Prediction of seismic p-wave velocity using machine learning (se-2019-58)

Ines Dumke and Christian Berndt

Responses to referee comments

Referee comments are given in black, author comments (**AC**) and changes in the manuscript (**CM**) in blue

Referee #1: Taylor Lee

General comments

Machine learning has been previously well established in other fields, but has not grasped attention in a similar way within the geosciences. This paper uses sparse p-wave velocity data from DSDP/ODP/IODP as training data in a machine learning algorithm (Random Forest) to predict p-wave velocity with depth. A thorough analysis was done to determine how effective machine learning is at predicting vertical velocity profiles. This analysis included comparison of p-wave velocity machine learning predictions with empirical estimates. A variety of appropriate methods were tested to improve the machine learning prediction (e.g. smoothing input data and prediction results, varying max_features and number of predictors used, 10-fold cross validation, predictor value scaling). As a result, this work provides valuable information on types of useful predictors and variables highly correlated to p-wave velocity. Additionally, this method shows in some case superior to using strictly empirical methods to estimate p-wave velocity with depth.

Results show this work is novel and useful. However, there is a major component of the analysis missing. This work contains many examples of validation of previously existing p-wave velocity but lacks demonstration on prediction of p-wave velocity in areas where no velocity data is available.

AC: As we explain in the Methods section, due to our leave-location-out approach all predictions are made for locations that were withheld from the training data and therefore act as unknown locations. Validation of the prediction involved a comparison against the true v_p data, but these data were in no case part of the prediction model. We explain below (last "specific comment") why we refrain from making predictions for completely new locations as it is beyond the scope of this paper, i.e. the purpose of this paper is to demonstrate the method and to discuss its advantages and limitations. When more training data become available the method can be used to make predictions elsewhere – probably first for limited areas and then globally.

Specific comments

Page 3 Section 2.1.2 (Predictors) Line 28 mentions that the continental crust was set at 1 billion years to represent significant older crust than that of the oceanic crust. If all the observed data (DSDP/ODP/IODP) are on oceanic crust, what is the importance/meaning of defining continental crust age?

AC: It is not true that all the data are from sites above oceanic crust. In fact, 142 of the 333 boreholes – 42% – were drilled on continental crust, e.g. in continental shelf regions. As the thermal regime of continental crust is different to that of oceanic crust – with old continental

crust being of lower temperatures than young oceanic crust –, which affects density and hence p-wave velocity, we thought it reasonable to differentiate between the two types of crust and their ages.
CM: no changes made in the manuscript

Page 7 Section 3.3 (Predictor importance) Line 20 states that categorical predictors generally do not have any importance in prediction performance. Additionally, it is again discussed in the discussion section (Section 4.2- lines 8-14 page 9). What is the variance of your sampled data set in categorical predictors? For example, for a given test data set (i.e. fold) are all of your categorical predictors for that run a 1 or 0? If all of your test data set has only one categorical value then that predictor would be of no importance.
AC: We do not claim that categorical predictors "generally" do not have "any importance" in prediction performance. In the referenced line (now p. 8 line 10-11), we use the term "negligible importance", i.e. almost zero, and we explain that this only refers to the results of our own study, not to studies involving categorical predictors in general.
The number of boreholes per predictor (for which the predictor is 1) varies between 2 (0.6%) and 191 (57%), on average, it is 42 (12.7%). We therefore agree that predictors with a very low representation will also be of low importance, and that this should be added as an explanation in the Discussion.
CM: We added the following sentence to the end of section 2.1 in the Methods: "Across the categorical predictors, the number of boreholes for which a predictor was set to 1 varied between 2 (0.6 %) and 191 (57.4 %); on average, the geological setting represented by a categorical predictor applied to 42 boreholes (12.7 %)."
We also included the sentence "The poor representation of some predictors, such as "cold_vent", "mud_volcano" and "hydroth_vent" in the dataset, causing these predictors to be 0 for all boreholes in some test folds, may likely explain the low importance of these predictors in the predictor ranking." in the last paragraph of section 4.2 in the Discussion (p. 10 lines 2-5).

Consider, if true, explicitly stating that predictions of this kind have not done with depth before. (page 2 ~ lines 16-20)
AC: As far as we know, predictions with depth have not been done before, and we agree that this should be stated in the text.
CM: We added a sentence to this paragraph (lines 21-23): "These studies were in general restricted to the prediction of one value per geographic location; the prediction of multiple values, such as depth profiles, has, to our knowledge, not been attempted before."

Minor suggestion to add in the abstract that this method is not designed to capture high variance in a p-wave velocity profile, but is instead intended to capture the overall trend of p-wave velocity profile.
AC: We agree that this should already be stated in the Abstract.
CM: We changed the sentence in line 9-10 (now lines 10-11) to read: "Here, we present a machine learning approach to predict the overall trend of seismic p-wave velocity (v_p) as a function of depth (z) for any marine location."

It is stated and supported (Line 1 page 7; Figure 3) that the RFE CV 16 predictors prediction (green) is better than CV, max_features =22, 38 predictors however the error in the prediction is significantly higher for the green prediction with roughly the same % boreholes labelled as

"good". Why do you consider green prediction to be so much better than yellow prediction? It might be useful if you explicitly state what your ultimate metric of correctness is (e.g. highest % correct or lowest error?)

AC: We did not mean to imply that one of the two runs provides better results than the other, and we also do not claim this anywhere. We merely stated the differences. However, we agree that this could have been easily misunderstood due to our ill use of the term "performance" – we meant performance to refer to both the highest % correct and the lowest error (i.e., in the same model), but we seem to have used it in other ways too, which must have been confusing. We now explain in more detail what we mean by performance and how our predictions were evaluated.

CM: Paragraph 3 in the Methods section 2.3 was changed to read "Performance of the RF model was evaluated in two ways: (1) by standard error metrics and (2) by the proportion of boreholes with predicted $v_p(z)$ superior to that of empirical functions. The standard error metrics root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R2) were calculated based on the comparison of the predicted and true $v_p(z)$ curves for each borehole in the test fold. RMSE, MAE and R2 of all test folds were then averaged to give final performance values."

Throughout the manuscript, we also replaced the term "performance" where necessary, to make its use consistent. In the sentence reference above (now lines 19-22), we replaced "performance" by "prediction scores".

What is the final global spatial resolution? E.g. prediction of p-wave velocity profile every 1-degree, 5-min, etc.?

AC: We do not want to go so far as to give a final global spatial resolution for the prediction of v_p. Our main aim was to investigate if it is at all possible to achieve realistic predictions of $v_p(z)$. We have shown that this is generally the case, however, our results also clearly indicate that more input data are required to overcome low prediction performance due to lack of suitable data. For this reason, we think that the prediction model needs to be improved further before a "final resolution" should be given. – In any case, one final resolution value likely would not be sufficient. Due to the heterogeneous depth distribution of the boreholes used (in addition to the heterogeneous spatial distribution), the resolution would vary with depth. Thus, separate resolution values would need to be determined for different depths (here: range 0-2500 m), which would likely be confusing and not very helpful for the reader.

CM: no changes made in the manuscript

Page 9 Section 4.2 (Most important predictors for the prediction of vp(z)) Lines 2-8 discuss how certain predictors are not used (porosity, density, pressure) as not all boreholes have depth associated measurements. However, some of the predictors used in the prediction do not have a depth component (e.g. crustage). Applying this logic, why do you not use seafloor porosity (i.e. depositional porosity) or likewise predictors?

AC: We did not mean that we can only use predictors with depth measurements, we obviously also used depth-independent predictors. The point here (which was not well explained in the text) was that we could only use predictors that were available (or could be determined) for every borehole location. This did not apply to many of the e.g. porosity measurements, which had been measured in boreholes (with a depth component) but often not at the borehole locations at which v_p had been measured. Even in the relatively few boreholes where both porosity (or density, pressure etc) and v_p had been measured, the depth ranges did not always

match - so there would have been depths with v_p data but no porosity data. It was impossible to also account for such cases, which is why we decided to leave these parameters out. As reviewer 2 also asked for an explanation regarding choice of predictors, we clarified this in the Methods section 2.1.2.
We also agree that a parameter like seafloor porosity, which is available as a global grid (we are assuming that the reviewer is referring to the grid by Martin et al., 2015), could easily have been added as a predictor. We did not do this at the time, and we hope the reviewer will understand that it is now too late to add new predictors to our study – as we state in section 4.2, there are several other predictors that could potentially be added, but this would have to be done in a future study.
CM: We have clarified our choice of predictors by adding the following passage to the Methods section 2.1.2 (p. 4 lines 5-13): "… These predictors were parameters that were assumed to influence p-wave velocity. However, only predictors that could be obtained for each of the 333 borehole locations were used. Predictors such as latitude (lat), longitude (long), and water depth (wdepth) were taken from the borehole's metadata, whereas other predictors were extracted from freely available global datasets and grids (Table 1). In addition, predictors describing the borehole's geological setting were determined from the site descriptions given in the proceedings of each drilling campaign. Some parameters known to influence seismic velocity – e.g. porosity, density, or pressure – had to be left out as suitable datasets were not available. Although some of these parameters had been measured in DSDP, ODP and IODP boreholes, they had not necessarily been logged at the same locations and depths at which v_p data had been measured, and therefore could not be obtained at all of the 333 boreholes used."

No supplemental material was provided for the global prediction of p-wave velocity with depth. This paper should include the final global prediction of p-wave velocity with depth.
AC: No, we do not agree. As with the final spatial resolution, providing a final global prediction of v_p at this stage (i.e. when the prediction model still requires optimization and is therefore not final yet) is neither feasible nor helpful. In fact, it would maybe give this method a bad reputation to deploy it prematurely. Furthermore, we show that one "final global prediction" would not be sufficient. We assume the reviewer expects a global map of final prediction values, similar to Fig. 4 in Taylor et al. (2019) or Fig. 1c in Martin et al. (2015). While such a map may be useful in cases with only one prediction value per location, in our case – taking into account the depth component of the predicted v_p – a whole range of prediction maps would seem necessary, one for each depth. However, none of these maps would be of much use on its own. It would only show the variation of velocity at a certain depth, but we are interested in the variation (or trend) of velocity with depth (i.e., a profile), which is much better illustrated by the predicted $v_p(z)$ profiles (of which we show sufficient examples). Thus, we do not think a final global prediction is useful.
CM: no changes made in the manuscript

Technical corrections
Page 8 delete "the" on line 21: "by the at least 60% of test locations"
CM: deleted "the" (now p. 9 line 8)

Page 8 line 3 consider changing "our results show that vp(z) profiles" to "our results show that the general trend of vp(z) profiles"
CM: We changed this sentence accordingly. (now line 24)

Page 16 Figure 2 caption (e) change “less good” to different word (substandard?)
CM: We changed this to “lower-quality prediction”.

Page 23 table 3, change words so they have consistent capitalization between table columns (e.g. Long and long)
CM: We changed the capitalized letters accordingly (also in Table 2).

Page 12 Lee et al., 2019 citation is missing the publication year.
AC: Sorry, this paper was fully published just before we submitted our manuscript and we forgot to update the reference correctly.
CM: Added publication year.

Referee #2: Anonymous referee

The manuscript, " Prediction of seismic p-wave velocity using machine learning", is a well-written description of a machine learning method –Random Forests – to predict seismic p-wave velocity as a function of depth for any a generic marine location. This manuscript is suitable for Copernicus, but the manuscript needs to be revised before it can be accepted. I have some suggestions here.

1 Introduction.
Page 2:
L24: You make the statement that the most widely used machine learning methods are ANNs, SVMs, and RFs. It is hard to convince people that these three algorithms are the most widely used. For specific problems, some algorithms may be more common than the other algorithms. You may say that the most widely used machine learning includes ANNs, SVMs, and RFs.
AC: We agree with the reviewer that this is probably problem-dependent and should not be generalized here.
CM: We replaced “are” by “include”. (now line 26)

L31: You mentioned that RF has been repeatedly found superior to other machine learning methods. You need to specify the particular problems that RF has been found superior to “other machine learning methods” in the text. And what other machine learning methods do you mean here? Please specify in the text.
AC: We have given more details on the particular studies and algorithms tested.
CM: We added 2 sentences after the first sentence of this paragraph (now p.3 lines 1-5): “For example, Li et al. (2011) tested 23 machine learning algorithms – including RF, SVM, and kriging methods – to predict mud content in marine sediments, and found that RF, along with RF combined with ordinary kriging or inverse distance squared, provided the best prediction results. Cracknell and Reading (2014) applied five machine learning methods to lithology classification of multispectral satellite data and reported higher classification accuracy for RF than for Naive Bayes, SVM, ANN, and k-Nearest Neighbors.”

2 Methods
Page 3, section 2.1.2:

L25: How do you come up with these 38 predictors? Could you specify the reason why you choose these 38 predictors in this section?
AC: We agree that further information would be helpful in this section.
CM: We shortened the first sentence to "A total of 38 geological and spatial variables were included as predictors (Table 1)." and added the following passage: "These predictors were parameters that were assumed to influence p-wave velocity. However, only predictors that could be obtained for each of the 333 borehole locations were used. Predictors such as latitude (lat), longitude (long), and water depth (wdepth) were taken from the borehole's metadata, whereas other predictors were extracted from freely available global datasets and grids (Table 1). In addition, predictors describing the borehole's geological setting were determined from the site descriptions given in the proceedings of each drilling campaign. Some parameters known to influence seismic velocity – e.g. porosity, density, or pressure – had to be left out as suitable datasets were not available. Although some of these parameters had been measured in DSDP, ODP and IODP boreholes, they had not necessarily been logged at the same locations at which v_p data had been measured, and therefore could not be obtained at all of the 333 boreholes used."

Page 4, section 2.2:
L14: How do you define "performance"? I saw you mentioned performance in the later section 2.3. But it is better to define that when you first mention that.
AC: We agree that it is not always clear what we mean by the term "performance", and we also used it inconsistently to refer to the standard error metrics, the proportion of well predicted boreholes, or both – this is obviously confusing. By performance, we mean both the error metrics and the proportion of well predicted boreholes. We now explain this in paragraph 3 of section 2.3 (p. 5 lines 20-25) and removed/replaced the term in the previous paragraphs.
CM: Paragraph 3 was changed to read "Performance of the RF model was evaluated in two ways: (1) by standard error metrics and (2) by the proportion of boreholes with predicted vp(z) superior to that of empirical functions. The standard error metrics root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R2) were calculated based on the comparison of the predicted and true vp(z) curves for each borehole in the test fold. RMSE, MAE and R2 of all test folds were then averaged to give final performance values." Throughout the manuscript, we replaced the term "performance" where necessary.

In addition, why do you choose 1000 trees? what is the maximum depth of each tree? How does the number of trees and depth affect the bias and variance of the prediction?
AC: Our study of relevant literature showed that most studies used either 500 or 1000 trees. In an early version of our prediction model, we ran RF repeatedly for numbers of trees between 2 and 1500 and evaluated model performance based on the OOB (out-of-bag) score. As the performance still varied after 500 trees but stabilized around 1000 trees, we chose 1000 trees. We did not repeat this procedure with our final prediction model (which no longer used the OOB approach), so it is possible that a lower number of trees might already have been sufficient. In that case, however, a higher number would not have decreased model performance. The depth of the trees was not defined and therefore not varied in the final prediction model. Early tests showed that performance was generally worse when maximum tree depths were specified (e.g. for max_depth = 5).
CM: no changes made in the manuscript

3 Results

Page 6, section 3.1:

The performance of an algorithm should be shown by both bias and variance. I only see the comparison of errors and percentage of boreholes with scores 2 and 3 in Fig. 3 and 4. How does the number of predictors and data smoothing affect the variance of the prediction?

AC: There is no strict rule that algorithm performance should always be evaluated by bias and variance. Many studies applying machine learning methods use other means to validate their results. We chose to evaluate performance by MAE, RMSE and R^2, which have been used as performance measures by several other studies that predicted environmental parameters (e.g. Gasch et al. (2015), Ließ et al. (2016), Meyer et al. (2015, 2016)). Our own borehole percentage value serves as an additional measure. We think that our performance evaluation approach is now well described in section 2.3. The effects of varying numbers of predictors and data smoothing in terms of prediction performance are already described in the text.

CM: see above for changes regarding clarification of prediction performance

Since you only have 333 boreholes, 2% change due to different model runs only change scores of 7 boreholes. I am curious about the location distributions of those boreholes which changed their scores, and why their scores changed by changing the number of predictors or data smoothing.

AC: Unfortunately, our applied prediction method does not allow determining which boreholes changed their scores across different model runs. We agree that this would be an interesting aspect to look into, but in this case our model cannot easily be adapted accordingly, so this would likely require setting up a completely new model. This is beyond the scope of these revisions.

CM: no changes made in the manuscript

4 Discussion

Page 10:

L1-5: You made a strong statement about performance of RF. As I suggested in your introduction section, the performance of a machine learning algorithm really depends on situations.

AC: We agree that this can also be misunderstood to mean that RF is always the perfect choice, which is of course not the case. What we actually meant to say was that due to the issues with our dataset (spatially inhomogeneous, varying depth ranges, etc), it is much more likely that the cases of poor performance are due to the dataset itself, and not due to the choice of machine learning algorithm.

CM: We rewrote the last sentence (now lines 31-33) to clarify this: "However, given the present dataset and its spatial inhomogeneity, we doubt that a different algorithm would lead to a significantly improved prediction performance for v_p."

5 Conclusion

Page 10:

L15: RF is hard to extrapolate to data outside the range they have been seen. I doubted that RF can be used for geophysical modeling in areas lacking v_p(z) from boreholes or seismic data.

AC: This is why we recommend more data to be added – to increase the data ranges within the RF model and the likelihood that when the RF model is applied to new data, these data are within the ranges known to RF. We agree that at present, this is not always the case, which likely

explains some of the lower-performing locations. However, we also point out that our RF model is not meant as a replacement for other sources of $v_p(z)$ data. It is meant only as an aid when no other means are available. We do not expect RF to ever replace or be superior (or even very close) to actual v_p measurements or v_p from seismic data (nor do we claim this in the manuscript). Our approach is only meant to provide an alternative to using an (unrealistic) constant velocity or empirically-derived $v_p(z)$ profiles, which are, as we show, often of lower quality than our predicted $v_p(z)$ profiles.

CM: no changes made in the manuscript

Prediction of seismic p-wave velocity using machine learning

Ines Dumke[1], Christian Berndt[1]

[1]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

Correspondence to: Christian Berndt (cberndt@geomar.de)

Abstract. Measurements of seismic velocity as a function of depth are generally restricted to borehole locations and are therefore sparse in the world's oceans. Consequently, in the absence of measurements or suitable seismic data, studies requiring knowledge of seismic velocities often obtain these from simple empirical relationships. However, empirically derived velocities may be inaccurate, as they are typically limited to certain geological settings, and other parameters potentially influencing seismic velocities, such as depth to basement, crustal age, or heatflow, are not taken into account. Here, we present a machine learning approach to predict the overall trend of seismic p-wave velocity (v_p) as a function of depth (z) for any marine location. Based on a training dataset consisting of v_p(z) data from 333 boreholes and 38 geological and spatial predictors obtained from publically available global datasets, a prediction model was created using the Random Forests method. In 60 % of the tested locations, the predicted seismic velocities were superior to those calculated empirically. The results indicate a promising potential for global prediction of v_p(z) data, which will allow improving geophysical models in areas lacking first-hand velocity data.

1 Introduction

Seismic p-wave velocities (v_p) and velocity-depth profiles are needed in many marine-geophysical applications, e.g. for seismic data processing, for time-depth conversions, or to estimate hydrate concentrations in gas hydrate modelling. Direct measurements of seismic velocities, however, are sparse and limited to borehole locations such as those drilled by the Deep Sea Drilling Project (DSDP), the Ocean Drilling Program (ODP), and the International Ocean Discovery Program (IODP).

Seismic velocities can also be obtained indirectly from seismic data. Approaches include derivation of 1D velocity profiles via refraction seismology using ocean bottom seismometers (OBS) (Bünz et al., 2005; Mienert et al., 2005; Westbrook et al., 2008; Plaza-Faverola et al. 2010a, 2010b, 2014), and velocity analysis of large-offset reflection seismic data (Crutchley et al., 2010, 2014; Plaza-Faverola et al., 2012). However, suitable seismic datasets are only available in certain areas, and OBS-derived velocity profiles are of relatively low spatial and vertical resolution.

In the absence of measurements and refraction seismic data, constant velocities are often used for time-depth conversions (e.g. Brune et al., 2010) or processing of reflection seismic data (Crutchley et al., 2010, 2011, 2013; Netzeband et al., 2010; Krabbenhoeft et al., 2013; Dumke et al., 2014), even though a constant velocity-depth profile is generally unrealistic and will thus lead to inaccurate results.

As an alternative, empirical velocity functions have been derived, which are based on averaged measurements and provide seismic velocity-depth relationships for different geological and geographical settings. For example, Hamilton (1979, 1980, 1985) used averaged v_p measurements from boreholes and sonobuoys to derive velocity-depth functions for different marine settings and sediment types. Velocities calculated from these empirical functions have been used e.g. for time-depth conversions (Lilly et al., 1993; Brune et al., 2010), brute stack processing of reflection seismic data, as well as local (Bünz et al., 2005) and regional (Scanlon et al., 1996; Wang et al., 2014) velocity models.

Although velocity profiles calculated from empirical functions may work well in some cases, empirical functions do not always produce accurate $v_p(z)$ profiles, due to their use of depth as the only input parameter and their limitation to certain regions or geological settings. Mienert et al. (2005) observed both agreements and disagreements between velocity profiles derived from OBS data and calculated from Hamilton functions, whereas Westbrook et al. (2008) argue that empirical functions are in general not representative for other areas due to variations in lithology and compaction history. Moreover, the Hamilton functions fail to provide correct velocities in areas containing gas hydrates or gas-saturated sediments (Bünz et al., 2005; Westbrook et al., 2008). Consequently, an alternative method is required to estimate $v_p(z)$ profiles for a larger variety of geological settings.

Over the last years, parameters in many different applications have been successfully predicted using machine learning techniques (e.g. Lary et al., 2016). In geosciences and remote sensing, machine learning methods have been used to predict soil properties (Gasch et al., 2015; Ließ et al., 2016; Meyer et al., 2018), air temperatures (Meyer et al., 2016a, 2018), biomass (Meyer et al., 2017), and the elasticity modulus of granitic rocks (Karakus, 2011). Applications also extended into marine settings, involving the prediction of sediment mud content off southwest Australia (Li et al., 2011), as well as parameters such as seafloor porosity (Martin et al., 2015; Wood et al., 2018), seafloor biomass (Wei et al., 2010), and seafloor total organic carbon (Wood et al., 2018; Lee et al., 2019) on a global scale. These studies were in general restricted to the prediction of one value per geographic location; the prediction of multiple values, such as depth profiles, has, to our knowledge, not been attempted before.

In machine learning, a prediction model is constructed from a training dataset consisting of the target variable to be predicted, and a set of predictor variables. A random subset of the data, the test set, is typically held back for testing and validation of the prediction model. The most widely used machine learning methods include~~are~~ Artificial Neural Networks (ANN; e.g. Priddy and Keller, 2005), Support Vector Machines (SVM; Vapnik, 2000), and Random Forests (RF; Breiman, 2001).

RF is an ensemble classifier based on the concept of decision trees, which are grown from the training set by randomly drawing a subset of samples with replacement (bagging or bootstrap approach) (Breiman, 2001). At each tree node, the data are split based on a random subset of predictor variables to partition the data into relatively homogenous subsets and maximize the differences between the offspring branches. Each tree predicts on all samples in the test set and the final prediction is obtained by averaging the predictions from all trees.

RF has been repeatedly found superior to other machine learning methods. For example, Li et al. (2011) tested 23 machine learning algorithms – including RF, SVM, and kriging methods – to predict mud content in marine sediments, and found that RF, along with RF combined with ordinary kriging or inverse distance squared, provided the best prediction results. Cracknell and Reading (2014) applied five machine learning methods to lithology classification of multispectral satellite data and reported higher classification accuracy for RF than for Naive Bayes, SVM, ANN, and k-Nearest Neighbors ~~(e.g. Li et al., 2011; Cracknell and Reading, 2014)~~. RF~~It~~ is robust to noise and outliers (Breiman, 2001), and it is also able to handle high-dimensional and complex data. Moreover, RF does not require any preprocessing of the input variables and provides variable importance measurements, making it the first choice method in many applications.

Here, we apply RF to predict seismic p-wave velocity-depth profiles on a global scale, based on a set of 38 geological and spatial predictors that are freely available from global datasets. Prediction performance is evaluated and compared to velocity-depth profiles calculated from empirical v_p functions. We also test additional methods for improvement of model performance and determine which predictors are most important for the prediction of v_p.

2 Methods

2.1 Dataset

2.1.1 $v_p(z)$ data

$v_p(z)$ profiles for training of the RF model were obtained from boreholes drilled by the DSDP, ODP and IODP campaigns between 1975 and 2016. All boreholes containing v_p measurements were used, excluding those with bad-quality logs according to the logging description notes. In total, 333 boreholes were included in the dataset, the distribution of which is shown in Fig. 1. All $v_p(z)$ data from these boreholes are available through http://www.iodp.org and were downloaded from the archive at http://mlp.ldeo.columbia.edu/logdb/scientific_ocean_drilling/.

A multitude of measuring methods and tools had been employed by the different drilling campaigns to obtain v_p measurements, including wireline logging tools (e.g. sonic digital tool, long-spacing sonic tool, dipole sonic imager, borehole compensated sonic tool) and logging-while-drilling tools (sonicVISION tool, ideal sonic-while-drilling tool). The majority of these methods provided v_p measurements at 0.15 m depth intervals. Lengths of the v_p logs varied greatly, ranging between 10 m and 1800 m (average: 370 m), with top depths of 0-1270 m (average: 138 m) and bottom depths of 16-2460 m (average: 508 m).

After exporting the $v_p(z)$ profiles for each borehole, the data were smoothed using a moving average filter with a window of 181 data points (corresponding to ca. 27 m for a 0.15 m depth interval). Smoothing was applied to remove outliers and to account for unknown and varying degrees of uncertainty associated with the different measurement tools. In addition, smoothing was expected to facilitate prediction, as the aim was to predict the general $v_p(z)$ trend at a given location, rather

than predicting exact v_p values at a certain depth. Following smoothing, the profiles were sampled to 5 m depth intervals, using the same depth steps in all boreholes.

2.1.2 Predictors

A total of 38 geological and spatial variables ~~obtained from the borehole metadata and freely available global datasets~~ were included as predictors (Table 1). These predictors were parameters that were assumed to influence p-wave velocity. However, only predictors that could be obtained for each of the 333 borehole locations were used. Predictors such as latitude (lat), longitude (long), and water depth (wdepth) were taken from the borehole's metadata, whereas other predictors were extracted from freely available global datasets and grids (Table 1). In addition, predictors describing the borehole's geological setting were determined from the site descriptions given in the proceedings of each drilling campaign. Some parameters known to influence seismic velocity – e.g. porosity, density, or pressure – had to be left out as suitable datasets were not available. Although some of these parameters had been measured in DSDP, ODP and IODP boreholes, they had not necessarily been logged at the same locations and depths at which v_p data had been measured, and therefore could not be obtained at all of the 333 boreholes used.

For predictor variables based on global grids, such as age of crust (crustage), sediment thickness (sedthick), and surface heatflow (heatflow), values were extracted for each borehole location in GMT (Wessel et al., 2013), using the command *grdtrack*. As the crustal age grid (Müller et al., 2008) contained only ages for oceanic crust, the age for locations above continental crust was set to 1 billion years to represent a significantly older age than that of oceanic crust. Depth to basement (depth2base) was calculated by subtracting the depth values from the (constant) sedthick value at each borehole location, so that depths below the basement were indicated by a negative depth2base value. The distance predictor variables, e.g. distance to the nearest seamount (dist2smt), were calculated based on the borehole location and the respective datasets (Table 1) via the GMT command *mapproject*.

Of the 38 predictors, 15 were of the type continuous, whereas 23 were categorical variables describing the type of crust and the geological setting at each borehole location (Table 1). The categorical predictors were encoded as either 0 or 1, depending on whether the predictor corresponded to the geological setting at a given borehole. Multiple categories were possible; for example, a borehole located in a fore-arc basin above continental crust would be described by 1 for the predictors "contcrust", "active_margin", "subduction" and "fore-arc", and 0 for all other categorical predictors. Across the categorical predictors, the number of boreholes for which a predictor was set to 1 varied between 2 (0.6 %) and 191 (57.4 %); on average, the geological setting described by a categorical predictor applied to 42 boreholes (12.7 %).

2.2 Random Forest implementation

RF was implemented using the *RandomForestRegressor* in Python's machine learning library scikit-learn (Pedregosa et al., 2011). Two parameters needed to be set: the number of trees (n_estimators) and the number of randomly selected predictors to consider for splitting the data at each node (max_features). Many studies used 500 trees (e.g. Micheletti et al., 2014;

Belgiu and Drăguţ , 2016; Meyer et al., 2017, 2018), but as performance still increased after 500 trees, we chose 1000 trees instead. The max_features parameter was initially set to all predictors (38), as recommended for regression cases (Pedregosa et al., 2011; Müller and Guido, 2017), although some studies suggest tuning this parameter to optimize model results (Micheletti et al., 2014; Ließ et al., 2016; Meyer et al., 2016b).

2.3 Model validation

A 10-fold cross-validation (CV), an approach frequently used in model validation (e.g. Li et al., 2011; Gasch et al., 2015; Ließ et al., 2016; Meyer et al., 2016b, 2018), was applied to validate the RF model. CV involved partitioning the dataset into ten equally sized folds. Nine of these folds acted as the training set used for model building, whereas the remaining fold was used for testing the model and evaluating the ~~performance~~predictions. This procedure was repeated so that each fold acted once as the test fold, and hence each borehole was once part of the test set. ~~Performances of all test folds were averaged to give a final model performance.~~

Partitioning into folds was not done randomly from all available data points but by applying a leave-location-out (LLO) approach (Gasch et al., 2015; Meyer et al., 2016a, 2018) in which the data remained separated into boreholes, i.e., locations, so that each fold contained 1/10 of the boreholes. With 33-34 boreholes per fold, the size of the training dataset thus varied between 20166 and 20784 data points. By using the LLO approach, whole locations were left out of the training set, thereby allowing the RF model to be tested on unknown locations through prediction of v_p for each borehole in the test fold. If the folds were chosen randomly from all data points, each borehole location would be represented in the training set by at least some data points, resulting in overoptimistic model performance due to spatial overfitting (Gasch et al., 2015; Meyer et al., 2016a, 2018).

Performance of the RF model was evaluated in two ways: (1) by standard error metrics and (2) by the proportion of boreholes with predicted $v_p(z)$ superior to that of empirical functions. ~~by comparing the predicted and true $v_p(z)$ curves for each borehole in the test fold and calculating t~~The standard error metrics root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2)~~.~~ were calculated based on the comparison of the predicted and true $v_p(z)$ curves for each borehole in the test fold. RMSE, MAE and R^2 of all test folds were then averaged ~~over the ten folds~~to give final performance values.

To determine the proportion of boreholes with better $v_p(z)$ trends than those from empirical functions, ~~In addition,~~ we ~~also~~ tested how well the predicted $v_p(z)$ curves performed compared to $v_p(z)$ curves calculated from empirical functions. Using the depth values of the respective test borehole, $v_p(z)$ profiles were therefore calculated from the five empirical functions presented by Hamilton (1985) for deep-sea sediments, i.e., for terrigenous silt and clays (termed H1 in the following), terrigenous sediments (H2), siliceous sediments (H3), calcareous sediments (H4), and pelagic clay (H5). These v_p functions were chosen because the deep-sea setting applied to the majority of the boreholes, or was the best choice in absence of empirical functions for other geological settings such as mid-ocean ridges. The resulting Hamilton curves were evaluated against the true $v_p(z)$ profile, and RMSE, MAE and R^2 were averaged over the five curves. The averaged error metrics were

then compared to the error metrics of the prediction, and each borehole was assigned a score between 0 and 3 as shown in Table 2. Scores 2 and 3 were interpreted as a good prediction, i.e., better than the Hamilton curves, whereas scores 0 and 1 represented generally bad predictions. The proportion of boreholes with good predictions was then~~,~~ averaged over the ten folds~~., served as another performance evaluation measure.~~

2.4 Predictor selection

To determine the most important predictors for v_p prediction, a predictor selection approach was performed. Although RF can deal with high data dimensionality, predictor selection is still recommended, not only to remove predictors that could cause overfitting but also to increase model performance (e.g. Belgiu and Drăguţ, 2016, and references therein). We applied Recursive Feature Elimination (RFE), which is based on the variable importance scores provided by the RF algorithm. After calculating and evaluating a model with all 38 predictors, the least important predictor according to the variable importance scores was removed and the model was calculated again. This procedure was repeated until only one predictor was left. By evaluating model performance for each run via CV, using the same ten folds as before, the optimum number of predictors was determined.

2.5 Tests to improve prediction performance

Additional tests to improve prediction performance included predictor scaling, variation of the max_features parameter, and stronger smoothing of the $v_p(z)$ curves. All models were evaluated via a 10-fold CV, using the same folds as in the previous model runs.

Predictor scaling was applied to account for the different data ranges of continuous and categorical features. Model performance may be negatively affected if different types of variables or data ranges are used (Otey et al., 2006; Strobl et al., 2007), even though RF does not normally require scaled input data. All continuous predictors were scaled to between 0 and 1 to match the range of the categorical predictors, and RFE was repeated.

As tuning of the max_features parameter, i.e., the number of predictors to consider at each split, is recommended by some studies (Ließ et al., 2016; Meyer et al., 2018), an additional model was run in which max_features was varied between 2 and 38 (all features) with an interval of 2. Performance was evaluated for each case to find the optimum number of predictors to choose from at each split.

A third attempt to improve model performance involved enhanced filtering of the $v_p(z)$ curves so that larger v_p variations were smoothed out and the curves indicated only a general trend, which would likely be sufficient for many applications requiring knowledge of v_p with depth. The v_p curves therefore underwent spline smoothing using Python's scipy function *UnivariateSpline*. Three separate RF models were calculated: (i) spline1, which involved spline smoothing of the predicted curve of each test borehole; (ii) spline2, in which the input $v_p(z)$ data were smoothed; and (iii) spline3, where both the input $v_p(z)$ curves and the predictions were smoothed. All three cases were run with the 16 most important predictors as determined from the RFE results, and compared to the previous models.

3 Results

3.1 Prediction performance

Overall, many $v_p(z)$ profiles were predicted well by the RF models. For the 38-predictor CV, about 59.5 % of the boreholes had prediction scores of 2 or 3, representing a prediction performance superior to that of the Hamilton functions.

Predictions of prediction score 3, which were characterised by lower RMSE and MAE values and a higher R^2 than the five empirical functions, often exhibited a good fit to the true $v_p(z)$ curve (Fig. 2a-d). Even for more complex velocity profiles, e.g. involving a velocity reduction at depth (Fig. 2d), or a strong increase such as that from 2.2 km s^{-1} to >4 km s^{-1} at the basement contact in Fig. 2b, the predicted $v_p(z)$ curves generally matched the true curves well. In some cases, score 3 predictions did not provide a good fit but still performed better than the empirical functions (Fig. 2e). Score 2 predictions generally indicated the correct trend of the true $v_p(z)$ profile (Fig. 2f), whereas score 1 and score 0 predictions failed to do so, with velocities often considerably higher or lower than the true velocities (Fig. 2g, h).

The RFE CV revealed best performance for 33 predictors, as indicated by the lowest RMSE and MAE values (Fig. 3a). The proportion of boreholes with prediction scores of 2 or 3 was 59.2 % and thus slightly lower than for the 38-predictor CV (59.5 %; Fig. 3b). The highest proportion of 61.9 % was achieved by the 16-predictor model (Fig. 3b), but this also led to the highest errors (Fig. 3a).

By scaling all predictors to between 0 and 1 and repeating RFE, RMSE and MAE were reduced further, with the best errors obtained for 35 predictors (Fig. 3a). These errors were only slightly lower than those of the 30-predictor case, which achieved a higher percentage of boreholes with good prediction (60.4 %; Fig. 3b).

Varying the number of predictors to consider for splitting the data at each tree node also improved the performance. For max_features = 22, RMSE and MAE were lower than in all previous RF cases (Fig. 3a), while the proportion of boreholes with good ~~performance~~ prediction scores was 61.3 % and thus only slightly lower than for the 16-predictor case in which all 38 predictors were considered (Fig. 3b).

The three attempts of stronger smoothing of the $v_p(z)$ profiles via splines resulted in overall worse performance than the 16-predictor case, both in terms of errors and proportion of well-predicted boreholes (Fig. 4a). An exception is the spline1 case (spline smoothing of the predicted $v_p(z)$ profile), for which 62.4 % of the boreholes had scores of 2 or 3 (Fig. 4b), although RMSE and MAE were slightly worse than for the other RF cases.

3.2 Score distribution

The global distribution of boreholes with different prediction scores, shown in Fig. 5 for the 16-predictor case without spline smoothing, did not indicate a clear separation into areas with relatively good (scores 2 and 3) or bad (scores 0 and 1) prediction ~~performance~~scores. Some areas contain clusters of >10 boreholes, many of which had a prediction score of 3. Examples included the Sea of Japan (area A in Fig. 5a), the Nankai Trough (B), the Ontong-Java Plateau (C), the

Queensland Plateau (D), and the Great Australian Bight (E). However, nearly all of these cluster areas also contained boreholes with bad prediction scores (Fig. 5b). Similarly, single boreholes in remote locations were often characterised by a prediction score of 0 (Fig. 5b), but there were also several remote boreholes with scores of 3, e.g. on the Mid-Atlantic Ridge (area F in Fig. 5a).

3.3 Predictor importance

For the 38-predictor CV, the five most important predictors were "depth2base", "crustage", "depth", "dist2smt", and "wdepth" (Fig. 6). Continuous predictors and categorical predictors were clearly separated in the predictor importance plot (Fig. 6), with continuous predictors being of high importance in the RF model, whereas categorical predictors appeared less important. The only exception was the categorical predictor variable "spreading_ridge", which had a slightly higher importance ranking than the continuous predictors "long" and "dist2transform". Many of the categorical predictors were of negligible (almost 0) importance (Fig. 6).

When the least important predictor was eliminated after each model run using RFE, the same trend was observed: in both the unscaled and scaled RFE cases, all categorical predictors were eliminated before the continuous predictors (Table 3). In the 16-predictor case, which had the highest proportion of well-predicted boreholes (61.9 %), the only categorical predictor included was "spreading_ridge".

In the unscaled RFE case, the five most important predictors were the same as in the feature importance plot of the 38-predictor case (Fig. 6). However, the order differed slightly, with "depth" being eliminated before "dist2smt", "wdepth", "depth2base", and "crustage" (Table 3). When using scaled predictors, the five top predictors included "heatflow" (ranked sixth in both the 38-predictor CV and unscaled RFE cases) instead of "crustage". "Crustage" dropped to position 15 and was thus the least important of the continuous predictors (Table 3). In general, however, the position ranking of most predictors varied only by up to five positions between the unscaled and the scaled RFE cases (Table 3).

4 Discussion

4.1 Prediction performance in comparison with empirical functions

Our results show that the general trend of $v_p(z)$ profiles can be predicted successfully using machine learning. Overall, the applied RF approach is superior to the empirical v_p functions of Hamilton (1985), as indicated by the 60 % of tested boreholes with prediction scores of 3 or 2. Although such a quantitatively better performance (i.e., lower RMSE and MAE, and higher R^2 than the Hamilton $v_p(z)$ profiles) does not always mean a perfect fit to the true $v_p(z)$ curve of the tested borehole, the RF approach has a promising potential for the prediction of v_p with depth.

Slight improvements of the prediction performance were achieved by applying RFE, resulting in a proportion of well-predicted boreholes of 61.9 % for the 16-predictor model. Smoothing the predicted $v_p(z)$ profiles via spline smoothing

(spline1 case) provided a further increase to 62.4 % of well-predicted boreholes. In addition, reducing the max_features parameter from 38 (all predictors) to 22 also resulted in a slight improvement (61.3 %), thus supporting other studies that recommended tuning the max_features parameter to improve results (Ließ et al., 2016; Meyer et al., 2018). However, to increase model performance even further, to a proportion of well-predicted boreholes well exceeding 60 %, other changes are required.

4.2 Most important predictors for the prediction of $v_p(z)$

Both the predictor importance ranking of RF and the RFE results revealed "depth" as one of the most important predictors. However, "depth" was not the most important predictor, which is surprising as empirical v_p functions, including those of Hamilton (1985), all use depth as the only input parameter. Our results showed that "depth2base" was always ranked higher than "depth", and often the predictors "wdepth", "dist2smt" and "crustage" also had higher importance scores than "depth". Although "depth" is obviously still an important parameter in the prediction of v_p, these observations imply that empirical functions using only depth as input and neglecting all other influences may not produce realistic v_p values, which is supported by ~~the~~ at least 60 % of test locations for which the RF approach produced better $v_p(z)$ profiles than the Hamilton functions.

The high importance of the predictors "depth2base", "wdepth", "dist2smt", "crustage", as well as "heatflow", seems reasonable. The depth to the basement, which is related to the sediment thickness, is relevant because of the rapid v_p increase at the basement contact and the associated transition from relatively low (<2.5 km s^{-1}) to higher (> 4 km s^{-1}) v_p values. Even though in the majority of boreholes, the basement was not reached, the depth to the basement strongly influences v_p. The high ranking of the distance to the neareast seamount is likely attributed to the associated change in heatflow at seamount locations. Higher heatflow and hence higher temperatures affect density, which in turn affects v_p. The predictor "crustage" indicates young oceanic crust, which is characterised by higher temperature and hence lower density, affecting v_p. Moreover, "crustage" differentiates between oceanic (<200 Myr) and continental (here: 1 Byr) crust, and apparently more effectively than the categorical predictors "oceancrust" and "contcrust", which are of considerably lower importance.

It has to be noted that the high-importance predictors discussed above only represent the most important of the 38 predictors used for prediction of v_p; they are not necessarily the parameters that most strongly influence v_p in general. If other parameters, such as porosity, density, pressure, or saturation, had been included as predictors, they would likely have resulted in a higher importance ranking than, e.g., "dist2smt" or "crustage". However, these parameters were not included in the model as they were restricted to measurements at borehole locations – not necessarily those from which $v_p(z)$ data were obtained – and are therefore not available for every location in the oceans. For the same reason, other geophysical parameters, e.g. electrical resistivity and magnetic susceptibility, were also not included.

A surprising finding in terms of predictor importance is the low importance of all categorical predictors. The clear separation between continuous and categorical predictors in the predictor importance plots may be due to biased predictor selection, as observed by Strobl et al. (2007) when different types of predictors were used. In such cases, categorical predictors may often

be neglected and ignored by the machine learning algorithm (Otey et al., 2006). Scaling the continuous predictors to the same range as the categorical predictors did not help to change the importance ranking, but bias cannot be excluded. The poor representation of some predictors, such as "cold vent", "mud volcano" and "hydroth vent" in the dataset, causing these predictors to be 0 for all boreholes in some test folds, may likely explain the low importance of these predictors in the predictor ranking. On the other hand, it is also possible that the geological setting described by the categorical predictors was simply not relevant to the prediction of v_p. This possibility appears to be supported by the RFE results, which reveal the best ~~performance~~ prediction scores (61.9 % well-predicted boreholes) when all but one categorical predictors were excluded (16-predictor case).

4.3 Suggestions for further improvement of performance

The fact that prediction performance could not be much improved by predictor selection, tuning the max_features parameter, or additional smoothing suggests that other measures are needed to further improve the prediction performance. The comparatively high proportion of boreholes with badly predicted $v_p(z)$ profiles (about 40 %) is likely due to the limited number of boreholes that were available in this study, but may also have been influenced by the choice of machine learning algorithm.

It is possible to add more predictors that potentially influence v_p, for example, seafloor gradient, bottom water temperature, and distance to the shelf edge. In addition, some of the predictors could be improved. For example, the age of the continental crust, currently set to the constant value of 1 Byr, could be adapted based on the crustal age grid by Poupinet and Shapiro (2009). Other studies also suggest including the first and second derivatives of predictors or other mathematical combinations of predictors (Obelcz and Wood, 2018; Wood et al., 2018; Lee et al., 2019).

Another way to extend the dataset is to include more $v_p(z)$ data. Given the relatively inhomogeneous global distribution of borehole locations used in this study (Fig. 1), adding more $v_p(z)$ data is highly recommended. On a much smaller scale, Gasch et al. (2015) noted that high spatial heterogeneity of input locations limits the prediction performance and increases prediction errors. Adding more $v_p(z)$ data, especially from regions such as the southern Pacific and Atlantic oceans that are presently not covered, will likely help to increase the prediction performance. For example, the $v_p(z)$ records from recent IODP expeditions may be added to the dataset as they become available. Additional v_p data could also be obtained from commercial boreholes and refraction seismic data from ocean bottom seismometers, although the latter would be of lower vertical resolution.

The choice of machine learning algorithm may also influence model performance. Studies comparing RF against other machine learning algorithms reported different trends: in some cases, RF was superior in terms of prediction performance (e.g. Li et al., 2011; Cracknell and Reading, 2014), whereas in other cases, no strong differences were observed between the different methods (e.g. Goetz et al., 2015; Meyer et al., 2016b). However, ~~G~~given the present dataset and its spatial inhomogeneity~~generally positive reputation of RF as a prediction method~~, we doubt that a different algorithm would lead to a significantly improved~~different~~ prediction performance for v_p.

5 Conclusions

In this study, we presented an RF model for the prediction of $v_p(z)$ anywhere in the oceans. In about 60 % of the tested locations, the RF approach produced better $v_p(z)$ profiles than empirical v_p functions. This indicates a promising potential for the prediction of $v_p(z)$ using machine learning, although some improvement is still required. In particular, the model input data should be extended to increase spatial coverage, which is expected to significantly improve prediction performance. Our results showed that depth, which is the only input in empirical v_p functions, is not the most important parameter for the prediction of v_p. Distance to the basement, water depth, age of crust, and distance to the nearest seamount are, in general, equally or even more important than depth. By including these parameters in the determination of v_p, the RF model is able to produce more accurate $v_p(z)$ profiles and can therefore be used as an alternative to empirical v_p functions. This is of particular interest for geophysical modelling applications, such as modelling gas hydrate concentrations, in areas lacking alternative $v_p(z)$ information from boreholes or seismic data.

Acknowledgements

This study was funded by the Helmholtz Association, grant ExNet-0021. We thank Taylor Lee and an anonymous reviewer for their constructive comments, which helped to improve the manuscript.

References

Belgiu, M., Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions, ISPRS J. Photogramm., 114, 24-31, doi:10.1016/j.isprsjprs.2016.01.011, 2016.

Breiman, L.: Random forests. Mach. Learn., 45, 5-32, doi:10.1023/A:1010933404324, 2001.

Brune, S., Babeyko, A. Y., Gaedicke, C., Ladage, S.: Hazard assessment of underwater landslide-generated tsunamis: a case study in the Padang region, Indonesia, Nat. Hazards, 53, 205-218, doi:10.1007/s11069-009-9424-x, 2010.

Bünz, S., Mienert, J., Vanneste, M., Andreassen, K.: Gas hydrates at the Storegga Slide: constraints from an analysis of multicomponent, wide-angle seismic data, Geophysics, 70, B19-B34, doi:10.1190/1.2073887, 2005.

Coffin, M. F., Gahagan, L. M., Lawver, L. A.: Present-day plate boundary digital data compilation, UTIG Technical Report No. 174, University of Texas Institute for Geophysics, Austin, TX, 1998.

Cracknell, M. J., Reading, A. M.: Geologial mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information, Comp. Geosci., 63, 22-33, doi:10.1016/j.cageo.2013.10.008, 2014.

Crutchley, G. J., Pecher, I. A., Gorman, A. R., Henrys, S. A., Greinert, J.: Seismic imaging of gas conduits beneath seafloor seep sites in a shallow marine gas hydrate province, Hikurangi Margin, New Zealand, Mar. Geol., 272, 114-126, doi:10.1016/j.margeo.2009.03.007, 2010.

Crutchley, G. J., Berndt, C., Klaeschen, D., Masson, D. G.: Insights into active deformation in the Gulf of Cadiz from new 3-D seismic and high-resolution bathymetry data, Geochem. Geophys. Geosyst., 12, Q07016. doi:10.1029/2011GC003576, 2011.

Crutchley, G. J., Karstens, J., Berndt, C., Talling, P. J., Watt, S., Vardy, M., Hühnerbach, V., Urlaub, M., Sarkar, S., Klaeschen, D., Paulatto, M, Le Friant, A., Lebas, E., Maeno, F.: Insights into the emplacement dynamics of volcanic landslides from high-resolution 3D seismic data acquired offshore Montserrat, Lesser Antilles, Mar. Geol., 335, 1-15, doi:10.1016/j.margeo.2012.10.004, 2013.

Crutchley, G. J., Klaeschen, D., Planert, L., Bialas, J., Berndt, C., Papenberg, C., Hensen, C., Hornbach, M. J., Krastel, S., Brueckmann, W.: The impact of fluid advection on gas hydrate stability: Investigations at sites of methane seepage offshore Costa Rica, Earth Planet. Sci. Lett., 401, 95-109, doi:10.1016/j.epsl.2014.05.045, 2014.

Davies, J. H.: Global map of solid Earth surface heat flow, Geochem. Geophys. Geosyst., 14, 4608-4622, doi:10.1002/ggge.20271, 2013.

Dumke, I., Berndt, C., Crutchley, G. J., Krause, S., Liebetrau, V., Gay, A., Couillard, M.: Seal bypass at the Giant Gjallar Vent (Norwegian Sea): Indications for a new phase of fluid venting at a 56-Ma-old fluid migration system, Mar. Geol., 351, 38-52, doi:10.1016/j.margeo.2014.03.006, 2014.

Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T., Brown, D. J.: Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D+T: The Cook Agronomy Farm data set, Spat. Stat., 14, 70-90, doi:10.1016/j.spasta.2015.04.001, 2015.

Goetz, J. N., Brenning, A., Petschko, H., Leopold, P.: Evaluating machine learning and statistical prediction techniques for landslide susceptibility modelling, Comp. Geosci., 81, 1-11, doi: 10.1016/j.cageo.2015.04.007, 2015.

Hamilton, E. L.: V_p/V_s and Poisson's ratios in marine sediments and rocks, J. Acoust. Soc. Am., 66, 1093-1101, doi:10.1121/1.383344, 1979.

Hamilton, E. L.: Geoacoustic modeling of the sea floor, J. Acoust. Soc. Am., 68, 1313-1337, doi:10.1121/1.385100, 1980.

Hamilton, E. L.: Sound velocity as a function of depth in marine sediments, J. Acoust. Soc. Am., 78, 1348-1355, doi:10.1121/1.392905, 1985.

Karakus, M.: Function identification for the intrinsic strength and elastic properties of granitic rocks via genetic programming (GP), Comp. Geosci., 37, 1318-1323, doi:10.1016/j.cageo.2010.09.002, 2011.

Kim, S.-S., Wessel, P.: New global seamount census from altimetry-derived gravity data, Geophys. J. Int., 186, 615-631, doi:10.1111/j.1365-246X.2011.05076.x, 2011.

Krabbenhoeft, A., Bialas, J., Klaucke, I., Crutchley, G., Papenberg, C., Netzeband, G. L.: Patterns of subsurface fluid-flow at cold seeps: The Hikurangi Margin, offshore New Zealand, Mar. Pet. Geol., 39, 59-73, doi:10.1016/j.marpetgeo.2012.09.008, 2013.

Lary, D. J., Alavi, A. H., Gandomi, A. H., Walker, A. L.: Machine learning in the geosciences and remote sensing. Geosci. Front., 7, 3-10, doi:10.1016/j.gsf.2015.07.003, 2016.

Lee, T. R., Wood, W. T., Phrampus, B. J.: A machine learning (kNN) approach to predicting global seafloor total organic carbon, Global Biochem. Cy., 33, 37-46, doi:10.1029/2018GB005992, 2019.

Li, J., Heap, A. D., Potter, A., Daniell, J. J.: Application of machine learning methods to spatial interpolation of environmental variables, Environ. Modell. Softw., 26, 1647-1659, doi:10.1016/j.envsoft.2011.07.004, 2011.

Ließ, M., Schmidt, J., Glaser, B.: Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches, PLoS ONE, 11, e0153673, doi:10.1371/journal.pone.0153673, 2016.

Lilley, F. E. M., Filloux, J. H., Mulhearn, P. J., Ferguson, I. J.: Magnetic signals from an ocean eddy, J. Geomagn. Geoelectr., 45, 403-422, doi: 10.5636/jgg.45.403, 1993.

Martin, K. M., Wood, W. T., Becker, J. J.: A global prediction of seafloor sediment porosity using machine learning, Geophys. Res. Lett., 42, 10640-10646, doi:10.1002/2015GL065279, 2015.

Meyer, H., Katurji, M., Appelhans, T., Müller, M. U., Nauss, T., Roudier, P., Zawar-Reza, P.: Mapping daily air temperature for Antarctica based on MODIS LST, Remote Sens., 8, 732, doi:10.3390/rs8090732, 2016a.

Meyer, H., Kühnlein, M., Appelhans, T., Nauss, T.: Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals, Atmos. Res., 169, 424-433, doi:10.1016/j.atmosres.2015.09.021, 2016b.

Meyer, H., Lehnert, L. W., Wang, Y., Reudenbach, C., Nauss, T., Bendix, J.: From local spectral measurements to maps of vegetation cover and biomass on the Qinghai-Tibet-Plateau: Do we need hyperspectral information?, Int. J. Appl. Earth Obs. Geoinf., 55, 21-31, doi:10.1016/j.jag.2016.10.001, 2017.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T.: Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation, Environ. Modell. Softw., 101, 1-9, doi:10.1016/j.envsoft.2017.12.001, 2018.

Micheletti, N., Foresti, L., Robert, S., Leuenberger, M., Pedrazzini, A., Jaboyedoff, M., Kanevski, M.: Machine learning feature selection methods for landslide susceptibility mapping, Math. Geosci., 46, 33-57, doi:10.1007/s11004-013-9511-0, 2014.

Mienert, J., Bünz, S., Guidard, S., Vanneste, M., Berndt, C.: Ocean bottom seismometer investigations in the Ormen Lange area offshore mid-Norway provide evidence for shallow gas layers in subsurface sediments, Mar. Pet. Geol., 22, 287-297, doi:10.1016/j.marpetgeo.2004.10.020, 2005.

Müller, A. C., Guido, S.: Introduction to Machine Learning with Python, O'Reilly Media, Sebastopol, CA, 2017.

Müller, D., Sdrolias, M., Gaina, C., Roest, W. R.: Age, spreading rates, and spreading asymmetry of the world's ocean crust, Geochem. Geophys. Geosyst., 9, Q04006, doi:10.1029/2007GC001743, 2008.

Netzeband, G. L., Krabbenhoeft, A., Zillmer, M., Petersen, C. J., Papenberg, C., Bialas, J.: The structures beneath submarine methane seeps: Seismic evidence from Opouawe Bank, Hikurangi Margin, New Zealand, Mar. Geol., 272, 59-70, doi:10.1016/j.margeo.2009.07.005, 2010.

Obelcz, J., Wood, W. T.: Towards a quantitative understanding of parameters driving submarine slope failure: A machine learning approach, EGU General Assembly, Vienna, Austria, 8-13 April 2018, EGU2018-9778, 2018.

Otey, M. E., Ghoting, A., Parthasarathy, S.: Fast distributed outlier detection in mixed-attribute data sets, Data Min. Knowl. Disc., 12, 203-228, doi:10.1007/s10618-005-0014-6, 2006.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 1825-2830, 2011.

Plaza-Faverola, A., Bünz, S., Mienert, J.: Fluid distributions inferred from P-wave velocity and reflection seismic amplitude anomalies beneath the Nyegga pockmark field of the mid-Norwegian margin, Mar. Pet. Geol., 27, 46-60, doi:10.1016/j.marpetgeo.2009.07.007, 2010a.

Plaza-Faverola, A., Westbrook, G. K., Ker, S., Exley, R. J. K., Gailler, A., Minshull, T. A., Broto, K.: Evidence from three-dimensional seismic tomography for a substantial accumulation of gas hydrate in a fluid-escape chimney in the Nyegga pockmark field, offshore Norway, J. Geophys. Res. Solid Earth, 115, B08104, doi:10.1029/2009JB007078, 2010b.

Plaza-Faverola, A., Klaeschen, D., Barnes, P., Pecher, I., Henrys, S., Mountjoy, J.: Evolution of fluid expulsion and concentrated hydrate zones across the southern Hikurangi subduction margin, New Zealand: An analysis from depth migrated seismic data, Geochem. Geophys. Geosyst., 13, Q08018, doi:10.1029/2012GC004228, 2012.

Plaza-Faverola, A., Pecher, I., Crutchley, G., Barnes, P. M., Bünz, S., Golding, T., Klaeschen, D., Papenberg, C., Bialas, J.: Submarine gas seepage in a mixed contractional and shear deformation regime: Cases from the Hikurangi oblique-subduction margin. Geochem. Geophys. Geosyst., 15, 416-433, doi:10.1002/2013GC005082, 2014.

Poupinet, G., Shapiro, N. M.: Worldwide distribution of ages of the continental lithosphere derived from a global seismic tomographic model, Lithos, 109, 125-130, doi:10.1016/j.lithos.2008.10.023, 2009.

Priddy, K. L., Keller, P. E.: Artificial Neural Networks: an Introduction, SPIE Press, Bellingham, WA, 2005.

Scanlon, G. A., Bourke, R. H., Wilson, J. H.: Estimation of bottom scattering strength from measured and modeled mid-frequency sonar reverberation levels, IEEE J. Ocean. Eng., 21, 440-451, doi:10.1109/48.544055, 1996.

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

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory, 2nd edn. Springer, New York, NY, 314 pp.

Wang, J., Guo, C., Hou, Z., Fu, Y., Yan, J.: Distributions and vertical variation patterns of sound speed of surface sediments in South China Sea, J. Asian Earth Sci., 89, 46-53, doi:10.1016/j.jseaes.2014.03.026, 2014.

Wei, C.-L., Rowe, G. T., Escobar-Briones, E., Boetius, A., Soltwedel, T., Caley, M. J., Soliman, Y., Huettmann, F., Qu, F., Yu, Z., Pitcher, C. R., Haedrich, R. L., Wicksten, M. K., Rex, M. A., Baguley, J. G., Sharma, J., Danovaro, R., MacDonald, I. R., Nunnally, C. C., Deming, J. W., Montagna, P., Lévesque, M., Weslawski, J. M., Wlodarska-Kowalczuk, M., Ingole, B. S., Bett, B. J., Billett, D. S. M., Yool, A., Bluhm, B. A., Iken, K., Narayanaswamy, B. E.: Global patterns and predictions of seafloor biomass using random forests, PLoS ONE, 5, e15323, doi:10.1371/journal.pone.0015323, 2010.

Wessel, P., Smith, W. H. F.: A global, self-consistent, hierarchical, high-resolution shoreline database, J. Geophys. Res., 101, 8741-8743, doi:10.1029/96JB00104, 1996.

Wessel, P., Smith, W. H. F., Scharroo, R., Luis, J., Wobbe, F.: Generic Mapping Tools: Improved version released, EOS Trans. Am. Geophys. Union, 94, 409-410, doi:10.1002/2013EO450001, 2013.

Westbrook, G. K., Chand, S., Rossi, G., Long, C., Bünz, S., Camerlenghi, A., Carcione, J. M., Dean, S., Foucher, J.-P., Flueh, E., Gei, D., Haacke, R. R., Madrussani, G., Mienert, J., Minshull, T. A., Nouzé, H., Peacock, S., Reston, T. J., Vanneste, M., Zillmer, M.: Estimation of gas hydrate concentration from multi-component seismic data at sites on the continental margins of NW Svalbard and the Storegga region of Norway. Mar. Pet. Geol., 25, 744-758, doi:10.1016/j.marpetgeo.2008.02.003, 2008.

Whittaker, J. M., Goncharov, A., Williams, S., Müller, R. D., Leitchenkov, G.: Global sediment thickness data set uploaded for the Australian-Antarctic Southern Ocean. Geochem. Geophys. Geosyst., 14, 3297-3305, doi:10.1002/ggge.20181, 2013.

Wood, W., Lee, T., Obelcz, J.: Practical quantification of uncertainty in seabed property prediction using geospatial KNN machine learning. EGU General Assembly, Vienna, Austria, 8-13 April 2018, EGU2018-9760, 2018.

Data availability: All $v_p(z)$ data used in this study are available through the website of the International Ocean Discovery Program at http://www.iodp.org. Additional datasets and grids used are referenced in the Methods section and in Table 1.

Author contributions: ID performed the modelling and analyses and CB acquired funding for the project. ID prepared the manuscript with contributions from CB.

Competing interests: The authors declare that they have no conflict of interest.

Figure captions

Figure 1. Distribution of the 333 boreholes from which $v_p(z)$ profiles were extracted. DSDP – Deep Sea Drilling Project, ODP – Ocean Drilling Program, IODP – International Ocean Discovery Program. Bathymetry (30 s resolution) is from the GEBCO_2014 grid (http://www.gebco.net).

Figure 2. Examples for true $v_p(z)$ curves, predicted $v_p(z)$ curves, and $v_p(z)$ calculated from the five Hamilton functions (Hamilton, 1985) used in model evaluation. (a)-(d) well predicted $v_p(z)$ curves of score 3, (e) lower-quality~~less good~~ prediction of score 3, (f) score 2, (g)-(h) bad predictions of scores 1 and 0. See 2.3 for a description of H1 to H5.

Figure 3. Comparison of (a) error metrics and (b) proportion of well predicted boreholes (scores 2 and 3) for different model runs. RMSE – root mean square error, MAE – mean absolute error, CV – cross-validation, RFE – Recursive Feature Elimination.

Figure 4. Comparison of (a) error metrics and (b) proportion of well predicted boreholes (scores 2 and 3) for model runs with different degrees of data smoothing. RMSE – root mean square error, MAE – mean absolute error, CV – cross-validation.

Figure 5. Distribution of boreholes with (a) good (scores 2 and 3) and (b) bad (scores 0 and 1) v_p predictions. Areas A-E mark clusters of boreholes in the Sea of Japan (A), the Nankai Trough (B), the Ontong-Java Plateau (C), the Queensland Plateau (D), and in the Great Australian Bight (E). Area F indicates an example for remote boreholes of score 3 on the Mid-Atlantic Ridge. Bathymetry (30 s resolution) is from the GEBCO_2014 grid (http://www.gebco.net).

Figure 6. Predictor importance ranking for the 38-predictor model run. For each predictor, the importance was averaged over the ten runs of the 10-fold CV. Categorical predictors are marked with an asterisk. Predictor names are explained in Table 1.

Table captions

Table 1. Overview of the 38 predictors and their sources.

Table 2. Scores for performance comparison between RF prediction and v_p calculated from the empirical functions of Hamilton (1985).

Table 3. Predictor ranking based on the RFE results for unscaled and scaled predictors. Categorical predictors are marked with an asterisk. See Table 1 for an explanation of predictor names.

Table 1

predictor	description	type	source description	reference
lat	latitude	continuous	DSDP/ODP/IODP data processing notes	
long	longitude	continuous	DSDP/ODP/IODP data processing notes	
wdepth	water depth	continuous	DSDP/ODP/IODP data processing notes	
depth	depth below seafloor	continuous	v_p logs	
crustage	age of crust	continuous	ocean crust: global crustal age grid (2 min res.)	Müller et al. (2008)
			continental crust: 1 Byr (const.)	
sedthick	sediment thickness	continuous	global sediment thickness grid (5 min res.)	Whittaker et al. (2013)
spreadrate	spreading rate	continuous	global spreading rate grid (2 min res.)	Müller et al. (2008)
heatflow	surface heatflow	continuous	global surface heatflow grid (2° res.)	Davies (2013)
depth2base	depth to acoustic basement	continuous	derived from sediment thickness and depth	
dist2smt	distance to nearest seamount	continuous	derived from global seamount dataset	Kim and Wessel (2011)
dist2hole	distance to nearest borehole	continuous	derived from borehole locations	

dist2coast	distance to nearest coast	continuous	derived from global shoreline dataset	Wessel and Smith (1996)
dist2trench	distance to nearest trench	continuous	derived from global trench dataset	Coffin et al. (1998)
dist2ridge	distance to nearest spreading ridge	continuous	derived from global spreading ridge dataset	Coffin et al. (1998)
dist2transform	distance to nearest transform boundary	continuous	derived from global transform boundary dataset	Coffin et al. (1998)
oceancrust	oceanic crust	categorical	derived from crustal age	
contcrust	continental crust	categorical	derived from crustal age	
active_margin	geological setting: active margin	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
passive_margin	geological setting: passive margin	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
spreading_ridge	geological setting: spreading ridge	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
subduction	geological setting: subduction zone	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
volcanic_arc	geological setting: volcanic arc	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
fore-arc	geological setting: fore-arc basin	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
accretion_wedge	geological setting: accretionary wedge	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
trench	geological setting: trench	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
cont_slope	geological setting: continental slope	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
shelf	geological setting: continental shelf	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
reef	geological setting: (former) reef	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
basin	geological setting: basin	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
struct_high	geological setting: structural high	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
cont_plateau	geological setting: continental plateau	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
aseismic_ridge	geological setting: aseismic ridge	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
seamount	geological setting: seamount	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
guyot	geological setting: guyot	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
mud_volcano	geological setting: mud volcano	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
ds_fan	geological setting: deep-sea fan	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
hydroth_vent	geological setting: hydrothermal vent	categorical	DSDP/ODP/IODP proceedings (site descriptions)	
cold_vent	geological setting: cold vent	categorical	DSDP/ODP/IODP proceedings (site descriptions)	

Table 2

Score	Description	Inferred prediction performance
3	all 3 error metrics of RF prediction indicate better fit than empirical functions	~~G~~good
2	2 of 3 error metrics of RF prediction indicate better fit than empirical functions	g~~G~~ood
1	1 of 3 error metrics of RF prediction indicate better fit than empirical functions	~~B~~bad
0	all 3 error metrics of empirical functions indicate better fit than RF prediction	~~B~~bad

Table 3

Position	Predictor	
	RFE unscaled	**RFE scaled**
1	~~C~~crustage	wdepth
2	depth2base	depth2base
3	~~W~~wdepth	dist2smt
4	dist2smt	depth
5	~~D~~depth	heatflow
6	~~H~~heatflow	sedthick
7	dist2hole	dist2trench
8	dist2coast	dist2hole
9	dist2trench	dist2coast
10	~~L~~lat	spreadrate
11	~~S~~sedthick	dist2ridge
12	~~S~~spreadrate	long
13	dist2ridge	lat
14	~~L~~long	dist2transform

15	dist2transform	crustage
16	spreading_ridge*	contcrust*
17	cont_plateau*	basin*
18	reef*	active_margin*
19	aseismic_ridge*	struct_high*
20	basin*	oceancrust*
21	struct_high*	passive_margin*
22	oceancrust*	subduction*
23	volcanic_arc*	reef*
24	active_margin*	accretion_wedge*
25	contcrust*	cont_plateau*
26	guyot*	cont_slope*
27	passive_margin*	spreading_ridge*
28	trench*	fore-arc*
29	subduction*	shelf*
30	seamount*	ds_fan*
31	fore-arc*	volcanic_arc*
32	hydroth_vent*	trench*
33	cont_slope*	seamount*
34	shelf*	aseismic_ridge*
35	accretion_wedge*	guyot*
36	ds_fan*	cold_vent*
37	mud_volcano*	hydroth_vent*
38	cold_vent*	mud_volcano*

5