Ore Geology Reviews 162 (2023) 105671

Contents lists available at ScienceDirect

Ore Geology Reviews

journal homepage: www.elsevier.com/locate/oregeorev

Application of random-forest machine learning algorithm for mineral predictive mapping of Fe-Mn crusts in the World Ocean

Pierre Josso [a,*], Alex Hall [a], Christopher Williams [a], Tim Le Bas [b], Paul Lusty [a], Bramley Murton [b]

[a] British Geological Survey, Environmental Science Centre, Keyworth, Nottingham NG12 5GG, UK
[b] National Oceanography Centre, Waterfront Campus, European Way, Southampton SO14 3ZH, UK

ARTICLE INFO

Keywords:
Ferromanganese crust
Deep-sea mineral
Prospectivity analysis
Mineral resources
Machine learning algorithm
Random forest

ABSTRACT

Mineral prospectivity mapping constitutes an efficient tool for delineating areas of highest interest to guide future exploration. Multiple knowledge-driven approaches have been applied for the creation of prospectivity maps for deep-sea ferromanganese (Fe-Mn) crusts over the last decades. The results of a data-driven approach making use of an extensive data collection exercise on occurrences of Fe-Mn crusts in the World Ocean and recent increase in global marine datasets are presented. A Random Forest machine learning algorithm is applied, and results compared with previously established expert-driven maps. Optimal predictive conditions for the algorithm are observed for (i) a forest size superior to a hundred trees, (ii) a training dataset larger than 10%, and (iii) a number of predictors to be used as nodes superior to two. The confusion matrix and out-of-bag errors on the remaining unused data highlight excellent predictive capabilities of the trained model with a prediction accuracy for Fe-Mn crusts of 87.2% and 98.2% for non-crusts locations, with a Kohen's K index of 0.84, validating its application for prediction at the World scale. The slope of the seafloor, sediment thickness, sediment type, biological productivity, and abyssal mountain constitute the five strongest explanatory variables in predicting the occurrence of Fe-Mn crusts. Most 'hand-drawn' knowledge-driven prospective areas are also considered prospective by the random forest algorithm with notable exceptions along the coast of the American continent. However, poor correlation is observed with knowledge-driven GIS-based criterion mapping as the Random Forest considers un-prospective most target areas from the GIS approach. Overall, the Random Forest prediction performs better in predicting a high chance of Fe-Mn crust occurrence in ISA licensed area than the GIS approach, which constitutes an external validation of the predictive quality of the random forest model.

1. Introduction

Mineral prospectivity mapping (MPM) is a complex multi-criteria decision task aimed at delineating prospective areas for exploring undiscovered mineral deposits (Carranza and Laborte, 2015). Over the last two decades the abundance of high-resolution remote sensing data and expansion of digitised regional geological and geophysical surveys have sparked the development of GIS-based solutions for MPM. Methods have evolved drastically from simple logical or arithmetic operators to complex mathematical functions incorporating an ever-increasing amount of geophysical, structural, geochemical, and environmental data, taking advantage of increasing computational power and diversity of tools in GIS-related applications (Rodriguez-Galiano et al., 2015; Wang et al., 2020). A broad range of methodologies exists giving more or less weight to the influence of experts and/or being dominantly data-driven. Each approach was initially developed for a specific exploration case reflecting various stages of mineral exploration (green or brown field), scale (national, local), and the amount of available data.

Knowledge-driven models rely on experts' deep understanding of ore deposits for the attribution of a weight to each data layer, reflecting a subjective, albeit informed, judgement on the association of spatial information with the mineral deposit investigated. As such, a knowledge-driven approach is considered best-suited for green-field exploration in geologically permissive terrains where no or few occurrences of mineral deposits are known (Carranza and Laborte, 2015; Lusty et al., 2012; McKay and Harris, 2016). This approach relies on inputs and knowledge of geologists with adequate experience. It is advantageous as it does not require any extensive datasets or controls but is at the expense of introducing human bias in the predictions. In contrast, data-driven approaches for MPM makes use of mathematical functions and algorithms

* Corresponding author.
E-mail address: piesso@bgs.ac.uk (P. Josso).

https://doi.org/10.1016/j.oregeorev.2023.105671
Received 14 June 2023; Received in revised form 11 September 2023; Accepted 12 September 2023
Available online 14 September 2023

using bivariate (weights-of-evidence, evidential belief modelling) or multivariate (logistic regression, artificial neural networks) analysis. The weights assigned to individual layers are estimated on the spatial association of datasets of prospective recognition criteria, also referred to as predictors, and known mineral deposit locations (Carranza and Laborte, 2015). Data-driven models are probabilistic and variations of multivariate logistic regression analyses such as gradient-boosting and bagging are the most common. The reason for such diversity is found in the capacity of these multivariate MLA to deal more efficiently with the non-linearity and high complexity of spatial relationship between predictors and mineral deposit occurrences. The Random Forest algorithm, making use of bagging methodology, currently dominate the methods available in machine learning algorithms (MLA) for MPM. However, not all approaches offer the same transparency and ease of training. Notably, artificial neural networks, considered as deep machine learning methods, rely on a 'black box' for the attribution of coefficients representing the degree of spatial association between deposits and evidential data that cannot be evaluated by the operator, whilst logistic regression analysis (shallow machine learning) can be scrutinized (Carranza and Laborte, 2015). Deep learning methods such as convolutional neural network (CNN) or graph convolutional network (GCN) have far more complex architectures containing multiple sequential layers (Zuo and Carranza, 2023). The CNN will be capable of integrating neighbouring pixel's information within its evidence base giving it more power to capture spatial coupling relationship between geological features. Alternatively, GCN are an emerging method relying on a pre-established graphic relationship of deposits and their evidential geological features enabling better capture of spatial anisotropic characteristics of the mineralisation (Zuo and Carranza, 2023). This is notably of interest when the mineral system of interest is related to anisotropic features such as faults and fluid circulations. In the context of this study and the mineral system of ferromanganese crusts, use of random forest is justified as a first application of machine-learning methodology for deposits unrelated to anisotropic structures. For more details on shallow and deep-learning methods, the reader is referred to a special edition dedicated to machine-learning-based mapping for mineral exploration (Zuo and Carranza, 2023).

The applicability and efficiency of the random forest algorithm for MPM has been recently evaluated against other knowledge- and data-driven methods such as regression trees, artificial neural networks, support vector machines, weights-of-evidence, logistic regression and evidential beliefs for epithermal gold deposit in Rodalquilar, Spain (Rodriguez-Galiano et al., 2015), the Baguio Gold District, Philippines (Carranza and Laborte, 2015) and gold deposits in Nunavut, Canada (McKay and Harris, 2016). All these studies conclude that Random Forest models are more stable and reproducible, using various dataset sizes and subsets, and outperform other methods in terms of success- and prediction-rate. Similar conclusions were reached when comparing performances of Random Forests with those methods using fuzzy weights of evidence for MPM for skarn and porphyry-epithermal deposits in the southwestern Fujian metallogenic belt, China (Gao et al., 2016; Zhang et al., 2016), and Yangtze River Valley metallogenic Belt, China (Xiang et al., 2020). The Random Forest algorithm is now commonly employed for spatial predictive mapping of mineral prospectivity on land for various commodities including Au, Cu, Fe, W (Carranza and Laborte, 2016; Hariharan et al., 2017; Li et al., 2020; Parsa et al., 2018; Wang et al., 2020; Xiang et al., 2020).

The continuously increasing demand for minerals and metals in the context of decarbonizing societies is pushing mineral exploration into new frontiers (Lusty and Murton, 2018). Notably, the large-scale deployment of technologies capable of harnessing renewable sources of energy and the ongoing electrification of transport and supporting infrastructures are creating an unprecedented demand for a range of critical raw materials such as Ni, Co, Cu, rare earths (REE), Mn, Li, graphite (C), Te, and Pt (Hein et al., 2013). Over the last decades, the interest in the potential for deep-ocean mineral deposits such as polymetallic nodules, ferromanganese (Fe-Mn) crusts, and seafloor massive sulfide deposits to make a significant contribution to global future raw material supply has increased dramatically. Industries have gained exploration licenses for polymetallic nodules, ferromanganese crusts and seafloor massive sulfide deposits in both the high-seas and exclusive economic zones (EEZ), with economic assessment produced for deposits of polymetallic nodules in the Clarion-Clipperton Zone (AMC Consultants, 2021a; AMC Consultants, 2021b) and seafloor massive sulfides in the EEZ of Papua New Guinea (AMC Consultants, 2018). The drive towards the deep-ocean is a response to the growing global demand as a matter of increasing population, urbanization and living standards, coupled to concerns about the security of supply of critical raw materials and the increasing challenges of land-based mining (Calas, 2017; Lusty and Murton, 2018). The ocean covers more than 70% of the surface of the planet and has an average depth of 3700 m (Charette and Smith, 2010), making exploration of deep-ocean mineral deposits a technologic and economic challenge. Consequently, determining areas of high prospectivity is of paramount importance to optimize exploration efforts and investments.

The access to a large range of high-resolution aerial and satellite remote sensing data has made data-driven MPM of on-land deposits an inexpensive and effective method for the determination of prospects in green- and brown-field exploration. Unfortunately, the deep ocean is opaque to most satellite remote sensed data, preventing large-scale, high-resolution, multi-physics imaging of the seafloor and rapid characterization of its lithologies, structures, and other physical properties. These obstacles play a significant role as to why solely knowledge-driven predictive maps exist at a global scale for deep-sea minerals. The most commonly used maps presenting prospective areas for hydrogenetic ferromanganese (Fe-Mn) crusts have been produced by Hein et al. (2013) and were recently updated by Mizell et al. (2022). These highlight regions in the global ocean where samples have been recovered and the 'permissive areas' are subsequently extended around the major local geomorphological features known to host these deposits. Petersen et al. (2016), refined these hand-drawn 'permissive areas' by GIS-mapping of all portions of the oceans cumulating exploration criteria commonly accepted for these deep-sea mineral deposits, i.e. for Fe-Mn crusts; (i) morphological features such as seamounts, guyots, and ridges peaking between 800 and 3000 mbsl, (ii) a sedimentation rate of less than 2 cm/1000 years with less than 500 m total sediment accumulation, and (iii) seafloor older than 10 Ma.

However, a compilation of known occurrences of ferromanganese crust deposits by the authors (Fig. 1) shows that numerous samples fall outside of knowledge-driven MPM areas defined previously for Fe-Mn crusts (1.7 million km^2, Petersen et al. (2016)), suggesting an underestimation of the prospective zones for future exploration of these deposits. The main reason for this difference relates to the choice of parameters which follow the traditional deposit model established for Pacific Seamounts, fitting a majority but not all geomorphological and oceanographic context in the world ocean. Furthermore, the depth range exclude any geomorphological feature shallower than 800 m or deeper than 3000 m based on economic considerations relative to variation in the content of metal of economic interests with depth, as well as feasibility of extraction (Hein et al., 2009).

This study presents the results of MPM based on Random Forest data-driven methods for Fe-Mn crust deposits at the global scale excluding constraints including economic feasibility that is constantly fluctuating as a matter of commodity prices, technological development, socio-economic and geopolitical factors. Establishing a map of known location and area of potential occurrences is necessary prior to further filtering locations based on economic factors relevant to time and location. Furthermore, establishing a predictive map of locations where ferromanganese crusts can be encountered in the world ocean, of economic interest or not, is of further interest to a larger community by providing an updated coverage of Fe-Mn oxide of the seabed. This could be used for evaluating global cycles of metals between continent and

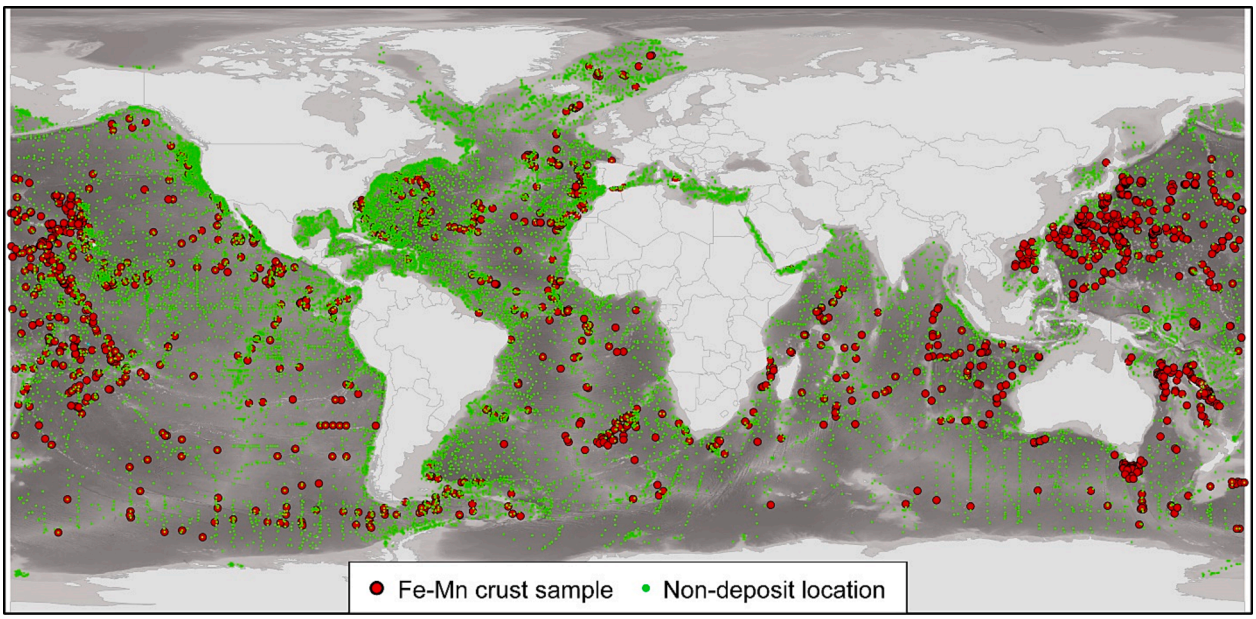

Fig. 1. Compilation of all locations used for the machine learning Random Forest algorithm with Fe-Mn crusts shown in red (n = 4,423) and other non-deposit locations; sediments, nodules, massive sulfides, presented in green (n = 42,600) (See main text for complete reference list for compilation, bathymetric data from GEBCO Compilation Group (2021)).

oceans, stocks of metal stored on the seabed, and informing future paleoceanographic campaigns given these deposits provide faithful records of ocean paleo-environments (Koschinsky and Hein, 2017). Here, we present the performance of the Random Forest machine-learning algorithm and compare its output with previously established knowledge-driven approaches.

2. Random Forest algorithm

The random forest algorithm was pioneered in its modern version by Breiman (2001) as an iterative and randomized succession of regression tree analysis. A decision tree constitutes a powerful tool to efficiently classify a dataset based on its most discriminative features. However, this approach is sensitive to over-fitting and predictions on unknown samples usually result in models with medium bias and high variance. The former represents the model's flexibility to fit unknown data to its trained model, whilst the latter informs on the robustness of the model based on varying training datasets. These parameters are akin to the precision (bias) and accuracy (variance) of any geochemical dataset. Although bias and variance can be optimized by pruning a decision tree, this usually reduces their predictive power. A Random Forest compensates for the shortfalls of individual regression tree classification via two processes. First, a 'bootstrapped dataset' is created for each tree of the Random Forest by sampling with replacement from the input training dataset. This creates a unique dataset where its size does not necessarily equal that of the whole data set and where a data point can be picked more than once. This randomness over many trees ensures that overfitting of each decision tree is compensated by their own uniqueness, therefore, making the model less sensitive to the original training dataset and potential outliers, achieving greater stability, and increasing prediction accuracy (Breiman, 2001; Rodriguez-Galiano et al., 2015). Secondly, the 'random feature selection' minimizes the statistical correlation between all decision trees by forcing them to use a randomly selected subset of predictors for their nodes. As the Random Forest algorithm grows a new tree, it uses the best feature within this randomised subset of predictors to split points, therefore increasing variation between trees and diversification in the classification and prediction process (Carranza and Laborte, 2015). Commonly, the number of nodes each tree uses in a Random Forest model is close to the square root of the total number of predictors. Although limiting the number of usable discriminative features tends to decrease the strength of each single tree, this also reduces the generalization error (Breiman, 2001; Rodriguez-Galiano et al., 2015).

The combined process of bootstrapping and aggregation of random feature selection (also called "bagging" for short) repeated on many decisions tree enforces a randomness and prediction diversity by committee that results in robust and more accurate models. The precision of this model is internally evaluated by the algorithm which makes use of randomly non-sampled data to create a new out-of-bag (OOB) subset and use it to evaluate the model's performance. This OOB approach is beneficial for two reasons; (i) it allows an examination of the performance of the model without the need for an external validation data set, and (ii) it permits an assessment of the relative importance of the different predictors (McKay and Harris, 2016). Breiman (2001) demonstrated that through bagging and the OOB errors, the RF models do not overfit the data as the generalisation error converges as the number of trees in the forest increases.

The next stage of the machine learning process concerns the aggregation of all the decision trees' predictions, which is made following two routes depending on the type of variable to be modelled. In a regression situation, when the variable to predict is numerical for instance, the average prediction of all decision tree is used. In a classification situation, where the data to predict is categorical, the prediction can be made either using the statistical mode of all decision trees or given as an index of relative proportion between the two categories which can be interpreted as an index of the prediction varying between 0 and 1. Either output can be used, although usually prospective vs. non-prospective maps are produced using a probability threshold of 0.5.

A RF algorithm presents the advantages of being simple to implement and capable of dealing with large, uneven datasets, and producing robust and accurate predictions with metrics of uncertainty available for evaluation of model's performance (Graw et al., 2021). However, the downside of this machine learning algorithm is the processing power required to produce the numerous decision trees and compute the predictions depending on the size of the forest, input data, scale of investigation, and desired resolution of output. In addition, as with all supervised or machine learning algorithms, a RF model will be limited by the range of the data it is trained with and cannot extrapolate out of

it, which constitutes an issue mostly for regression modelling rather than for classification modelling.

3. Ferromanganese crusts

Ferromanganese crusts are chemical precipitates formed by the accumulation of Fe and Mn oxides colloids precipitating from ambient cold seawater and depositing on any indurated substrate on the seafloor (Lusty and Murton, 2018). This hydrogenetic precipitation constitutes one of the slowest processes on the planet with accumulation rates on the order of 1–10 mm/Ma with crusts ranging in thickness from 1 to 400 mm (Friedrich and Schmitz-Wiechowski, 1980; Hein et al., 2013; Josso et al., 2019; Lusty et al., 2018). Therefore, Fe-Mn crusts develop in stable environments with low sedimentation rates and non-erosive physical or chemical conditions over tens of millions of years anywhere in the ocean between 400 and 7000 m (Hein et al., 2013; Josso et al., 2020b). They commonly form pavements or encrustation on seamounts, ridges, and plateaus where precipitation, deposition and preservation of Fe and Mn oxide is possible (Josso et al., 2020a; Lusty et al., 2018). Open-ocean settings constitute an optimal environment with low sedimentation rate and strong upwelling currents on the flanks of seamounts keeping hard rock substrate exposed and clean of particles, although reports on occurrences of Fe-Mn crusts in more diversified oceanic setting have increased recently (Charles et al., 2020; Conrad et al., 2017; Konstantinova et al., 2017; Yeo et al., 2018; Zhong et al., 2017). Owing to Fe and Mn oxides physio-chemical properties, slow accumulation rate on the seafloor and high porosity, Fe-Mn crusts constitute effective scavenger of dissolved metals in seawater, building up to economically attractive concentrations of Co, Cu, Ni, REE, Pt and Te (Hein et al., 2013; Josso et al., 2021; Lusty et al., 2018). Regional trends in the composition of Fe-Mn crusts are now clearly identified as more data becomes available from the Atlantic, Arctic, and Indian Ocean, allowing comparison with the large body of work available on occurrences from the Pacific Ocean. Fe-Mn crust composition is influenced by their depth of formation as a reflection of the natural distribution of dissolved metals in the water column, oceanic water properties, biological activity, proximity to other sources of metals into the ocean such as continental land masses and hydrothermal systems, in turn influenced by climate evolution through time (Josso et al., 2021; Mizell et al., 2020; Verlaan and Cronan, 2022). Notably, deposits closer to continental masses or proximal to major sources of detrital material (riverine, aerosols) tends to be richer in Fe, Ti, V, Li, Pt, Te, and Ta, whilst open-ocean settings are richer in Mn and Co (Hein et al., 2017; Hein et al., 2013; Josso et al., 2021).

There are currently five contractors holding International Seabed Authority (ISA) exploration licences for Fe-Mn crusts, sponsored by Japan (JOGMEC), China (COMRA), Brazil (CPRM), Russia (Ministry of Natural Resources and Environment of the Russian Federation), and The Republic of Korea. Although Fe-Mn crusts are considered a complex seabed mineral deposit to explore for and potentially extract due to their two-dimensional occurrence on rough terrains, partial cover by sediment, and indurated substrate, preliminary collector and excavation tests have already been carried out in the Pacific (JOGMEC, 2020). Despite the interest in Fe-Mn crusts, their distribution on the many thousands of seamounts, submarine ridges and plateaus is poorly understood largely as a result of the sparse data and lack of detailed seabed exploration required to verify the presence and volume of Fe-Mn crusts present. Hence the need to develop a quantifiably reliable predictive method to focus future exploration effort.

4. Data sources and predictors

The objective of the Random Forest model is to predict the occurrence of ferromanganese crusts in the world ocean using spatial association between known location of Fe-Mn crusts and other environmental predictors. Therefore, the target variable and inputs constitute a categorical variable, either a deposit or a non-deposit. For deposit locations, a review of available literature was conducted and incorporate the databases from GeoERA - MINDeSEA (2019), NOAA (Frazer and Fisk, 1981), USGS (Manheim and Lane-Bostwick, 1988), and JAMSTEC (JAMSTEC, 2021), as well as samples from peer-reviewed literature in key locations (Baturin and Dubinchuck, 2011; Charles et al., 2020; Frank et al., 2002; Guan et al., 2017; Hein et al., 2017; Konstantinova et al., 2020; Konstantinova et al., 2021; Ren et al., 2019; Zhong et al., 2017) and locations of samples currently under study and not yet published obtained via personal communications. In total, more than four thousand Fe-Mn crust occurrences were compiled and cover most oceans of the planet (Fig. 1).

Non-deposit locations are as important for the algorithm to train the model in identifying the range of spatial signatures where deposits are not present and do not solely correspond to locations where data is lacking. Any ocean sampling location having not recovered a Fe-Mn crust forms a potential non-deposit candidate. These include sediment samples from the world ocean (n = 41,340, Diesing (2020)) as well as locations of polymetallic nodules (n = 856, Frazer and Fisk (1981)) and seafloor massive sulfides (n = 405, Petersen et al. (2016)). Note that only samples contained within the common geographic extent covered by all continuous predictors (Figs. S1-S6) are kept for this analysis and sources cited above may provide data outside of the range presented in Fig. 1. Available predictors limited the geographic extent of the prediction to roughly 80 N to 65S. Despite reports of Fe-Mn crusts in the Arctic and Antarctic Oceans (Hein et al., 2017; Konstantinova et al., 2017; Konstantinova et al., 2020), these locations are excluded from the study due to the lack of data coverage for a reliable prediction of the model.

In this configuration, the number of non-deposits outnumber the amount of deposit locations 10:1, although this parameter is compensated in the random forest algorithm for training the model and constitution of the out-of-bag dataset. To optimise identification of the multivariate spatial data signature of Fe-Mn crust locations, any non-deposit locations within a 20 km radius of each Fe-Mn crust were filtered out (Carranza and Laborte, 2015). A good areal coverage is obtained in the Pacific, Atlantic and Indian Oceans, less in the Southern Oceans and the Arctic Ocean is poorly or not covered.

In total, thirty-two predictors were used in the random forest algorithm for their general relevance to exploration criteria for Fe-Mn crusts and global coverage (Table 1). These include the categorical layers from the geomorphic features map of the global ocean (Harris et al., 2014), providing the major subdivision of oceanic domains and location of

Table 1
List of predictors and data source used in the random forest model (Figs. S1-S6).

Predictors	Variable	Source
Geomorphological classification of the seafloor (n = 23; abyssal hills, abyssal plains, abyssal mountains, canyons, seamounts, guyots, troughs, glacial troughs, trenches, bridges, sills, shelf valleys, rift valleys, ridges, spreading ridges, terraces, fans, rises, plateaus, escarpments, shelf, hadal, shelf slope)	Categorical	Harris et al. (2014)
Average seafloor kinetic energy	Continuous	pers. Comm. Andrew Coward (NOC)
Surface productivity	Continuous	NASA Ocean Biology (OB.DAAC) (2014)
GEBCO Bathymetry	Continuous	GEBCO Compilation Group (2021)
Slope	Continuous	Derived from GEBCO Compilation Group (2021)
Sediment thickness	Continuous	Straume et al. (2019)
Dissolved oxygen	Continuous	Garcia et al. (2006)
Seafloor lithologies (n = 5; Radiolarian ooze, lithic sediment, diatom ooze, clay sediment, calcareous sediment)	Continuous	Diesing (2020)

geomorphological features to be associated with deposits and non-deposit spatial signature by the model. These are complemented by continuous datasets on bathymetry (GEBCO Compilation Group, 2021), from which slope gradients were derived, data on surface bio-productivity (NASA Ocean Biology (OB.DAAC), 2014), sediment thickness (Straume et al., 2019), dominant sediment types (Diesing, 2020), dissolved oxygen (Garcia et al., 2006), and a model of the average seafloor kinetic energy (Coward Pers. Com.). These evidence layers (Figs. S1-S6) were used for their relevance in the formation process of Fe-Mn crusts. Forming in oxidising areas onto exposed rocky surface with low sedimentation rate, evidence layers on the type of outcropping seabed (geomorphological features), surface bio-productivity, sediment types and their cumulated thickness on the seafloor, the dissolved oxygen content of seawater, as well as ocean current strength and slopes affecting deposition or transportation of sediments, are relevant parameters to the mineral system of Fe-Mn crusts and, more importantly, are available in the published literature.

Although commonly considered an important parameter for the exploration of Fe-Mn crusts in relation to their slow growth, the age of the seafloor was excluded from the model for three reasons. Firstly, the age of geomorphological features commonly hosting Fe-Mn crusts such as seamounts, ridges, and plateaus, cannot be assigned to that of their surrounding basaltic seafloor because they are always built on existing seafloor, and no coherent dataset exists for the ages of seamounts in the world ocean. Including this variable from a global age model of the seafloor would bias the dataset as non-representative ages would be associated to deposits and non-deposit locations. Secondly, seafloor age models established on magnetic reversals recorded by the oceanic crust at spreading centres can only cover the oceanic floor. This leaves large swaths of the seabed without age constraints, such as continental margins and intra-continental basins. Whilst historically most Fe-Mn crusts were recovered in open-ocean settings, an increasing number of studies have reported data on Fe-Mn crusts proximal to continental settings for which the age of the substratum is in doubt or not known (Hein et al., 2017; Staszak et al., 2022; Zhong et al., 2017). Finally, the age of the seafloor is commonly used as a cut-off for economically viable deposits as an indirect proxy for the thickness of the deposits. This work aims at modelling areas where Fe-Mn crust might be actively forming, and therefore such limits are irrelevant for this study.

The continuous raster predictors have varying degrees of pixel resolution from 2 arcminutes (~3.7 km) to ¼ arcminutes (~0.5 km). All raster datasets were snapped and resampled at the highest resolution of ¼ arcminutes with a bilinear interpolation.

5. Random forest model

Random Forest models are ensemble learners, with final predictions being made based on the most common prediction made from a forest of separate model runs. The number of these separate models – or rather the number of decision trees (*ntrees*) within the forest – is pre-defined by the user. A key component of the random forest algorithm is that each individual tree considers a random selection of the input data available to it, therefore establishing each tree as being unique. The random forest model approach applied for this study is adapted from that developed in Williams et al. (in. prep). The implementation of the algorithm was through the 'H2O' Python library (LeDell and Poirier, 2020). The code as used in this study will be available under an open license (Williams et al., in. prep).

The input data with which the model was trained consisted of an input table for deposit and non-deposit location with the value or attribute of both numerical and categorical predictors; examples of the former being bathymetric terrain slope and of the latter, terrain feature classifications. Our implementation of the random forest algorithm H20 enabled these different data types to be used with no pre-processing necessary. More than 46,000 data records were available to develop the model for the purposes of predicting Fe-Mn crust locations. Optimisation of the model's hyperparameter such as the test/train data ratio, number of trees and number of predictors is described in the following paragraphs. Output predictions were generated over a grid of 0.03 decimal degree (3.3 km) spatial resolution for which attributes of the predictors were extracted. This resolution was selected as a compromise between processing time and resolution at a world scale.

6. Results and discussion

6.1. Effect of training sample set size and forest size

Predictive bias and variance of Random Forest models are sensitive to hyperparameters of the algorithm. The main parameters the operator influences, and which impact the quality of the model output are the number of trees, the split between the training dataset and data kept for model evaluation (OOB), and the number of predictors each tree can use for its nodes. Using a suite of metrics including the confusion matrix, Cohen's K index and the OOB, Breiman (2001) demonstrated that all models converged in their predictive capability beyond certain hyperparameter thresholds. The spatial distribution of data points selected for being part of the training and validation set, as well as the spatial distribution of the deposit vs non-deposit datapoint are presented in supplementary material (Figs. S7, S8). Both shows the randomise selection maximise spatial coverage of the training and validation (OOB) datasets. During the random selection of the data subset from the training set, the algorithm is coded to maintain the relative proportion of deposit and non-deposit locations.

The influence of the forest size ('ntrees') was evaluated against the training log-loss score (Fig. 2), indicative of how close the model's prediction probability is to the true value associated with the OOB samples. The algorithm was tested for a forest size up to 1,000 trees using 90% of the data randomly selected by the algorithm. The log-loss score sharply decreases over the first increments of testing to stabilise close to a value of 0.2 for a forest of 100 trees or more (Fig. 2).

The size of the randomly selected training set to be used by the algorithm, expressed as a per cent of the total input data, was tested between 1 and 90 %. The log-loss score stabilises close to 0.2 for all models run with more than 10% of the training data set (Fig. 2). The stability of the Random Forest models on data sets of largely different sizes is likely derived from (i) an initially large datasets (n = 46000) assuring statistical representativity of spatial association even using 10% of the data, (ii) utilisation of an optimised number of trees (i.e. 'ntrees' > 100), and (iii) bagging of the training dataset, consistent with findings from Carranza and Laborte (2015). However, for this series of tests the Cohen's k index, a measure of the prediction's reliability accounting for the possibility of the agreement between the OOB data and the prediction to occur by chance, demonstrate a continuous improvement increasing from 0.59 to 0.86 over the 1 – 90 % range (Fig. 2). This trend transcribes the more robust statistical treatment of observed values and plausible causal underlying spatial relationship with predictors, drastically improving the precision of the prediction and reducing rates of false positive and false negatives (Gazis et al., 2018).

Testing for the effect of the number of predictors ('mtries' = 1 – 10) using hyperparameters above identified thresholds for the number of trees and training set size, we find no significant improvement in the log-loss score or Cohen's K index. As a result, the final model was run with the commonly accepted 'mtries' value equivalent to the square root of the number of predictors used by the model, in this case 5 (Breiman, 2001; Carranza and Laborte, 2015).

6.2. Model performance

The model was parameterized to train 600 trees bagging on 75% of the data and compensating for imbalance between categorical data classes. A value well-above the 100 trees threshold was selected to compensate for the larger OOB dataset, which was set at 25% to improve

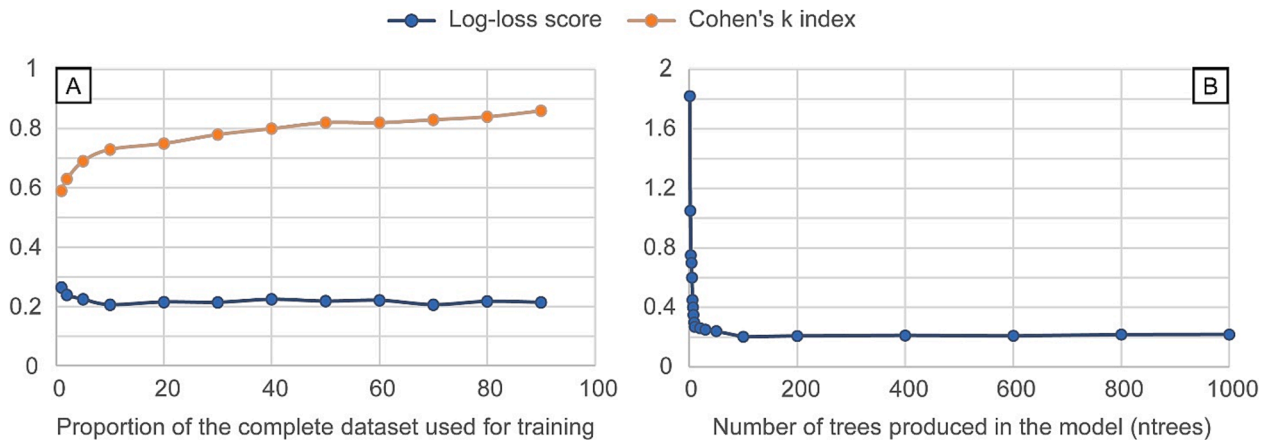

Fig. 2. (a) Effect of the training sample set size in log-loss score and Cohen k index on the Random Forest model. (b) Effect of the forest size on the log-loss score of the random forest model.

reliability of the model's evaluation (Table 2). Iterations of the training process were done to successively eliminate strongly correlated predictors from the model to avoid bias in the prediction. From the geomorphological classification (Harris et al., 2014), "Escarpments" and "Glacial troughs" were removed due to strong correlation ($r > 0.8$) with "Slope" and "Shelf Valley", respectively. The "Bridge" predictor was excluded by default by the model as its limited geographic extent resulted in no location having the attribute. Once removed, all r values in the Pearson correlation matrix of predictors on the training data set were $r^2 < 0.25$, except for the "Bathymetry" and "Lithic sediments" variables, correlated at $r^2 = 0.34$, and considered acceptable for use in a predictive model.

The relative importance of each predictor is evaluated internally by the model by ranking the effect each predictor has on the prediction whilst holding other predictors constant, which is similar to a bi-variate correlation coefficient between the predictor and the predicted variable (Carranza and Laborte, 2015). The ranking of all predictors used in the model is presented in Fig. 3. The "Slope" predictor derived from GEBCO Compilation Group (2021) appears as the strongest explanatory variable in predicting the occurrence of Fe-Mn crusts, followed by "Sediment thickness", "Radiolarian oozes" and "Productivity" (Fig. 3). Of the categorical geomorphological features, "Abyssal mountains", "Ridges", and "Seamounts" are the most impactful in the model. These predictors are in good agreement with the general exploration model for Fe-Mn crusts emphasizing areas of rocky outcrops (e.g., strong slopes, abyssal mountains, ridges, seamounts) with low sedimentation (slope again, negative correlation with sediment thickness and productivity).

To evaluate the model's performance, an output confusion matrix was generated using the remaining 25% of the data (OOB, n = 11,517) not used for the model's creation (Table 2, Figs. S8, S9). The trained model correctly predicts 949 Fe-Mn crusts samples out of 1,088 yielding a prediction accuracy of 87.2% (Fig. S9). No spatial clustering of false positive or false negative is observed (Fig. S9). The prediction accuracy for non-deposit locations reaches 98.2% with therefore a 1.8% false positive rate (Table 2, Fig. S9). For this model, the machine learning algorithm returned a Cohen's k index equals to 0.84. Although there is no general agreement on cut-offs for the interpretation of this index, commonly used scales consider thresholds > 0.75 or > 0.81 to represent strong, or excellent agreement of the model with the OOB data set (Fig. S10), validating the use of this model for prediction of Fe-Mn crusts occurrence in in the world ocean (Bakeman and Quera, 2011; Landis and

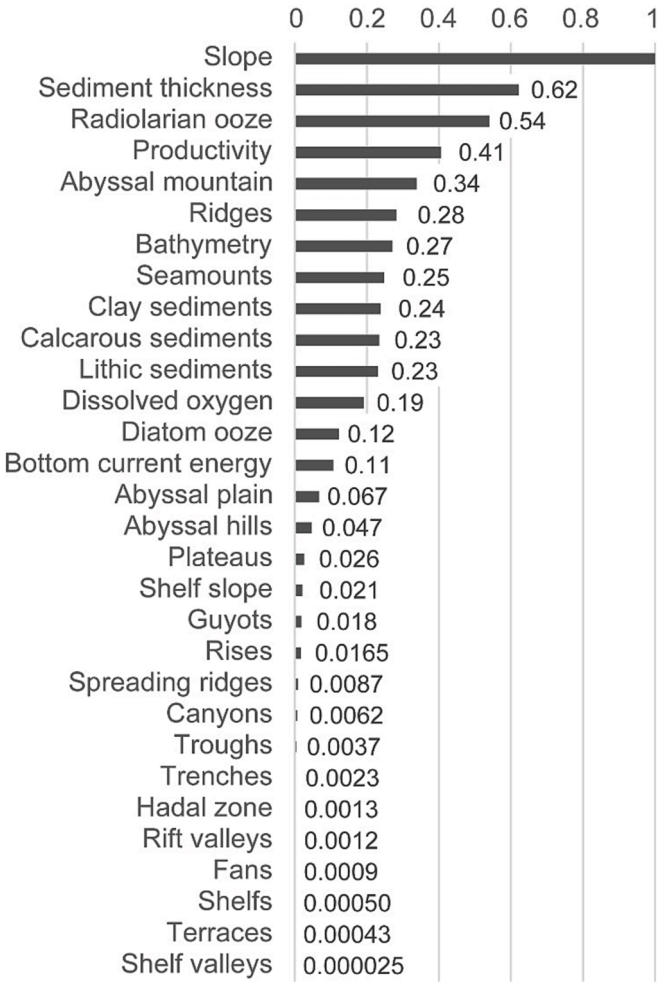

Fig. 3. Rescaled predictor importance in the trained random forest model.

Table 2
Confusion matrix produced on the OOB testing subset by the random forest model.

		Prediction of the model (n)	
		Non-deposit	Fe-Mn crust
Actual (n)	Non-deposit	10,244	185
	Fe-Mn crust	139	949
Prediction accuracy for "Non-deposit" = 98.2 %			False negative = 12.8 %
Prediction accuracy for "Fe-Mn crust" = 87.2 %			False positive = 1.8 %

Koch, 1977).

6.3. Random Forest prediction

The predictions of the random forest model are presented as a probability gradient for the occurrence of Fe-Mn crusts, and as a categorical prospective/non-prospective map using a 50% probability, or 0.5, threshold (Figs. 4, 5). The raster of the prediction is accessible for download from the National Geoscience Data Centre (NGDC) repository (https://doi.org/10.5285/4c8419b9-5ee4-4db4-b279-18d3ec75c3c4). Although the prospective/non-prospective approach constitutes an easy result to interpret, associating an uncertainty to the prediction and delivering the information in an accessible way for such a scale is complex. Therefore, the probability gradient map carries more information highlighting how many trees from the Random Forest concluded on the prospectivity of a given location based on the combination of the model's predictors, allowing to better prioritise exploration targets (Fig. 4). This notably differentiates areas classified as prospective with likelihood of Fe-Mn crust occurrence ranging anywhere between 50 and 91% (Figs. 4, 5). For each pixel of the grid, the prediction is made independently of the presence, or absence, of a sample in the training dataset. An extraction of the random forest model probability of occurrence for each location of the training dataset (Fig. 6), demonstrates the strong consistency of the Random Forest model with field observations and sampling with Fe-Mn crusts having higher probability than non-crust locations.

Relatively high probability of occurrence for Fe-Mn crusts are associated with geomorphological features (seamounts, ridges, cliffs, rift valleys, abyssal mountains) that generate important bathymetric gradients, or slope (Fig. 3). Thus, findings of the Random Forest model corroborate the generally accepted model of occurrence of Fe-Mn crusts, which in this case includes most mid-oceanic ridges as no age limitation have been imposed. Another indirect validation of the accuracy of the Random Forest model is the zonal statistic for the prediction of pixels located in the ISA licenced areas for Fe-Mn crust exploration (Fig. 7). These zones were requested for mineral exploration based on their high potential for Fe-Mn crusts by contractors. The Random Forest model predicts that the licenced areas have a probability of Fe-Mn crust occurrence of 77% on average (median = 81%, standard deviation = 11%, Fig. 7). This result can be regarded as an excellent validation of the accuracy of the Random Forest model as a combination of environmental parameters and geomorphological features. By comparison, close up maps on the ISA licensed areas (Figs. 8, 9) highlights the occasional mismatch between the GIS-based criteria analysis (Petersen et al., 2016) and locations of exploration licences on the Rio Grande Rise and in the Pacific Prime Crust Zone. All ISA licences containing pixels considered as non-prospective by the Random Forest model (i.e. inferior to 50%, Fig. 7) are located on the Rio Grande Rise, except for one licenced block in the Pacific Ocean. The general agreement between the model and licenced areas (Figs. 8, 9) demonstrates the improvement this data-driven method represents over previous knowledge-driven approaches at higher resolution and scales. Whilst this is clear in both Pacific and Atlantic ISA licence locations, the supervised classification of the seabed at Rio Grande Rise produced by Lisniowski et al. (2019) (Fig. 8) shows an excellent match between their Fe-Mn crust outcrop distribution ground-truthed by video transects with the predicted presence of Fe-Mn crusts from the Random Forest model. This match further enforces the quality of the world-scale prediction from the Random Forest model and the difference with the GIS-based approach.

It is notably interesting to observe that the flat top summits of large seamounts are generally considered of lower prospectivity than the edges and flanks (Figs. 8, 9). This is in good agreement with direct seabed observations by remotely operated vehicles (Lusty et al., 2018; Usui et al., 2017; Yeo et al., 2019) showing the common presence of sedimented areas in these regions. Although the presence of Fe-Mn crusts buried under a veil of sediments has been demonstrated

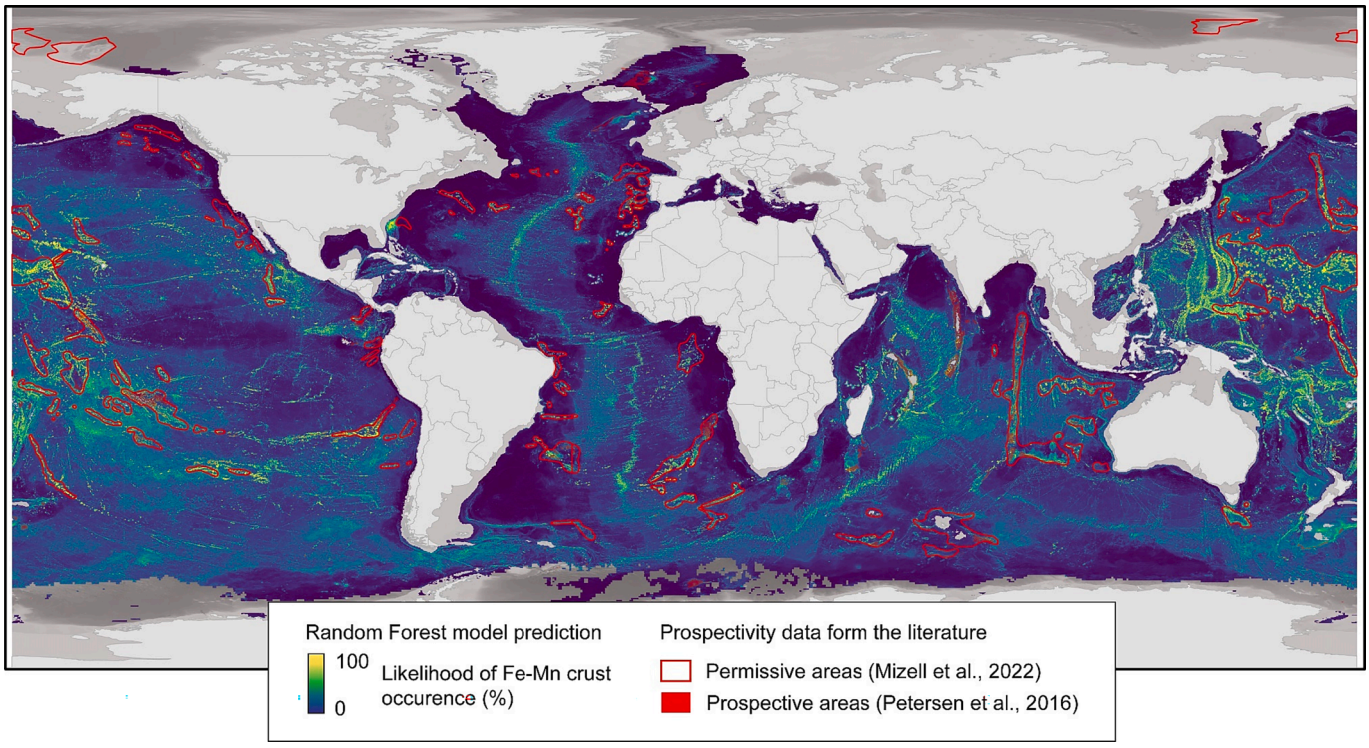

Fig. 4. Probability of Fe-Mn crust occurrence (data available at https://doi.org/10.5285/4c8419b9-5ee4-4db4-b279-18d3ec75c3c4). Previously published mineral prospective maps for Fe-Mn crusts from Mizell et al. (2022) and Petersen et al. (2016) (raw data provided by the authors) are shown in map A. Note that data from Petersen et al. (2016) is displayed with some transparency and may therefore appears in different shades depending on background. Bathymetric data from GEBCO Compilation Group (2021).

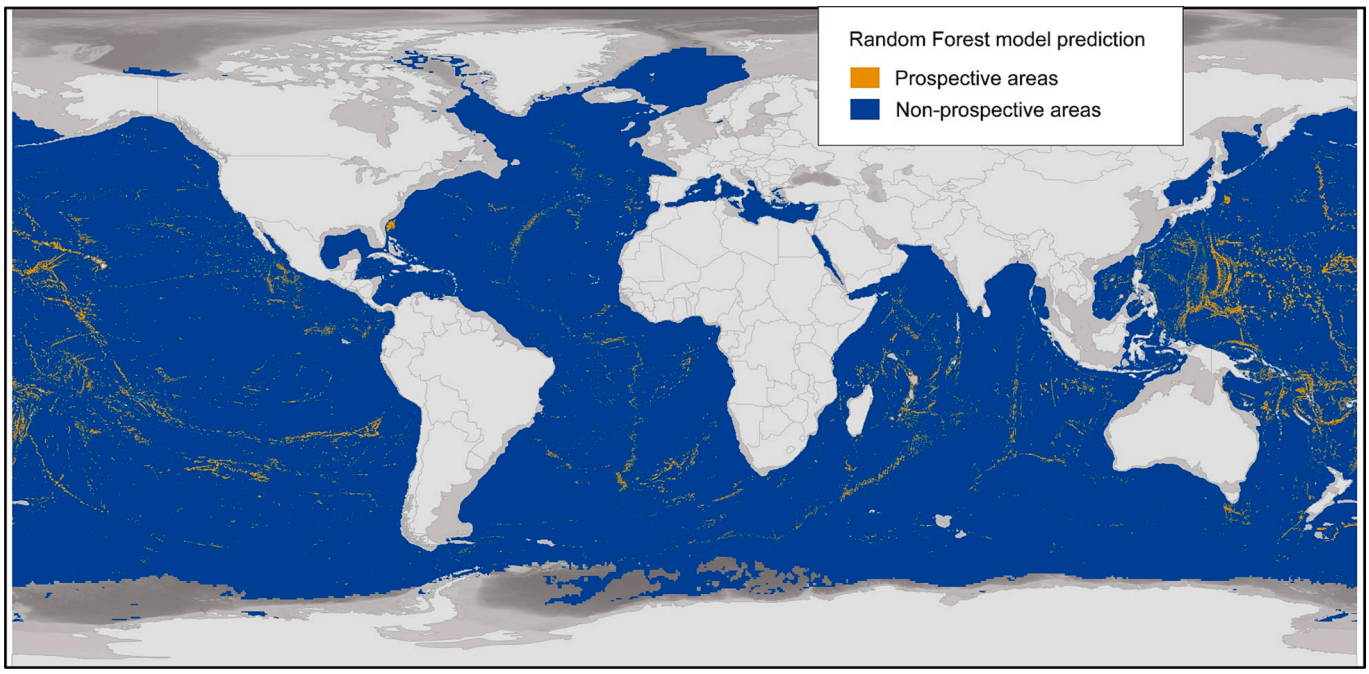

Fig. 5. Prospective versus non-prospective area as defined by the random forest model. The threshold for prospectivity is a probability of Fe-Mn crust occurrence superior to 50%. Bathymetric data from GEBCO Compilation Group (2021).

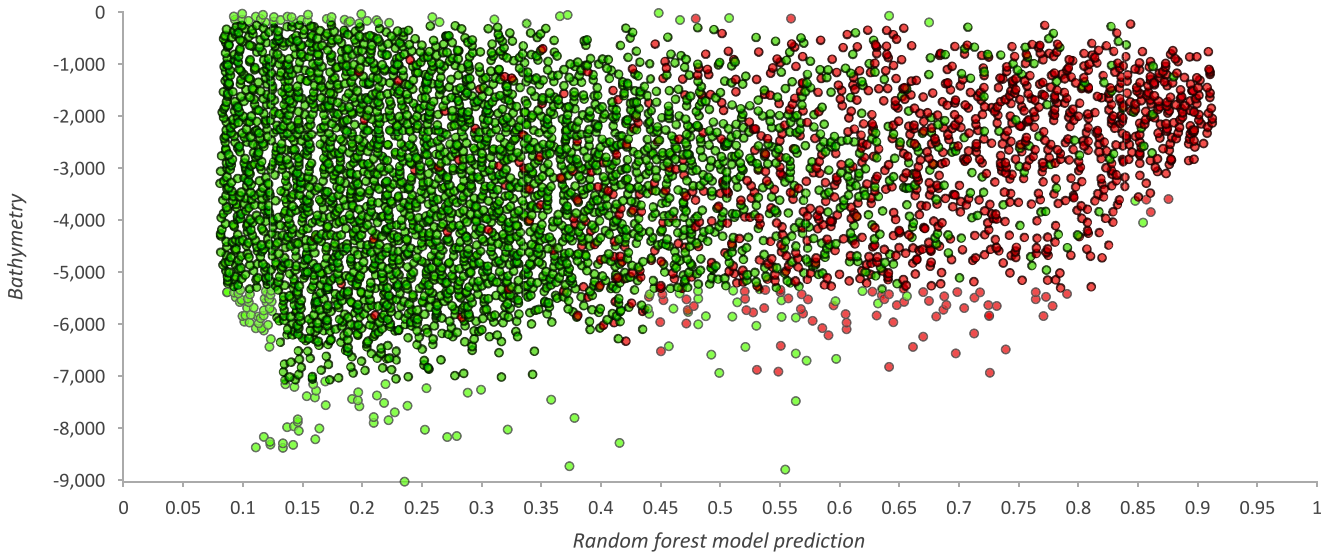

Fig. 6. Random Forest prediction for training locations for Fe-Mn crusts (red) and non-deposit locations (green) as function of bathymetry.

previously and efforts are being made to detect and measure their extent with AUV geophysics (Neettiyath et al., 2022), the model reflects direct observations, but also potentially the sampling bias that exists for these locations. Indeed, apart for locations visited by ROV equipped with a drill or saw blade, most dredging operation will fail to grab a purchase on flat Fe-Mn crust pavements (Lusty et al., 2018).

The model predicts that 34 million km^2 of seafloor have a probability $>$ 50% of hosting Fe-Mn crusts, of which 16 million km^2 are in countries' exclusive economic zones (EEZ). Comparing these values with previous estimates of prospective area for Fe-Mn crusts from Hein et al. (2013) and Mizell et al. (2022) (23 million km^2) and Petersen et al. (2016) (3.1 million km^2) is difficult given differences in scope and methodology. Notably, the surface reported by Mizell et al. (2022) represents that of the hand-drawn polygons, in which large portions of the seabed are abyssal plain unlikely to host Fe-Mn crusts. Predictions from Petersen et al. (2016) were limited to geomorphological features located between 800 and 3000 mbsl, and areas with seafloor older than 10 Ma on the basis of economic criteria not considered in this data-driven approach. It is worth noting that Fig. 5 highlights a limitation of previous approaches as numerous samples fall outside of the 800–3000 mbsl range and could constitute a mineral resource under the appropriate market and legal context.

Overall, a good agreement is observed with the general 'permissive areas' from Mizell et al. (2022). Most polygons contain high Fe-Mn crusts occurrence probability from the random forest model, albeit at various density, directly reflecting the difference between the pixel-by-pixel prediction and the hand-drown delineation of prospective areas. However, noticeable disagreements include 'permissive areas' located off the West coast of South America close to Ecuador and Peru, and the Blake Ridge in the northwest Atlantic where the Random Forest model

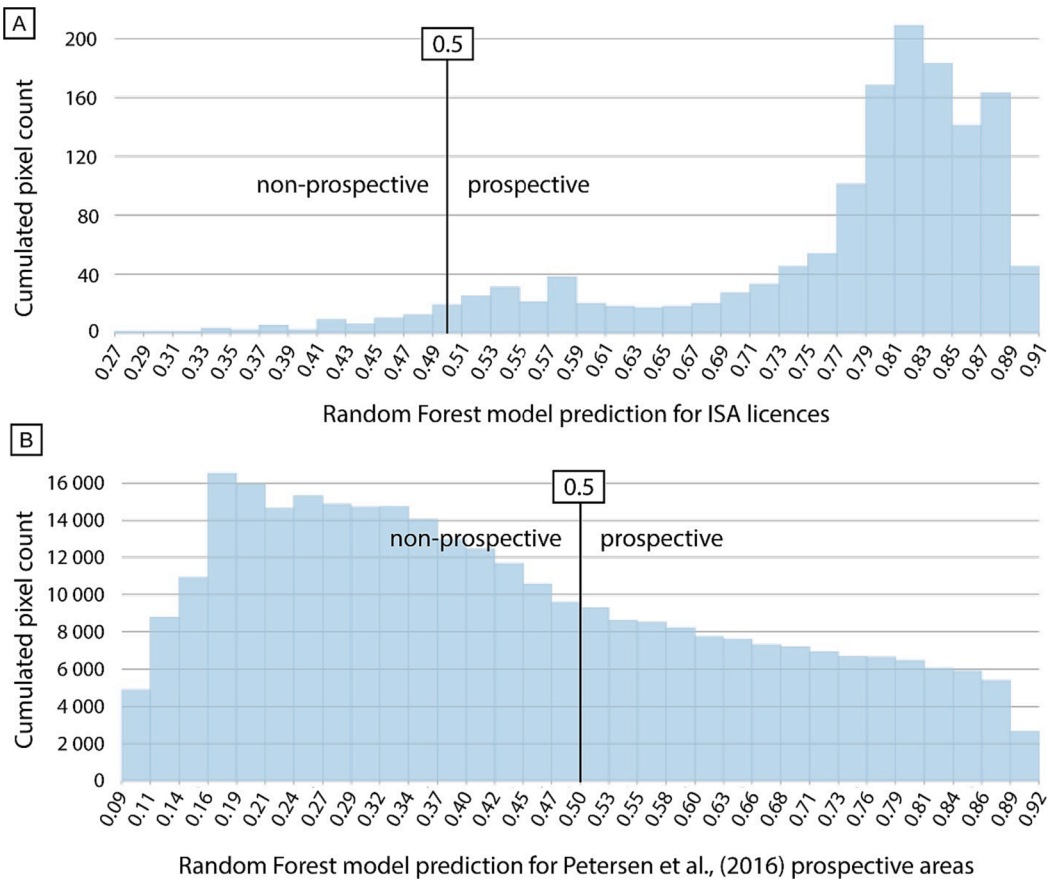

Fig. 7. Frequency of the probability of occurrence of Fe-Mn crusts by the Random Forest model in (A) the ISA licences, and (B) the areas from the GIS-study from Petersen et al. (2016).

predict very low probability of Fe-Mn crust occurrence. These might reflect the legacy use of lower resolution bathymetric maps and other predictors when such areas were first delimited decades ago, highlighting the need of critically considering the presence/absence, but also extent of the other 'permissive areas' reported in these maps.

Whilst assessing the fit between the 'permissive areas' (Mizell et al., 2022) and the Random Forest model remains qualitative, the output generated by the method used by Petersen et al. (2016) allows for a more robust evaluation of the concordance between the two approaches. The match between the Random Forest model and the knowledge-driven GIS selection from Petersen et al. (2016) is statistically poor (Fig. 7). This disagreement on the designation of zones of high prospectivity is evidenced by the overall low probability of Fe-Mn crust occurrence predicted by the Random Forest model in areas from the selective criteria GIS analysis (Figs. 7, 8, 9). Indeed, most of the surface covered by the polygons overlay areas considered as low probability (<50%) for the occurrence of Fe-Mn crusts according to the Random Forest model (Fig. 7). A possible reason for this divergence in prediction may come from the criteria for the exploration of Fe-Mn crusts commonly established for the mining of Fe-Mn crusts from Pacific Seamounts, though such cumulated criteria do not necessarily represent the optimal geographic selection process at the world scale and true combination of environmental, geologic, and geographic factors favouring the formation of Fe-Mn crusts. It is likely that such a combination of criteria is not discriminative enough and large portion of oceans are wrongly incorporated into potential prospective areas. This is further reinforced by the fact that out of all samples catalogued in this compilation, 2,765 Fe-Mn crusts are shallower than 3000 mbsl but only 931 intersect with the polygons defined as prospective by Petersen et al., 2016. Therefore, two third of known crust occurrences, within the depth restrictions imposed by their selection, are not represented by the GIS approach (Petersen et al., 2016). In addition, most ISA licenced areas for exploration are not covered by the polygons resulting from this GIS selection (Figs. 8, 9), further questioning the validity of the approach, or that of the parameters used in a cumulative fashion and reinforce the need to explore alternative methods including data-driven approaches.

Large areas of plateaus or ridges, considered as promising target for the exploration of Fe-Mn crusts by Petersen et al. (2016) are notably discarded by the Random Forest model and highlighted in purple in Fig. 10. This is the case for most of the Ninety East Ridge, the Seychelles Arc, south of Madagascar, the northeastern end of the Walvis Ridge, the Hatton Bank, the North of Iceland, the Azores, offshore Peru – Ecuador – California, the Manihiki Plateau, and a large portion of the Tuamotu Archipelago. The reasons for which each of these regions were not considered prospective by the Random Forest model, albeit presenting favourable geomorphological attributes, is complex to evaluate. Deciphering if this divergence relates to the model approaching its limits due to input data resolution, absence of existing sample in these regions, added value from an expert perspective or represent a true improvement in our capacity to classify an area with a data-driven approach, remains to be validated by direct observations in these locations.

7. Discussion

7.1. Model limitations and improvements

Any data-driven model is inherently limited by the quality and amount of data used for the training as well as the quality, and resolution of the predictors used for fingerprinting spatial associations. The model is notably restricted by the shared geographic extent of all

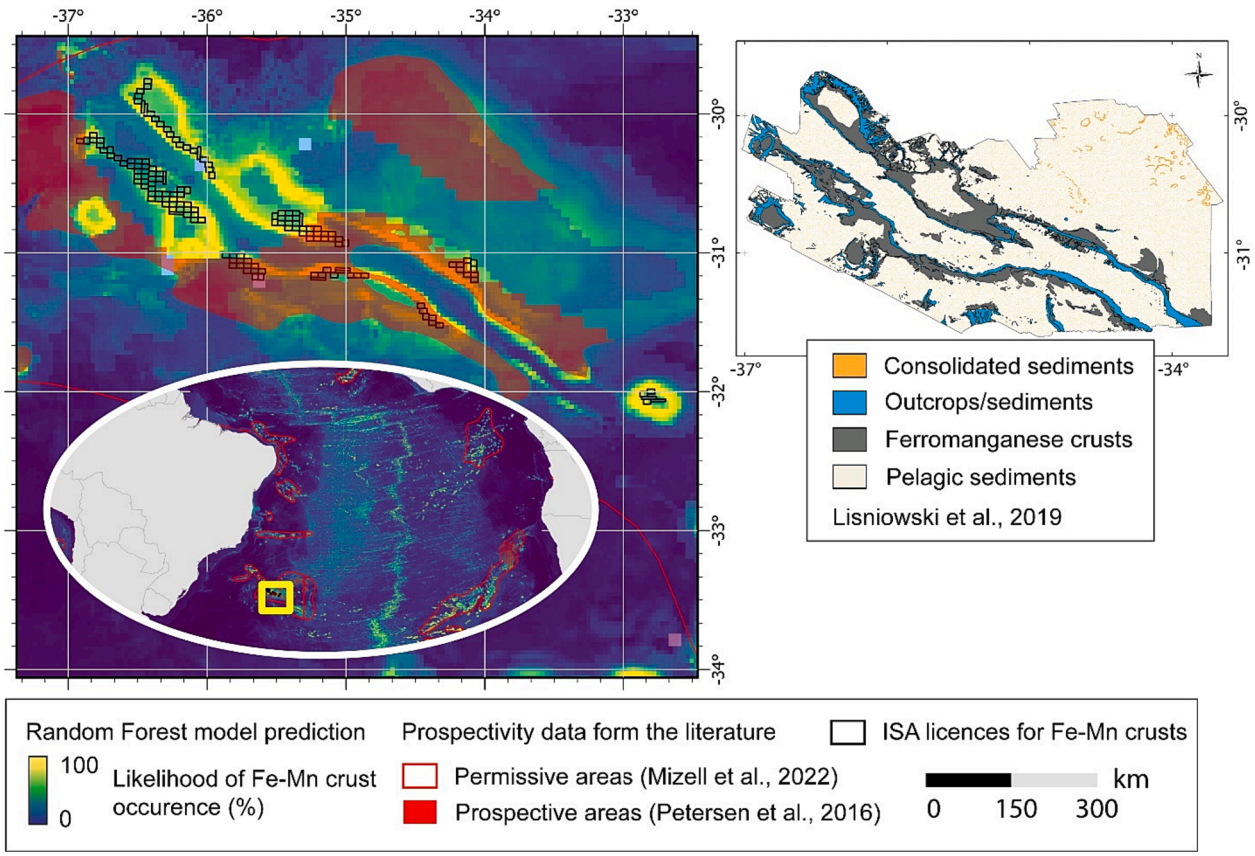

Fig. 8. Close up on the ISA licensed area for Fe-Mn crust exploration at the Rio Grande Rise, South Atlantic Ocean. Note that data from Petersen et al. (2016) (provided by the authors) is displayed with some transparency and may therefore appears in different shades depending on background. Top right inset (to scale with map on the left hand side, aligned with latitude) presents the substrate type of Rio Grande Rise obtained from supervised classification of slope, backscatter intensity ground-truthed by video transect (Lisniowski et al., 2019).

continuous predictors. Whilst the prediction covers a respectable portion of most oceans and represents a large diversity of environments, the Arctic and Antarctic Oceans are poorly represented and locations from these regions cannot be included in the training set, creating a bias as data is excluded. Although these two oceans could be considered minor in terms of geographic extent, oceanic properties in these regions are vastly different (surface productivity, bottom-current strength, salinity) from the more temperate, tropical, and equatorial settings. This therefore limits the predictor's amplitude, or range, against which spatial relationship are evaluated and consequently the conclusion that could be drawn from the output predictions in regions not covered by the model.

Despite these issues, a positive aspect of this data-driven method is that the model and prediction can be easily and rapidly updated as new data are made available, either from samples or other predictors, offering a large flexibility and responsive updates. For example, the model could be further improved by adding new relevant oceanographic datasets to the list of predictors which would refine the quality of the prediction. The model could be taken into other directions by adding more restrictive datasets, potentially related to economic considerations. An important way forward would be to complement this data compilation with information on sample thickness, geochemistry, and age derived from robust dating methods when available. This would provide the opportunity to produce predictive maps of metal content and thicknesses, which cross-referenced with the presented prediction of occurrence would further delineate the strongest prospects for the exploration of elements of interest. The random forest model could be improved by implementing gradient-boosted trees using complementary algorithm such as XGBoost or LightGBM.

The prediction of occurrence between 0 and 100% could be considered as a representation of the spatial coverage of crusts on the seafloor, therefore balancing the outcrop density versus sedimented areas commonly observed at the outcrop scale (Lusty et al., 2018; Yeo et al., 2018). Global estimates of metal content could then be derived from the combination of these three parameters providing another improvement on the latest calculations from Mizell et al. (2022).

7.2. *Exploration recommendations*

Most licenced areas for the exploration of Fe-Mn crusts have an equivalent prediction from the Random Forest model superior to 70%. Although all areas with a Fe-Mn crust occurrence probability superior to 50% are considered prospective, recommendation for prospects of equivalent interest to the existing licenced areas could be determined using a threshold $>$ 70%. Numerous discrete locations in the world ocean are underlined by the model as highly prospective but cannot be all referenced here, and the reader is referred to the online dataset (https://doi.org/10.5285/4c8419b9-5ee4-4db4-b279-18d3ec75c3c4) to explore them. Major 'underexplored areas' could be regarded as large locations with probability of Fe-Mn crust occurrence $>$ 70% that do not fall within the 'permissive areas' (Mizell et al., 2022). These underexplored areas constitute mostly geographic extensions of these polygons and include the Pitcairn Islands, the Louisville Seamount Chain, the northeast of Kiribati, the Musician Seamount Chain, north of the Seychelle Plateau, the south-eastern junction of the Ninetyeast Ridge and Broken Ridge, the north of the New Caledonian basin, the Mariana Ridge, the Blake Plateau, and southern seamounts of the Walvis Ridge,

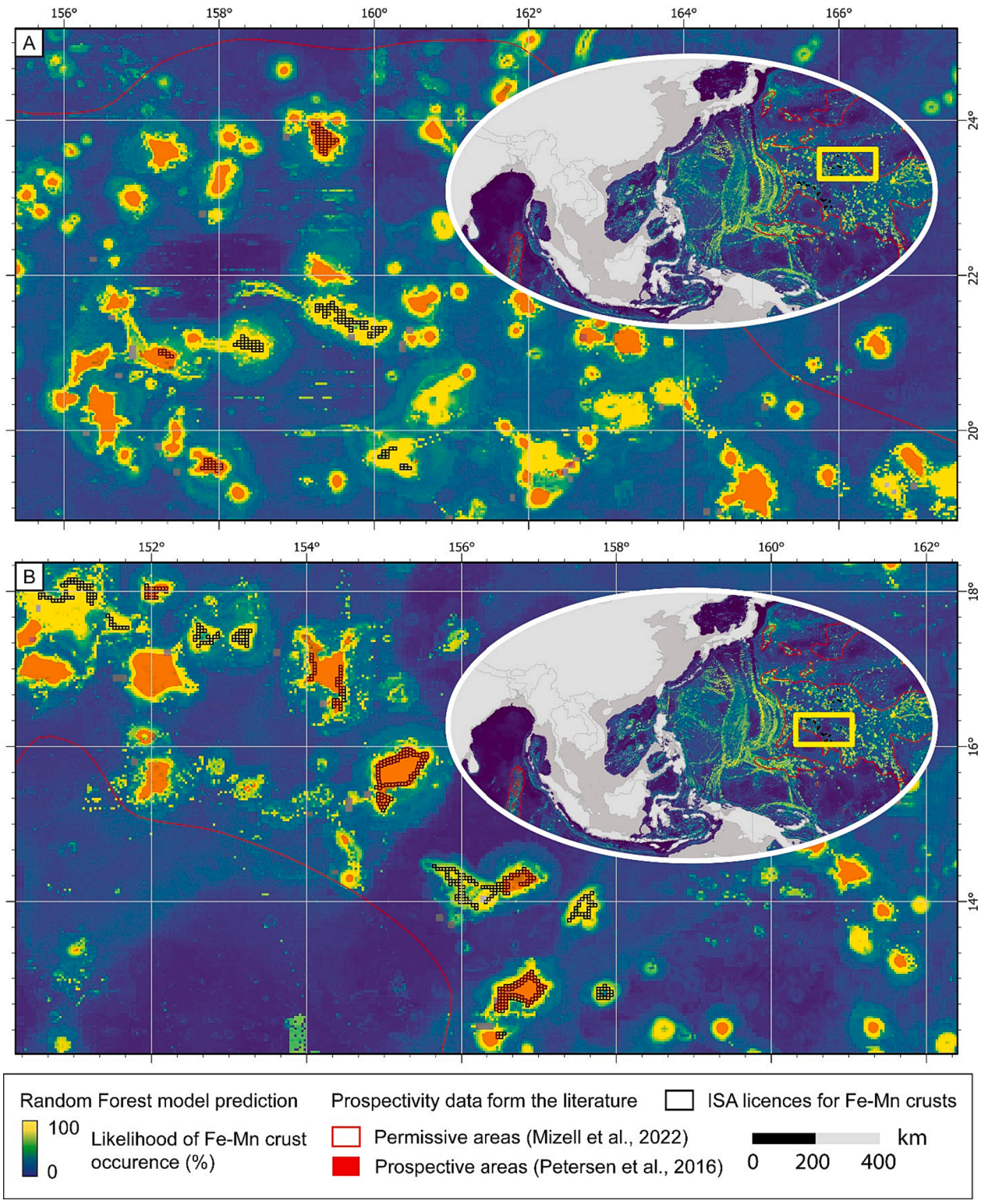

Fig. 9. Close up on the northern (A) and southern (B) Pacific Ocean ISA licensed area for Fe-Mn crust exploration. Note that data from Petersen et al. (2016) (provided by the authors) is displayed with some transparency and may therefore appears in different shades depending on background.

8. Conclusion

The Random Forest algorithm has been demonstrated by numerous studies to capture successfully spatial relationships between predictors and deposit/non-deposit locations for applications in mineral predictive mapping and prospectivity analysis (Carranza and Laborte, 2015; Gazis et al., 2018; Li et al., 2020; Xiang et al., 2020; Zhang et al., 2016). This study presents the first application of this data-driven machine learning approach to the occurrence of Fe-Mn crusts in the World Ocean. A compilation of more than 4,000 Fe-Mn crusts locations and more than 42,000 non-deposit locations was used to train the model against 30 predictors, 11 of which were continuous datasets. Owing to the large size of the input dataset, the bias and variance of the trained model were found to be stable when using 100 or more trees in the forest and a

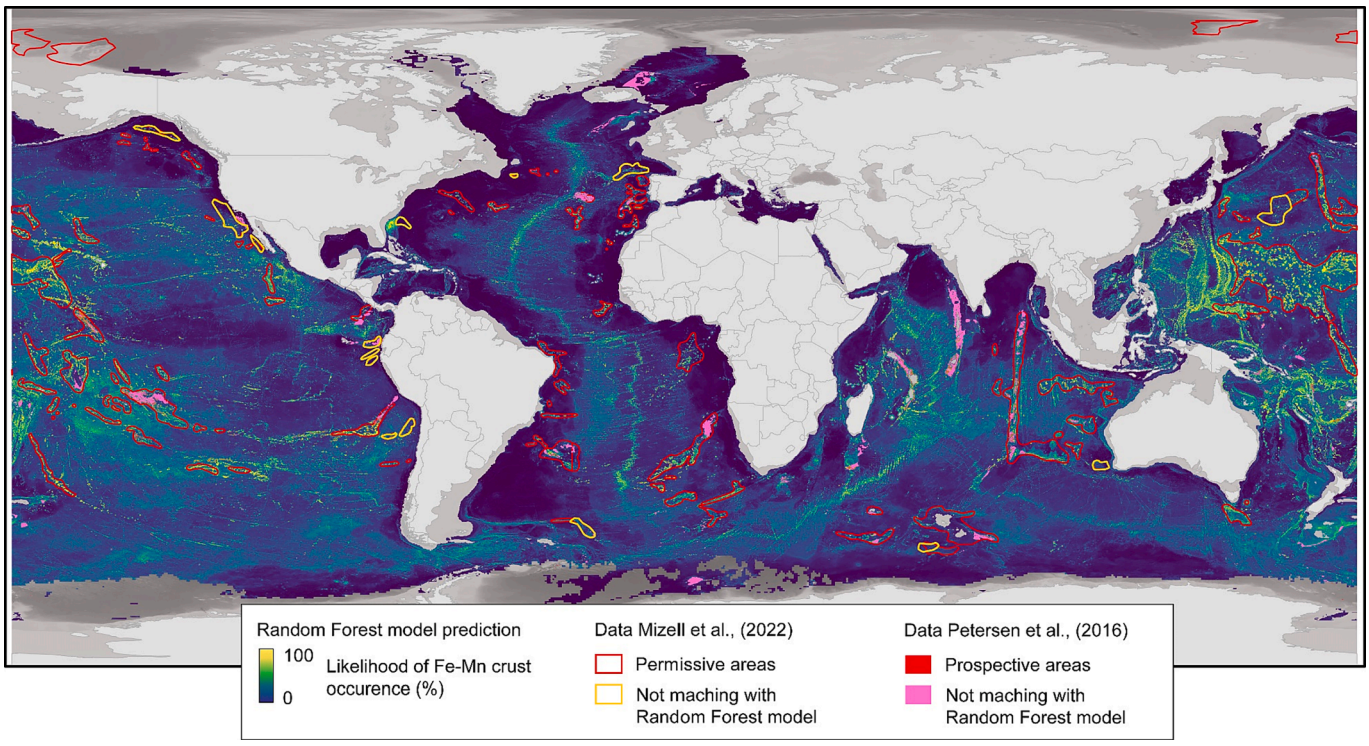

Fig. 10. Comparison of the Random Forest prediction with previous knowledge-driven prospective maps from Petersen et al. (2016) and Mizell et al. (2022). Note that areas from the GIS-based criteria analysis from Petersen et al. (2016) considered as non-prospective by the Random Forest model are highlighted in purple.

training set size superior to 10% of the complete data set. When using these two hyperparameters above identified thresholds, the number of predictors used for the nodes of the trees was observed to be of marginal importance to improve the prediction power of the model and a default setting of the square root of the number of predictors was used in the algorithm. Evaluation of the model using the out-of-bag dataset demonstrate a predictive accuracy of 87.2% on deposit-location and 98.2% on non-deposit location, with a Cohen's k index of 0.84.

The Random Forest model highlight the importance of the bathymetric slope as a major explanatory variable for the occurrence of Fe-Mn crusts. This in turn is consistent with the strong association of geomorphological features such as abyssal mountains, seamounts, and ridges with deposit locations.

Comparison of the Random Forest prospectivity output (https://doi.org/10.5285/4c8419b9-5ee4-4db4-b279-18d3ec75c3c4) with previous knowledge-driven approaches demonstrates the improved success-rate of mineral prospectivity prediction on ISA licenced areas for Fe-Mn crust exploration compared to previous knowledge-driven GIS-selection approaches. Therefore, owing to the stability of the model and high success rate in its predictions, the Random Forest constitutes a net improvement in delineating areas of Fe-Mn crust occurrences, which can be continuously improved as more sample or predictors are added to the model.

Funding

This work was supported by internal funding of the British Geological Survey and UKRI Exploring the frontier grant NE/X011690/1 awarded to Pierre Josso.

CRediT authorship contribution statement

Pierre Josso: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Validation, Project administration. **Alex Hall:** Methodology, Data curation, Validation. **Christopher Williams:** Methodology, Writing – review & editing. **Tim Le Bas:** Methodology, Writing – review & editing. **Paul Lusty:** Writing – review & editing. **Bramley Murton:** Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: A full disclosure of Bramley Murton deep-sea mineral related activities is produced in the interest of transparency.

Data availability

Data used in the model is publicly available in the literature. The machine learning code and output will be made public upon acceptation of the manuscript in the NGDC and in an upcoming publication.

Acknowledgements

PJ, AH, CH, and PL publish with the permission of the Executive Director, British Geological Survey (UKRI). The authors thanks Kira Mizell, Sven Petersen, and Uni Árting for their authorisation to use some of their unpublished data on Fe-Mn crust locations as well as the work of reviewers and editors who contributed to the publication process of this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.oregeorev.2023.105671.

References

AMC Consultants, 2018. Preliminary Economic Assessment of the Solwara Project Bismarck Sea, PNG for Nautilus Minerals Niugini Ltd. Technical Report AMC Project 317045, 274 pp.

AMC Consultants, 2021a. Initial assessment of the NORI Property, Clarion-Clipperton Zone. Technical Report AMC Project 321012, 338 pp.

AMC Consultants, 2021b. TOML Mineral Resource, Clarion-Clipperton Zone, Pacific Ocean. Technical Report AMC Project 321012, 223 pp.

Bakeman, R., Quera, V., 2011. Sequential Analysis and Observational Methods for the Behavioral Sciences. Cambridge University Press.

Baturin, G.N., Dubinchuck, V.T., 2011. Mineralogy and chemistry of ferromanganese crusts from the Atlantic Ocean. Geochem. Int. 48 (6), 578–593.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Calas, G., 2017. Mineral resources and sustainable development. Elements: Int. Mag. Mineral. Geochem. Petrol. 13 (5), 301–306.

Carranza, E.J.M., Laborte, A.G., 2015. Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm. Ore Geol. Rev. 71, 777–787.

Carranza, E.J.M., Laborte, A.G., 2016. Data-driven predictive modeling of mineral prospectivity using random forests: a case study in Catanduanes Island (Philippines). Nat. Resour. Res. 25 (1), 35–50.

Charette, M.A., Smith, W.H., 2010. The volume of Earth's ocean. Oceanography 23 (2), 112–114.

Charles, C., Pelleter, E., Révillon, S., Nonnotte, P., Jorry, S.J., Kluska, J.-M., 2020. Intermediate and deep ocean current circulation in the Mozambique Channel: New insights from ferromanganese crust Nd isotopes. Mar. Geol. 430, 106356.

Conrad, T., Hein, J.R., Paytan, A., Clague, D.A., 2017. Formation of Fe-Mn crusts within a continental margin environment. Ore Geol. Rev. 87, 25–40.

Diesing, M., 2020. Deep-sea Sediments of the Global Ocean Mapped with Random Forest Machine Learning Algorithm. PANGAEA.

Frank, M., Whiteley, N., Kasten, S., Hein, J.R., O'Nions, K., 2002. North Atlantic Deep Water export to the Southern Ocean over the past 14 Myr: Evidence from Nd and Pb isotopes in ferromanganese crusts. Paleoceanography 17 (2), 12-1.

Frazer, J., Fisk, M., 1981. Scripps Institution of Oceanography Ferromanganese Nodule Analysis File - IDOE Portion. NOAA National Centers for Environmental Information.

Friedrich, G., Schmitz-Wiechowski, A., 1980. Mienralogy and chemistry of a ferromanganese crust from a deep-sea hill, Central Pacific, "Vladivia" Cruise VA 13/2. Mar. Geol. 37, 71–90.

Gao, Y., Zhang, Z., Xiong, Y., Zuo, R., 2016. Mapping mineral prospectivity for Cu polymetallic mineralization in southwest Fujian Province, China. Ore Geol. Rev. 75, 16–28.

Garcia, H.E., Locarini, R.A., Boyer, T.P., Antonoc, J.I., 2006. World Ocean Atlas 2005, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation. In: Levitus, S. (Editor), NOAA Atlas NESDIS 63, U.S. Government Printing Office, Washington DC.

Gazis, I.Z., Schoening, T., Alevizos, E., Greinert, J., 2018. Quantitative mapping and predictive modeling of Mn nodules' distribution from hydroacoustic and optical AUV data linked by random forests machine learning. Biogeosciences 15, 7347–7377.

GEBCO Compilation Group, 2021. GEBCO 2021 Grid.

GeoERA - MINDeSEA, 2019. Seabed Mineral Deposits in European Seas: Metallogeny and Geological Potential for Strategic and Critical Raw Materials. In: GeoERA - MINDeSEA (Editor), https://data.geus.dk/egdi/?mapname=egdi_geoera_mindesea#baslay=baseMapGEUS&extent=-1116610,91370,10904830,6667390&layers=emodnet_mineral_occurrences.

Graw, J., Wood, W., Phrampus, B., 2021. Predicting global marine sediment density using the random forest regressor machine learning algorithm. J. Geophys. Res.: Solid Earth, 126(1): e2020JB020135.

Guan, Y., Sun, X., Jiang, X., Sa, R., Zhou, L.i., Huang, Y.i., Liu, Y., Li, X., Lu, R., Wang, C., 2017. The effect of Fe-Mn minerals and seawater interface and enrichment mechanism of ore-forming elements of polymetallic crusts and nodules from the South China Sea. Acta Oceanol. Sin. 36 (6), 34–46.

Hariharan, S., Tirodkar, S., Porwal, A., Bhattacharya, A., Joly, A., 2017. Random forest-based prospectivity modelling of Greenfield terrains using sparse deposit data: an example from the Tanami Region, Western Australia. Nat. Resour. Res. 26 (4), 489–507.

Harris, P.T., Macmillan-Lawler, M., Rupp, J., Baker, E.K., 2014. Geomorphology of the oceans. Mar. Geol. 352, 4–24.

Hein, J.R., Conrad, T.A., Dunham, R.E., 2009. Seamount characteristics and mine-site model applied to exploration- and mining-lease-block selection for cobalt-rich ferromanganese crusts. Mar. Georesour. Geotechnol. 27 (2), 160–176.

Hein, J.R., Mizell, K., Koschinsky, A., Conrad, T.A., 2013. Deep-ocean mineral deposits as a source of critical metals for high- and green-technology applications: Comparison with land-based resources. Ore Geol. Rev. 51, 1–14.

Hein, J.R., Konstantinova, N., Mikesell, M., Mizell, K., Fitzsimmons, J.N., Lam, P.J., Jensen, L.T., Xiang, Y., Gartman, A., Cherkashov, G., Hutchinson, D.R., Till, C.P., 2017. Arctic deep water ferromanganese-oxide deposits reflect the unique characteristics of the Arctic Ocean. Geochem. Geophys. Geosyst. 18 (11), 3771–3800.

JAMSTEC, 2021. Japanese Agency for Marine-Earth Science and Technology Databases, https://www.jamstec.go.jp/e/database/.

JOGMEC, 2020. JOGMEC Conducts World's First Successful Excavation of Cobalt-Rich Seabed in the Deep Ocean; Excavation Test Seeks to Identify Best Practices to Access Essential Green Technology Ingredients While Minimizing Environmental Impact, https://www.jogmec.go.jp/english/news/release/content/300368332.pdf.

Josso, P., Parkinson, I., Horstwood, M., Lusty, P., Chenery, S., Murton, B., 2019. Improving confidence in ferromanganese crust age models: A composite geochemical approach. Chem. Geol. 513, 108–119.

Josso, P., Horstwood, M.S.A., Millar, I.L., Pashley, V., Lusty, P.A.J., Murton, B., 2020a. Development of a correlated Fe-Mn crust stratigraphy using Pb and Nd Isotopes and its application to paleoceanographic reconstruction in the Atlantic. Paleoceanogr. Paleoclimatol. 35 (10).

Josso, P., Rushton, J., Lusty, P., Matthews, A., Chenery, S., Holwell, D., Kemp, S.J., Murton, B., 2020b. Late Cretaceous and Cenozoic paleoceanography from north-east Atlantic ferromanganese crust microstratigraphy. Mar. Geol. 422, 106122.

Josso, P., Lusty, P., Chenery, S., Murton, B., 2021. Controls on metal enrichment in ferromanganese crusts: Temporal changes in oceanic metal flux or phosphatisation? Geochim. Cosmochim. Acta 308, 60–74.

Konstantinova, N., Cherkashov, G., Hein, J.R., Mirão, J., Dias, L., Madureira, P., Kuznetsov, V., Maksimov, F., 2017. Composition and characteristics of the ferromanganese crusts from the western Arctic Ocean. Ore Geol. Rev. 87, 88–99.

Konstantinova, N., Hein, J.R., Mizell, K., Cherkashov, G., Dreyer, B., Hutchinson, D.R., 2020. Changes in sediment source areas to the Amerasia Basin, Arctic Ocean, over the past 5.5 million years based on radiogenic isotopes (Sr, Nd, Pb) of detritus from ferromanganese crusts. Mar. Geol. 428, 106280.

Konstantinova, N.P., Khanchuk, A.I., Mikhailik, P.E., Skolotnev, S.G., Ivanova, E.V., Bich, A.S., Cherkashev, G.A., 2021. Ferromanganese crusts of the doldrums fracture zone, Central Atlantic: new data on the chemical composition. Dokl. Earth Sci. 496 (2), 125–129.

Koschinsky, A., Hein, J.R., 2017. Marine ferromanganese encrustations: archives of changing oceans. Elements 13 (3), 177–182.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.

LeDell, E., Poirier, S., 2020. H2o automl: Scalable automatic machine learning, Proceedings of the AutoML Workshop at ICML.

Li, T., Xia, Q., Zhao, M., Gui, Z., Leng, S., 2020. Prospectivity mapping for tungsten polymetallic mineral resources, Nanling Metallogenic Belt, South China: Use of random forest algorithm from a perspective of data imbalance. Nat. Resour. Res. 29 (1), 203–227.

Lisniowski, M.A. et al., 2019. Multibeam and video data applied to seabed mapping in the Rio grande rise, SW Atlantic, GEOHAB Marine Geological Biological Habitat Mapping Conference, Saint-Petersburg.

Lusty, P., Hein, J.R., Josso, P., 2018. Formation and occurrence of ferromanganese crusts: Earth's storehouse for critical metals. Elements 14 (5), 313–318.

Lusty, P.A.J., Murton, B.J., 2018. Deep-ocean mineral deposits: metal resources and windows into earth processes. Elements 14 (5), 301–306.

Lusty, P.A.J., Scheib, C., Gunn, A.G., Walker, A.S.D., 2012. Reconnaissance-scale prospectivity analysis for gold mineralisation in the Southern Uplands-Down-Longford Terrane, Northern Ireland. Nat. Resour. Res. 21 (3), 359–382.

Manheim, F.T., Lane-Bostwick, C.M., 1988. Chemical Composition Of Ferromanganese Crusts in the World Ocean: A Review and Comprehensive Database. U.S. Geological Survey Open-File Report; 89-20,1988. .

McKay, G., Harris, J., 2016. Comparison of the data-driven random forests model and a knowledge-driven method for mineral prospectivity mapping: A case study for gold deposits around the Huritz Group and Nueltin Suite, Nunavut, Canada. Nat. Resour. Res. 25 (2), 125–143.

Mizell, K., Hein, J.R., Lam, P.J., Koppers, A.A.P., Staudigel, H., 2020. Geographic and Oceanographic Influences on Ferromanganese Crust Composition Along a Pacific Ocean Meridional Transect, 14 N to 14S. Geochem. Geophys. Geosyst., 21(2): e2019GC008716.

Mizell, K., Hein, J.R., Au, M., Gartman, A., 2022. Estimates of metals contained in abyssal manganese nodules and ferromanganese crusts in the global ocean based on regional variations and genetic types of nodules. In: Sharma, R. (Ed.), Perspectives on Deep-Sea Mining: Sustainability, Technology, Environmental Policy and Management. Springer International Publishing, Cham, pp. 53–80.

NASA Ocean Biology (OB.DAAC), 2014. Mean annual sea surface chlorophyll-a concentration for the period 2009-2013 (composite dataset created by UNEP-WCMC). Data obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua Ocean Colour website (NASA OB.DAAC, Greenbelt, MD, USA), Cambridge (UK): UNEP World Conservation Monitoring Centre.

Neettiyath, U. et al., 2022. Automatic Detection of Buried Mn-crust Layers Using a Sub-bottom Acoustic Probe from AUV Based Surveys, OCEANS 2022 - Chennai, pp. 1-7.

Parsa, M., Maghsoudi, A., Yousefi, M., 2018. Spatial analyses of exploration evidence data to model skarn-type copper prospectivity in the Varzaghan district, NW Iran. Ore Geol. Rev. 92, 97–112.

Petersen, S., Krätschell, A., Augustin, N., Jamieson, J., Hein, J.R., Hannington, M.D., 2016. News from the seabed – Geological characteristics and resource potential of deep-sea mineral resources. Mar. Policy 70, 175–187.

Ren, Y., Sun, X., Guan, Y., Xiao, Z., Liu, Y., Liao, J., Guo, Z., 2019. Distribution of rare earth elements plus yttrium among major mineral phases of marine Fe–Mn crusts from the South China Sea and Western Pacific Ocean: A comparative study. Minerals 9 (1), 8.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818.

Staszak, P., Collot, J., Josso, P., Pelleter, E., Etienne, S., Patriat, M., Cheron, S., Boissier, A., Guyomard, Y., 2022. Origin and composition of ferromanganese deposits of New Caledonia exclusive economic zone. Minerals 12 (2), 255.

Straume, E.O., Gaina, C., Medvedev, S., Hochmuth, K., Gohl, K., Whittaker, J.M., Abdul Fattah, R., Doornenbal, J.C., Hopper, J.R., 2019. GlobSed: updated total sediment thickness in the World's Oceans. Geochem. Geophys. Geosyst. 20 (4), 1756–1772.

Usui, A., Nishi, K., Sato, H., Nakasato, Y., Thornton, B., Kashiwabara, T., Tokumaru, A., Sakaguchi, A., Yamaoka, K., Kato, S., Nitahara, S., Suzuki, K., Iijima, K., Urabe, T., 2017. Continuous growth of hydrogenetic ferromanganese crusts since 17 Myr ago on Takuyo-Daigo Seamount, NW Pacific, at water depths of 800–5500 m. Ore Geol. Rev. 87, 71–87.

Verlaan, P.A., Cronan, D.S., 2022. Origin and variability of resource-grade marine ferromanganese nodules and crusts in the Pacific Ocean: A review of biogeochemical and physical controls. Geochemistry 82 (1), 125741.

Wang, J., Zuo, R., Xiong, Y., 2020. Mapping mineral prospectivity via semi-supervised random forest. Nat. Resour. Res. 29 (1), 189–202.

Williams, C. et al., in. prep. Capturing exposed bedrock in the upland regions of Great Britain: A geomorphometric focussed random forest approach. Earth Surf. Process. Landforms.

Xiang, J., Xiao, K., Carranza, E.J.M., Chen, J., Li, S., 2020. 3D mineral prospectivity mapping with random forests: A case study of Tongling, Anhui, China. Nat. Resour. Res. 29 (1), 395–414.

Yeo, I.A., Dobson, K., Josso, P., Pearce, R.B., Howarth, S.A., Lusty, P.A.J., Bas, T.P.L., Murton, B.J., 2018. Assessment of the mineral resource potential of Atlantic ferromanganese crusts based on their growth history, microstructure, and texture. Minerals 8 (8), 327.

Yeo, I.A., Howarth, S.A., Spearman, J., Cooper, A., Crossouard, N., Taylor, J., Turnbull, M., Murton, B.J., 2019. Distribution of and hydrographic controls on ferromanganese crusts: Tropic Seamount, Atlantic. Ore Geol. Rev. 114, 103131.

Zhang, Z., Zuo, R., Xiong, Y., 2016. A comparative study of fuzzy weights of evidence and random forests for mapping mineral prospectivity for skarn-type Fe deposits in the southwestern Fujian metallogenic belt, China. Sci. China Earth Sci. 59 (3), 556–572.

Zhong, Y.i., Chen, Z., González, F.J., Hein, J.R., Zheng, X., Li, G., Luo, Y., Mo, A., Tian, Y., Wang, S., 2017. Composition and genesis of ferromanganese deposits from the northern South China Sea. J. Asian Earth Sci. 138, 110–128.

Zuo, R., Carranza, E.J.M., 2023. Machine learning-based mapping for mineral exploration. Math. Geosci.