west (Figure 3b). Lowest abundance is predicted next to the seamount areas at a longitude between −118.5 and −118.0 and in the south-east of the GLA (Figure 3b). The most important spatial predictors for the random forest model are backscatter (Figure 4a), as an indicator for the presence or absence of nodules as well as for the dominating nodule size class, and the modelled abundance of polymetallic nodules (Figure 4b). Generally, an increase in both leads to a stepwise decrease in meiofauna abundance.

Comparing overall abundance to diversity according to Simpson's Index, the predicted distribution is largely opposite (Figure 3a), which can be confirmed by Pearson's correlation (*r*: −0.70; *p* value: ~0). Partial dependence plots of the random forest models suggest that diversity is higher in areas where nodules are present (Figure 4c,d). Spatially, diversity is predicted to be highest directly next to the seamounts, and lowest in the north-western region of the GLA (Figure 3a). The predicted distribution of the Simpson Index is almost identical to the distribution of the evenness, which can be confirmed by a high Pearson correlation (*r*: 0.98, *p* value: ~0).

The predicted distribution of the most abundant taxon Nematoda is almost identical to the distribution of the general meiofauna abundance (Figure 3c) with high correlation of 0.99 (*p* value: ~0) according to Pearson's correlation coefficient The abundance of the taxon Kinoryncha is low in the vicinity of the large seamounts but heightened in the abyssal plains at longitudes between −180 and −119.5 (Figure 3d). This is opposite of the distribution of the taxon Ostracoda, that is especially abundant close to the large seamounts (Figure 3e). The abundance of Tardigrada

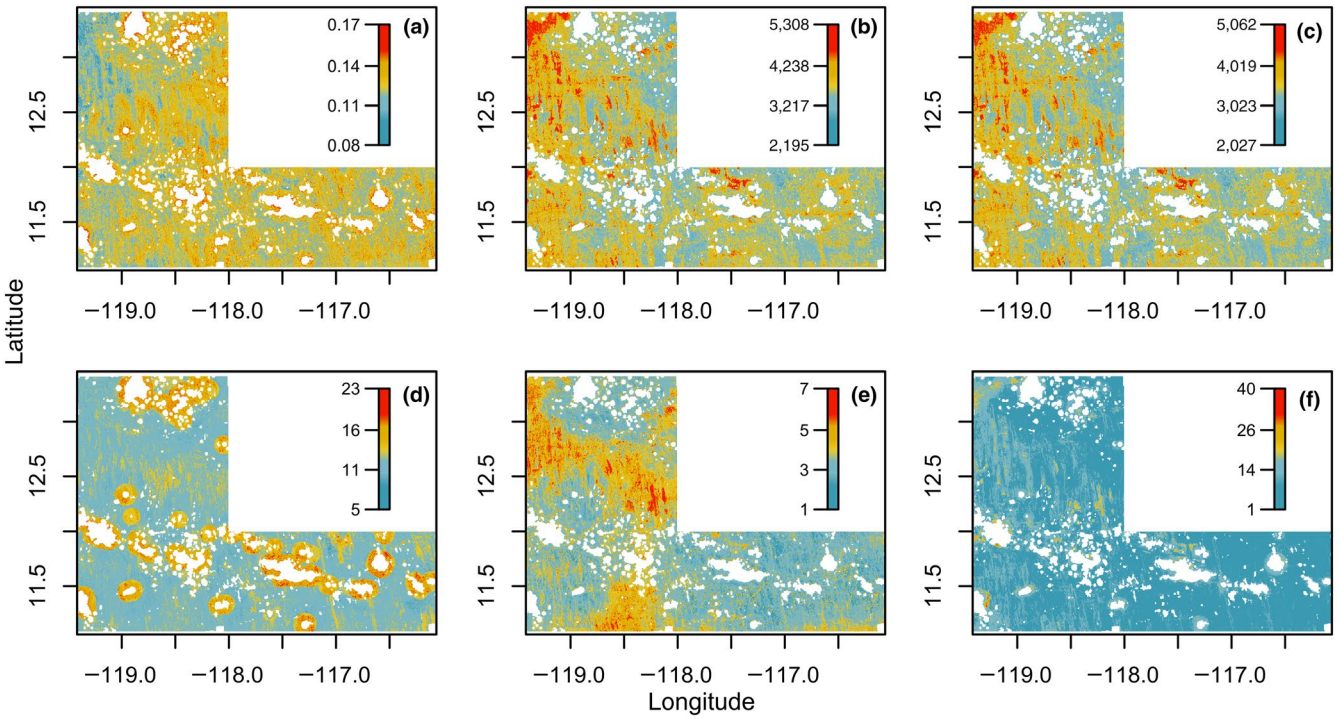

FIGURE 3 Distribution maps for the German license area for (a) Simpson's Index on high taxonomic level, (b) meiofauna abundance, (c) Nematoda abundance, (d) Kinoryncha abundance, (e) Ostracoda abundance, (f) Tardigrada abundance; all abundances are given for 100 cm^2

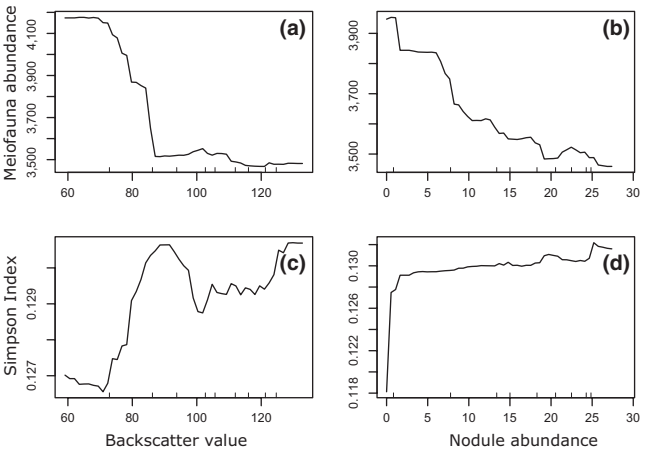

FIGURE 4 Partial dependence of the random forest regression of (a) backscatter signal on meiofauna abundance per 100 cm^2, (b) nodule abundance on meiofauna abundance per 100 cm^2, (c) backscatter signal on Simpson's Index on high taxonomic level, (d) nodule abundance on Simpson's Index on high taxonomic level

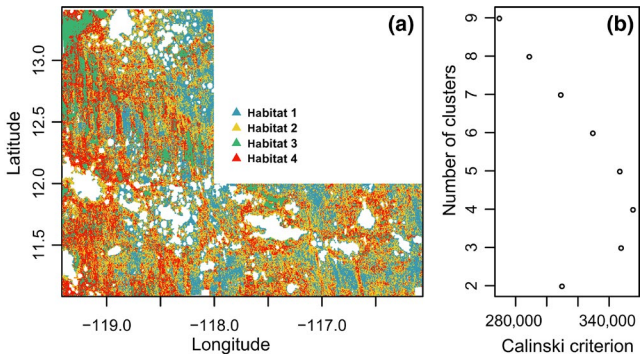

FIGURE 5 (a) Habitat map dividing the German license area in four clusters, (b) Calinski criterion given for varying numbers of *k*-means clusters dividing the habitat

is comparably low across the whole GLA, but high peaks are predicted directly next to seamounts at longitudes between −119.0 and −119.4 (Figure 3f).

3.2 | Spatial habitat mapping

The computed distributions of meiofauna abundance and diversity have been combined with the available continuous environmental variables in order to carry out a habitat classification of the entire area. The Calinski criterion suggests the use of four clusters (Figure 5b), and therefore the area was divided into four habitats (Table 3). The most common habitat is habitat 2, covering 37.1% of the whole GLA. This habitat is spread all over the area without clearly defining larger continuous areas (Figure 5a). It is characterized by high polymetallic nodule abundance and medium meiofauna abundance. The potential for mining in this habitat is high. It is followed by habitat 1 which covers 27.7% of the GLA, specifically in the vicinity of small seamounts (e.g. in the north-western part of the GLA) (Figure 5a). This habitat is characterized by very high nodule abundance and lowest meiofauna abundance. Habitat 4, covering 25.4% of the GLA, is also spread out all over the area (Figure 5a). It

has low nodule abundance and high meiofauna abundance. The least common habitat 3 (9.9% of the GLA) (Figure 5a) covers most areas where the abundance of nodules is predicted to be very low (Table 3) and is therefore unimportant from a mining perspective, but it harbours the highest meiofauna densities.

3.3 | Modelling with varying numbers of parallel samples

Computing all of the models with a varying number of parallel samples, that is, cores obtained with one multicore deployment, the mean standard deviation between predicted points decreases with an increasing number of cores (Figure 6). For the taxa Tardigrada, Kinoryncha and Ostracoda, the standard deviation decreases by one-third when integrating data from four cores; for all other meiofauna attributes, standard deviation decreases by one-third using five cores.

Investigating spatial variation of the predictions for meiofauna abundance in detail, we find that it is not evenly distributed across the whole area (Figure 7a–e). Areas with high variation are observed especially across the rippled area in the northwest of the GLA at latitudes between 12 and 13, and extending into the area north of the seamount chain that covers the GLA from east to west starting at a longitude of −117.5 (Figure 7a–e). Although variation generally decreases with the number of cores integrated into the models, higher variation is still observed at the same positions.

The habitat maps based on predictions integrating varying numbers of cores are almost identical (Figure 7f–j). The mean percentage area of the correctly chosen habitat increases from 64.7 ± 12.9% when integrating one core to 78.5 ± 14.9% when integrating five cores.

The spatial distribution of these variations between clusters is highest in the vicinity of large seamounts (Figure 7k–o). Highest consensus of habitat maps can be observed in the northwest of the GLA (Figure 7k–o). Consensus also increases slightly using an increasing number of parallel samples per deployment.

TABLE 3 Summary of the different habitats computed with *k*-means clustering

	% Coverage	Meiofauna abundance per 100 cm^2	Nodule abundance	Potential for mining
Habitat 1	27.7	3,328 ± 120	Very high	High
Habitat 2	37.1	3,631 ± 94	High	High
Habitat 3	9.9	4,477 ± 180	Very low	Very low
Habitat 4	25.4	3,962 ± 121	Low	Low

3.4 | Modelling with varying numbers of deployments

The variation between models decreases continuously when an increasing number of deployments are integrated into the models. For the abundance of the taxon Tardigrada, the mean of the standard deviation between prediction points decreases by half when using more than 42 deployments (Figure 8b,e–f). For Simpson's diversity, overall abundance and the abundance of the taxa Nematoda, Kinoryncha and Ostracoda, the mean of standard deviation decreases by half when using more than 63 deployments (Figure 8a,c–d).

As already observed during modelling with a varying number of cores per deployment, positions where a high variation is observed remain the same irrespective of the number of deployments (Figure 9a–d). Variation between the models is also especially high in

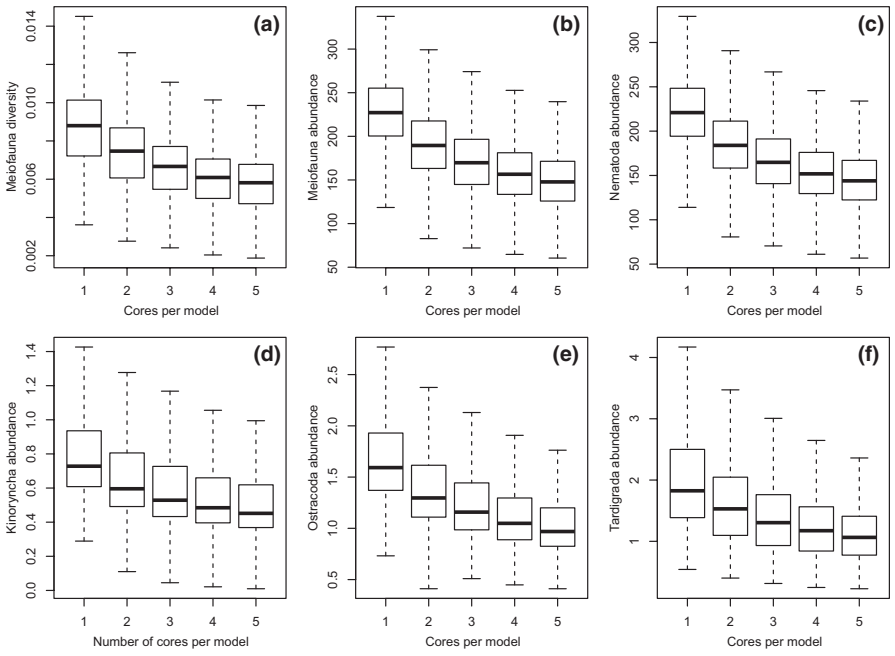

FIGURE 6 Boxplot of the standard deviation excluding outliers computed across all predictions computed with 1,000 replicates for varying numbers of cores: (a) Simpson's Index on high taxonomic level, (b) meiofauna abundance per 100 cm^2, (c) abundance of Nematoda per 100 cm^2, (d) abundance of Kinoryncha per 100 cm^2, (e) abundance of Ostracoda per 100 cm^2, (f) abundance of Tardigrada per 100 cm^2

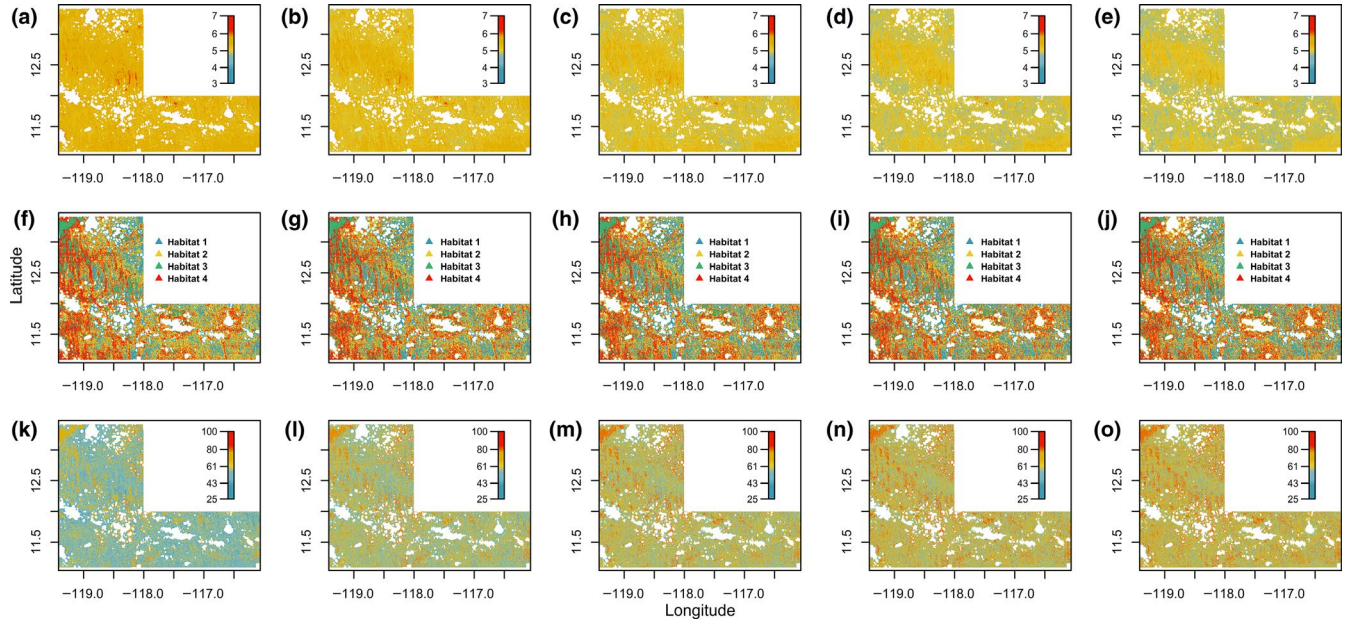

FIGURE 7 Standard deviation across the predictions for overall meiofauna abundance computed with 1,000 replicates including (a) 1 core, (b) 2 cores, (c) 3 cores, (d) 4 cores, (e) 5 cores per deployment; Habitat map dividing the German license area into four clusters computed including (f) 1 core, (g) 2 cores, (h) 3 cores, (i) 4 cores, (j) 5 cores per deployment; Percentage confirmation of the habitat maps across the 1,000 replicates including (k) 1 core, (l) 2 cores, (m) 3 cores, (n) 4 cores, (o) 5 cores per deployment

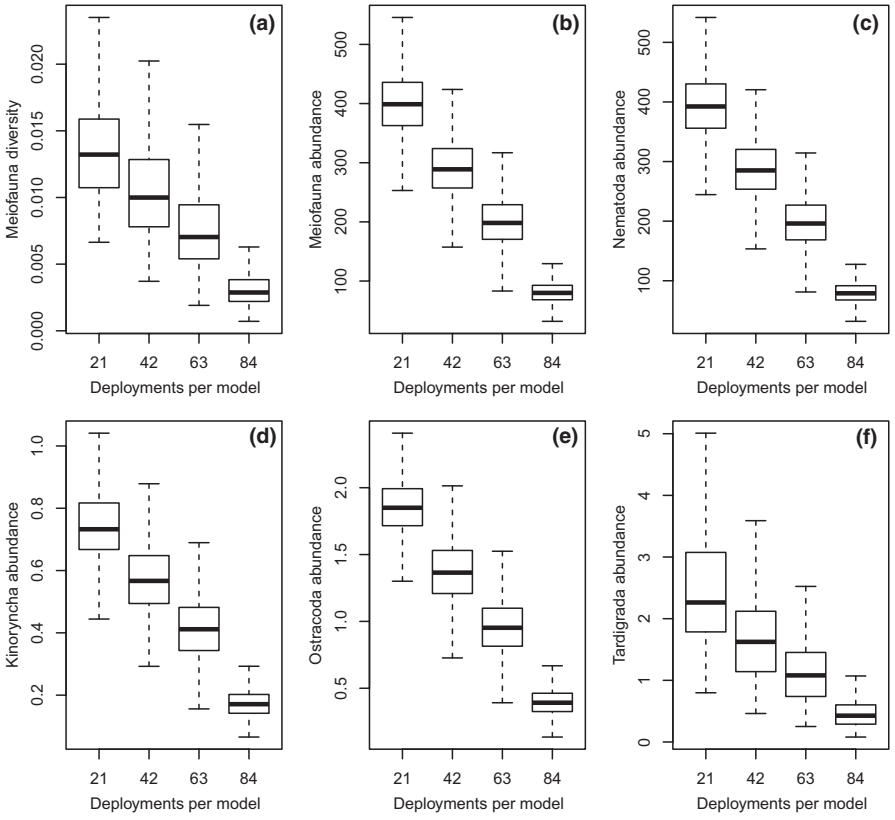

FIGURE 8 Boxplot of the standard deviation excluding outliers computed across all predictions computed with 1,000 replicates for varying numbers of deployments: (a) Simpson's Index on high taxonomic level, (b) meiofauna abundance per 100 cm^2, (c) abundance of Nematoda per 100 cm^2, (d) abundance of Kinoryncha per 100 cm^2, (e) abundance of Ostracoda per 100 cm^2, (f) abundance of Tardigrada per 100 cm^2

the vicinity of large seamounts and low in the very northwest of the GLA (Figure 9a–d).

In contrast to the varying numbers of cores used per deployment, higher numbers of deployments are needed to find consensus between habitat maps (Figure 9i–l). High deviation occurs when only small numbers of deployments are included, but with at least 63 deployments the habitat maps reach consensus again (Figure 9e–h).

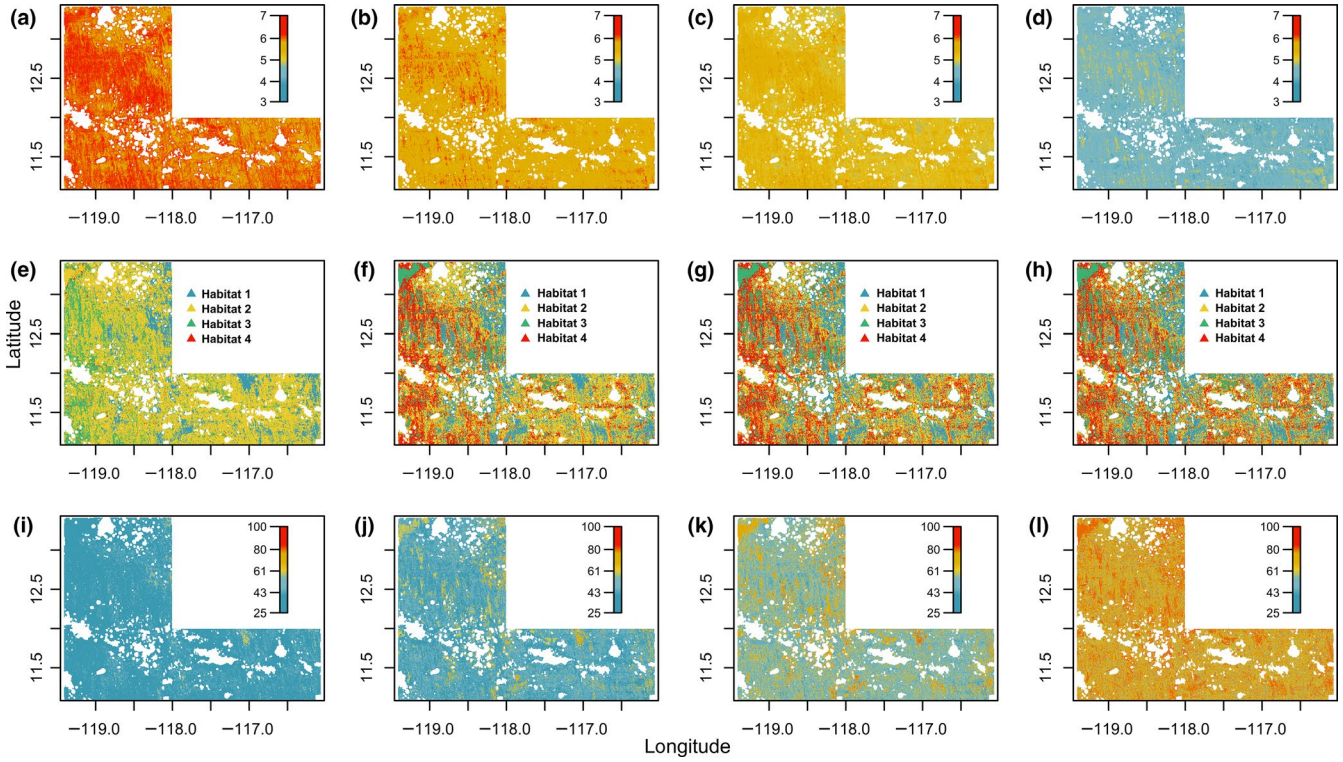

FIGURE 9 Standard deviation across the predictions for overall meiofauna abundance computed with 1,000 replicates including (a) 21 deployments, (b) 42 deployments, (c) 63 deployments, (d) 84 deployments; Habitat map dividing the German license area into four clusters computed including (e) 21 deployments, (f) 42 deployments, (g) 63 deployments, (h) 84 deployments; Percentage confirmation of the habitat maps across the 1,000 replicates including (i) 21 deployments, (j) 42 deployments, (k) 63 deployments, (l) 84 deployments

4 | DISCUSSION

The polymetallic nodule belt in the CCZ is in the focus of exploration activities and therefore, plans for monitoring and conservation of areas threatened by potential future mining activities are of great importance (Kaiser et al., 2017). Contractors are required to prepare an environmental management plan for the area subjected to exploitation before a plan of work can be approved by the ISA. However, information on the deep-sea habitat, its fauna and especially the functioning of ecosystems is still limited and punctual (Ramirez-Llodra et al., 2010). Under these circumstances, contractors have to define potential mining sites and designate equally diverse and representative preservation zones within their exploration areas.

In this contribution, we advocate for a classification of the whole exploration license area into habitats with similar biotic and abiotic characteristics as basic information that should be included into any environmental management plan. We were able to show that the information derived from ship-based multibeam echosounders is a reliable and adequate source for a series of environmental predictors that can be used to produce models of meiofauna communities useful for decision-making purposes (see Table 2). Obtaining such multibeam data for entire license areas is not extremely cost- or time-consuming and forms the basis of any exploration strategy, as the prospective nodule fields themselves are also identified using this technique (Kuhn, Rühlemann, & Wiedicke-Hombach, 2012).

Spatial distribution models are commonly used for conservation planning (Guisan & Thuiller, 2005) as they are cost-effective (Kennedy & Jacoby, 1999) and can be accessed relatively easily (Reiss et al., 2015). More specifically, we show here that distribution models computed with random forest regression prove to be useful for investigating spatial trends in meiofauna abundance and habitat mapping across the GLA and can hence be used for spatial conservation and management issues in this area.

Although modelling of meiofauna distribution was successful, the model performance (evaluated as the percentage of explained variance; Liaw & Wiener, 2002) is sometimes low. This is due to the relatively high variance in the dataset itself, that is, between sets of cores taken from the same multicorer deployment. Variance explained is computed as:

$$1 - \frac{\text{mean of squared residuals}}{(\text{standard deviation})^2} \qquad \text{(Liaw \& Wiener, 2002)}.$$

Therefore, if the mean of squared residuals is larger than the squared standard deviation, values even become negative. However, very high variation in abundance is commonly observed for deep-sea meiofauna (Ostmann & Martínez Arbizu, 2018) and is also mirrored in the correlations between predicted and observed values using Pearson's correlation as suggested by Ostmann et al. (2014) and Ostmann and Martínez Arbizu (2018).

The main problem of modelling meiofauna in the CCZ is the very high variability on small spatial scales (Rosli, Leduc, Rowden, & Probert, 2018), being observed even between cores obtained during one multicore deployment. Sources of variability are not only attributable to the natural patchiness of the meiofauna but also to variability of nodule size and abundance, which introduces a bias in meiofauna sampling. In distribution modelling, this high point variability leads to overdispersion, meaning that observed dispersion is greater than expected from the probability distribution (Guisan & Thuiller, 2005). Therefore, overall variability between models decreases when more cores are integrated as training data into the random forest regressions. Spatially, however, variability between predictions remains highest at the same positions with no regard to how many cores are integrated. This is to be expected, as in summary the multiple models integrate all available data and hence reveal positions where the model is less appropriate.

Such areas where differences between model predictions are comparably high can also be observed when comparing the models computed with varying numbers of deployments. Thus, to further improve the models for the abyssal plains, it would be useful to sample these spots and integrate them as training data into the random forest regressions. Although the CCZ is known to be a heterogeneous environment (Kaiser et al., 2017), all samples used in this study were retrieved from comparable depths (4,000–4,500 m) and similar bathymetric conditions in the deep-sea plains. Including areas that are not suitable for mining or more variable areas such as seamount sites would possibly result in better model performance.

The large APEIs in the CCZ have mainly been defined based on environmental proxies such as bathymetry, seamount positions and nitrogen flux (Wedding et al., 2013). The benthic communities of the CCZ, however, are usually compared in spatially distant, relatively small areas (e.g. Pape et al., 2017). By modelling these point source observations on benthic community distribution over larger spatial scales and integrating them with continuous environmental information into habitat maps, it is possible to define preservation zones or larger conservation areas such as marine-protected areas or vulnerable marine ecosystems including data on the benthic communities.

Although meiofauna is an important component of deep-sea environments supporting high levels of biodiversity (Ramirez-Llodra et al., 2010), it is important to now also include larger organisms of megafauna and macrofauna size into modelling exercises such as the one presented here. Furthermore, the environmental proxies should be supplemented with environmental variables that are known to be of significant influence on the community, such as particulate organic matter concentration and other sediment properties.

5 | CONCLUSIONS

The definition of mining areas (impact zones) and ecologically similar preservation zones within manganese nodule license areas in the CCZ should be based on scientific evidence. We advocate the use of ship-based, derived predictors and statistical learning to model the distribution of nodule abundance and benthic communities on the seafloor over large spatial scales. This information can then be classified into spatially defined deep-sea habitats, which in turn should be used to determine the appropriate location and extent of mining and preservation zones. To improve model performance, we recommend the use of sampling positions that cover as many different environmental conditions as possible as well as the use of several parallel samples per location to account for the high small-scale variability. If a baseline dataset is already available, additional sampling points can be set at positions where highest variability is observed between predictions of multiple models.

ACKNOWLEDGEMENTS

We thank the captains and crew of the research vessels 'Sonne', 'Kilo Moana' and 'L'Atalante' for their help and expertise during the scientific cruises 'Mangan2010 (SO-205)', 'BIONOD2012', 'Mangan2013', 'Mangan2014', 'FLUM (SO-240)', 'EcoResponse (SO-239)' and 'Mangan2016'. We especially thank Lena Albers, Sarah Menke, Dr. Annika Janssen and Focke Weerts for their help with processing and sorting the meiofauna. The cruises with RV Sonne were financed by the German Federal Ministry of Education and Research (BMBF). We acknowledge funding from BMBF contracts 03F0707E and 03F0812E as a contribution to the European project JPI-Oceans 'Ecological Aspects of Deep-Sea Mining'.

AUTHORS' CONTRIBUTIONS

P.M.A. and K.U. conceived the ideas and designed methodology; all authors collected the data; K.U. and P.M.A. analysed the data; K.U. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

Data on meiofauna, distribution and habitat maps are available via the PANGEA data publisher and information system https://doi.org/10.1594/PANGAEA.912217 (Uhlenkott, Vink, Kuhn, & Martínez Arbizu, 2020). The geological data (bathymetry, backscatter, modelled nodule abundance) were obtained in the framework of exploration for polymetallic nodules in the Clarion Clipperton Fracture Zone (CCZ), Pacific. The exploration license was granted to the Federal Institute for Geosciences and Natural Resources (BGR), Germany, by the International Seabed Authority (ISA). As these data are used for the evaluation and resource classification of prospective mining sites in the area in addition to answering research questions, the data are classified as being confidential. Bathymetry and backscatter maps (Wiedicke-Hombach & Shipboard Scientific Party, 2009) and the map of the predicted distribution of nodule abundance (Knobloch et al., 2017) are stored at the Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), Hannover, Germany (contact Thomas.Kuhn@bgr.de).

ORCID

Katja Uhlenkott https://orcid.org/0000-0002-1346-2621

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Galéron, J., Sibuet, M., Mahaut, M.-L., & Dinet, A. (2000). Variation in structure and biomass of the benthic communities at three contrasting sites in the tropical Northeast Atlantic. *Marine Ecology Progress Series*, *197*, 121–137.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York, NY: Springer Science & Business Media.

Heip, C., Vincx, M., & Vranken, G. (1985). The ecology of marine nematodes. *Oceanography and Marine Biology – An Annual Review*, *23*, 399–489.

Hijmans, R. J. (2017). raster: Geographic data analysis and modeling (Version 2.6-7). Retrieved from https://CRAN.R-project.org/package=raster

Jones, D. O. B., Kaiser, S., Sweetman, A. K., Smith, C. R., Menot, L., Vink, A., … Clark, M. R. (2017). Biological responses to disturbance from simulated deep-sea polymetallic nodule mining. *PLoS ONE*, *12*(2), e0171750. https://doi.org/10.1371/journal.pone.0171750

Kaiser, S., Smith, C. R., & Martínez Arbizu, P. (2017). Editorial: Biodiversity of the Clarion Clipperton Fracture Zone. *Marine Biodiversity*, *47*(2), 259–264. https://doi.org/10.1007/s12526-017-0733-0

Kennedy, A. D., & Jacoby, C. A. (1999). Biological indicators of marine environmental health: Meiofauna – A neglected benthic component? *Environmental Monitoring and Assessment*, *54*(1), 47–68. https://doi.org/10.1023/A:1005854731889

Knobloch, A., Kuhn, T., Rühlemann, C., Hertwig, T., Zeissler, K.-O., & Noack, S. (2017). Predictive mapping of the nodule abundance and mineral resource estimation in the Clarion-Clipperton Zone using artificial neural networks and classical geostatistical methods. In R. Sharma (Ed.), *Deep-sea mining: Resource potential, technical and environmental considerations* (pp. 189–212). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-52557-0_6

Kuhn, T., Rühlemann, C., & Wiedicke-Hombach, M. (2012). Developing a strategy for the exploration of vast seafloor areas for prospective manganese nodule fields. In H. Zhou & C. L. Morgan (Eds.), *Marine minerals: Finding the right balance of sustainable development and environmental protection* (pp. K1–9). Shanghai, China: The Underwater Mining Institute.

Kuhn, T., Uhlenkott, K., Vink, A., Rühlemann, C., & Martínez Arbizu, P. (2020). Manganese nodule fields from the Northeast Pacific as benthic habitats. In P. T. Harris & E. Baker (Eds.), *Seafloor geomorphology as benthic habitat: GeoHab Atlas of seafloor geomorphic features and benthic habitats* (2nd ed., pp. 933–947). Amsterdam, The Netherlands: Elsevier.

Lamarche, G., Orpin, A. R., Mitchell, J. S., & Pallentin, A. (2016). Benthic habitat mapping. In M. R. Clark, M. Consalvey, & A. A. Rowden (Eds.), *Biological sampling in the deep sea* (pp. 80–102). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118332535.ch5

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.

Lundblad, E. R., Wright, D. J., Miller, J., Larkin, E. M., Rinehart, R., Naar, D. F., … Battista, T. (2006). A benthic terrain classification scheme for American Samoa. *Marine Geodesy*, *29*(2), 89–111. https://doi.org/10.1080/01490410600738021

Miljutina, M. A., Miljutin, D. M., Mahatma, R., & Galéron, J. (2010). Deep-sea nematode assemblages of the Clarion-Clipperton Nodule Province (Tropical North-Eastern Pacific). *Marine Biodiversity*, *40*(1), 1–15. https://doi.org/10.1007/s12526-009-0029-0

Miller, K. A., Thompson, K. F., Johnston, P., & Santillo, D. (2018). An overview of seabed mining including the current state of development, environmental impacts, and knowledge gaps. *Frontiers in Marine Science*, *4*, 1–24. https://doi.org/10.3389/fmars.2017.00418

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2018). vegan: Community ecology package (Version 2.5-2). Retrieved from https://CRAN.R-project.org/package=vegan

Ostmann, A., & Martínez Arbizu, P. (2018). Predictive models using randomForest regression for distribution patterns of meiofauna in Icelandic waters. *Marine Biodiversity*, *48*(2), 719–735. https://doi.org/10.1007/s12526-018-0882-9

Ostmann, A., Schnurr, S., & Martínez Arbizu, P. (2014). Marine environment around Iceland: Hydrography, sediments and first predictive models of Icelandic deep-sea sediment characteristics. *Polish Polar Research*, *35*(2), 151–176. https://doi.org/10.2478/popore-2014-0021

Pape, E., Bezerra, T. N., Hauquier, F., & Vanreusel, A. (2017). Limited spatial and temporal variability in meiofauna and nematode communities at distant but environmentally similar sites in an area of interest for deep-sea mining. *Frontiers in Marine Science*, *4*, 1–16. https://doi.org/10.3389/fmars.2017.00205

R Core Team. (2019). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Ramirez-Llodra, E., Brandt, A., Danovaro, R., De Mol, B., Escobar, E., German, C. R., … Vecchione, M. (2010). Deep, diverse and definitely different: Unique attributes of the world's largest ecosystem. *Biogeosciences*, *7*(9), 2851–2899. https://doi.org/10.5194/bg-7-2851-2010

Reiss, H., Birchenough, S., Borja, A., Buhl-Mortensen, L., Craeymeersch, J., Dannheim, J., … Degraer, S. (2015). Benthos distribution modelling and its relevance for marine ecosystem management. *ICES Journal of Marine Science*, *72*(2), 297–315. https://doi.org/10.1093/icesjms/fsu107

Rosli, N., Leduc, D., Rowden, A. A., & Probert, P. K. (2018). Review of recent trends in ecological studies of deep-sea meiofauna, with focus on patterns and processes at small to regional spatial scales. *Marine Biodiversity*, *48*(1), 13–34. https://doi.org/10.1007/s12526-017-0801-5

Sayre, R., Wright, D., Breyer, S., Butler, K., Van Graafeiland, K., Costello, M., … Stephens, D. (2017). A three-dimensional mapping of the ocean based on environmental data. *Oceanography*, *30*(1), 90–103. https://doi.org/10.5670/oceanog.2017.116

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*(4148), 688–688. https://doi.org/10.1038/163688a0

Stefanoudis, P. V., Bett, B. J., & Gooday, A. J. (2016). Abyssal hills: Influence of topography on benthic foraminiferal assemblages. *Progress in Oceanography*, *148*, 44–55. https://doi.org/10.1016/j.pocean.2016.09.005

Uhlenkott, K., Vink, A., Kuhn, T., & Martínez Arbizu, P. (2020). *Meiofauna abundance and distribution predicted with random forest regression in the German exploration area for polymetallic nodule mining, Clarion Clipperton Fracture Zone, Pacific*. https://doi.org/10.1594/PANGAEA.912217

Vanreusel, A., Hilario, A., Ribeiro, P. A., Menot, L., & Martínez Arbizu, P. (2016). Threatened by mining, polymetallic nodules are required to preserve abyssal epifauna. *Scientific Reports*, *6*(1), 26808. https://doi.org/10.1038/srep26808

Wedding, L. M., Friedlander, A. M., Kittinger, J. N., Watling, L., Gaines, S. D., Bennett, M., … Smith, C. R. (2013). From principles to practice: A spatial approach to systematic conservation planning in the deep sea. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1773), 20131684. https://doi.org/10.1098/rspb.2013.1684

Wedding, L. M., Reiter, S. M., Smith, C. R., Gjerde, K. M., Kittinger, J. N., Friedlander, A. M., … Crowder, L. B. (2015). Managing mining of the deep seabed. *Science*, *349*(6244), 144–145. https://doi.org/10.1126/science.aac6647

Wegorzewski, A. V., & Kuhn, T. (2014). The influence of suboxic diagenesis on the formation of manganese nodules in the Clarion Clipperton nodule belt of the Pacific Ocean. *Marine Geology*, *357*, 123–138. https://doi.org/10.1016/j.margeo.2014.07.004

Weydert, M. M. P. (1990). Measurements of the acoustic backscatter of selected areas of the deep seafloor and some implications for the assessment of manganese nodule resources. *The Journal of the Acoustical Society of America*, *88*(1), 350–366. https://doi.org/10.1121/1.399910

Wiedicke-Hombach, M., & Shipboard Scientific Party. (2009). *Campaign "MANGAN 2008" with R/V Kilo Moana* [Cruise report]. Hannover, Germany: Bundesanstalt für Geowissenschaften und Rohstoffe (BGR).

Wilson, M. F. J., O'Connell, B., Brown, C., Guinan, J. C., & Grehan, A. J. (2007). Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. *Marine Geodesy*, *30*(1–2), 3–35. https://doi.org/10.1080/01490410701295962

How to cite this article: Uhlenkott K, Vink A, Kuhn T, Martínez Arbizu P. Predicting meiofauna abundance to define preservation and impact zones in a deep-sea mining context using random forest modelling. *J Appl Ecol*. 2020;00:1–12. https://doi.org/10.1111/1365-2664.13621