Received: 16 May 2020 | Accepted: 6 May 2021

DOI: 10.1111/2041-210X.13650

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Predicting into unknown space? Estimating the area of applicability of spatial prediction models

Hanna Meyer[1] | Edzer Pebesma[2]

[1]Institute of Landscape Ecology, Westfälische Wilhelms-Universität Münster, Münster, Germany

[2]Institute for Geoinformatics, Westfälische Wilhelms-Universität Münster, Münster, Germany

Correspondence
Hanna Meyer
Email: hanna.meyer@uni-muenster.de

Handling Editor: Robert Freckleton

Abstract

1. Machine learning algorithms have become very popular for spatial mapping of the environment due to their ability to fit nonlinear and complex relationships. However, this ability comes with the disadvantage that they can only be applied to new data if these are similar to the training data. Since spatial mapping requires predictions to new geographic space which in many cases goes along with new predictor properties, a method to assess the area to which a prediction model can be reliably applied is required.
2. Here, we suggest a methodology that delineates the 'area of applicability' (AOA) that we define as the area where we enabled the model to learn about relationships based on the training data, and where the estimated cross-validation performance holds. We first propose a 'dissimilarity index' (DI) that is based on the minimum distance to the training data in the multidimensional predictor space, with predictors being weighted by their respective importance in the model. The AOA is then derived by applying a threshold which is the (outlier-removed) maximum DI of the training data derived via cross-validation. We further use the relationship between the DI and the cross-validation performance to map the estimated performance of predictions. We illustrate the approach in a simulated case study chosen to mimic ecological realities and test the credibility by using a large set of simulated data.
3. The simulation studies showed that the prediction error within the AOA is comparable to the cross-validation error of the trained model, while the cross-validation error does not apply outside the AOA. This applies to models being trained with randomly distributed training data, as well as when training data are clustered in space and where spatial cross-validation is applied. Using the relationship between DI and cross-validation performance showed potential to limit predictions to the area where a user-defined performance applies.
4. We suggest to add the AOA computation to the modeller's standard toolkit and to present predictions for the AOA only. We further suggest to report a map of DI-dependent performance estimates alongside prediction maps and complementary to (cross-)validation performance measures and the common uncertainty estimates.

KEYWORDS
machine learning, model transferability, predictive modelling, Random Forest, remote sensing, spatial mapping, uncertainty

1 | INTRODUCTION

Spatial mapping is an important task in environmental science to reveal spatial (and spatio-temporal) patterns and changes in the environment. Predictive modelling is a common method in this context, where field data are used to train statistical models using spatially continuous predictor variables, for example derived from remote sensing imagery. The resulting model is then used to make predictions for the entire area of interest, i.e. beyond the geographic locations of training data. In the last years, machine learning algorithms have become the most popular tool in predictive modelling being able to capture nonlinear and complex relationships. In this way, a large variety of different environmental variables have been mapped even ambitiously on a global scale, such as global tree restoration potential (Bastin et al., 2019), soil properties (Hengl et al., 2017), distribution of nematodes (van den Hoogen et al., 2019) and soil bacteria (Delgado-Baquerizo et al., 2018), global leaf-freezing resistance (Zohner et al., 2020) or plant species Red List status (Pelletier et al., 2018) to mention just a few. However, the reliability of machine learning-based global prediction maps is increasingly called into question, leading to a loss of confidence in these maps (e.g. see comments to the highly discussed paper of Bastin et al., 2019). Improved analysis and communication of uncertainties of spatial predictions is therefore required. This is important to identify locations where predictions are too uncertain to be considered for further action, for example in the context of prioritizing conservation assessment (Pelletier et al., 2018), reserve design or if predictions are used as input for subsequent modelling where propagation of large errors should be avoided.

The performances of machine learning models are typically communicated via (cross-) validation estimates where the cross-validation strategy should always be designed according to the purpose of the model. In the context of spatial mapping, the relevance of accounting for spatial dependencies for reliable performance estimation via cross-validation has been recently highlighted by many studies (Brenning, 2012; Meyer et al., 2018; Ploton et al., 2020; Pohjankukka et al., 2017; Roberts et al., 2017; Schratz et al., 2019; Valavi et al., 2018). Spatial (cross-)validation provides a general error estimate for the predictions that is less sensitive to spatial dependence than cross-validation based on random partitioning; however, we argue that this is not sufficient to communicate the performance of prediction maps—typically field samples used as training data for predictive mapping are not evenly distributed over study areas and often predictions are made for areas that are lacking a support of training data. For example, in the global map of soil nematode densities of van den Hoogen et al. (2019), central Africa as well as North East Asia are lacking any training data, but predictions are made for these areas. By transferring the model beyond the training locations (i.e. to new geographic space), it is assumed that the learned relationships between predictors and responses still hold. However, especially in heterogeneous landscapes, the new geographic space might differ considerably in its environmental properties from what has been observed in the training data. This leads to a question that is not addressed by cross-validation so far—what happens if the algorithm has never 'seen' such environmental properties? This is relevant as most machine learning algorithms can fit very complex relationships; however, this ability comes with the disadvantage that they can only be applied to new data if these are similar to the training data. Therefore, gaps in the predictor space where there is no support of training data must be considered problematic because the algorithm was not enabled to learn about the relationships in these environments.

Since spatial mapping requires predictions to new geographic space which in many cases goes along with new predictor properties, we need to measure how dissimilar predictors at new locations are from those in the training data. Based on this, a delineation of the area to which a prediction model can reliably be applied is required. We call this the 'area of applicability' (AOA) of a prediction model. Similar concepts have been discussed mainly in the field of chemical modelling (Quantitative Structure-Activity Relationship (QSAR) models, see e.g. Mathea et al., 2016, Toplak et al., 2014, Gadaleta et al., 2016, where the concept is usually referred to as 'domain of applicability') and have been addressed in the field of species distribution modelling (Elith et al., 2010; Mesgaran et al., 2014; Zurell et al., 2012) and soil mapping (Zhu et al., 2015). Also, extrapolation conditions have been occasionally mapped (Bastin et al., 2019); however, these approaches usually consider minima and maxima of individual predictors or the regressor variable hull (Montgomery et al., 2012), hence do neither account for gaps in the predictor space, nor for unobserved combinations of predictors. Overall, applications in the field of spatial predictive modelling are rare. Instead, models are often assumed to be applicable to the entire area of interest (e.g. globally). An exception in many global prediction maps is that Antarctica is often masked from the predictions (e.g. van den Hoogen et al., 2019) probably because, by expert knowledge, the models can clearly not be applied to this new environment. However, other environments might have to be considered equally unsuitable for model application. This can be very obvious (e.g. high mountain ranges using a model trained in lowlands) but also hard to assess by expert knowledge when areas feature combinations of environmental variables that are not covered by training data.

This aspect is not addressed by common approaches of uncertainty estimation in machine learning, which are usually based on the variance of predictions made by ensembles of models (e.g. Bastin et al., 2019; Coulston et al., 2016; van den Hoogen et al., 2019, in the field of spatial mapping). Such a measure is very obvious for

ensemble-based algorithms like Random Forests, where each tree is regarded as a model of an ensemble, and the variation in predictions among individual trees is used to quantify uncertainty (e.g. standard deviations of individual predictions or prediction intervals). Missing knowledge about environments is also not addressed by the quantile regression forests (Meinshausen, 2006) that are occasionally suggested in the context of spatial mapping (Hengl et al., 2018; Vaysse & Lagacherie, 2017) to derive prediction intervals based on the conditional distribution of the response variable. Figure 1 shows an uncertainty estimation for a linear regression model and a Random Forest model. Clearly, both provide intervals that only make sense in the context of the respective models being valid, but it shows that Random Forest prediction intervals estimated from variability of predictions of the ensemble do not acknowledge that prediction gets harder when one moves further away from the training data, outside the data range or into significant gaps. While these uncertainty estimates give valuable information on the variance in predictions and hence on locations where predictions are robust within the model, these approaches give no information about 'unknown environments' because dissimilarities in the predictor space between training and new data are not considered.

A way to measure how dissimilar predictors at new locations are from those in the training data is hence required. One option is to look at distances in the (multidimensional) predictor space between training data and a new data point (e.g. Sheridan et al., 2004). Here, we suggest calculating distances to the nearest training data point (in the predictor space) which allows accounting for values outside the range or hull of predictors, as well as significant gaps in the predictor space, something that cannot be accounted for by methods like hulls (Montgomery et al., 2012; Netzeva et al., 2005) or average distances. Using raw predictor space distances, however, may be problematic because in a machine learning model, typically certain variables have a high importance while others may be completely irrelevant (i.e. they differ in the degree to which they drive the prediction patterns). To handle this, Janet et al. (2019) suggest to use the distance in the latent predictor space of a neural network. However, this approach is specific to neural networks and not generically applicable to, for example, Random Forests. Instead, we suggest weighting predictors according to their relevance in the model. Using the minimum distance to training data in the weighted predictor space, normalized by average distances between training data, such a dissimilarity index (DI) allows mapping the dissimilarity of predictor variables to the values in the training data in a continuous way.

To identify areas that are too different from the training data to be considered reliable for predictions, hence to derive the AOA, a threshold on the DI is required. We suggest deriving this threshold from the training data by identifying the (outlier-removed) maximum dissimilarity of the training data via cross-validation. As a consequence, the AOA is not only the area where we enabled the model to learn about relationships based on the training data, but also the area to which, on average, the performance measure estimated by cross-validation of the model applies. Just like for tuning and performance assessment, deriving the threshold from cross-validation means that this threshold is sensitive to the cross-validation strategy being used. The cross-validation strategy should always be designed according to the purpose of the model. Not considering this and choosing for instance randomly assigned folds for cross-validation with geographically clustered training data might lead to optimistic model performance estimates (e.g. Ploton et al., 2020) and to a very small AOA since the performance measure as well as the threshold is derived from training data with very strong similarities. Choosing the cross-validation strategy according to the purpose of the model is hence essential.

However, since the assessment of when a prediction is reliable or not depends on its purpose, and hence on the targeted user's requirements, it is hard to argue for a one-fits-all AOA. As we define the AOA as the area for which a certain cross-validation performance measure holds, and since different cross-validation strategies lead to different performance measures with different AOAs, we propose to use this and establish a relationship between DI and performance by adapting the cross-validation strategy. By creating cross-validation folds using a clustering of observations in predictor space, and by

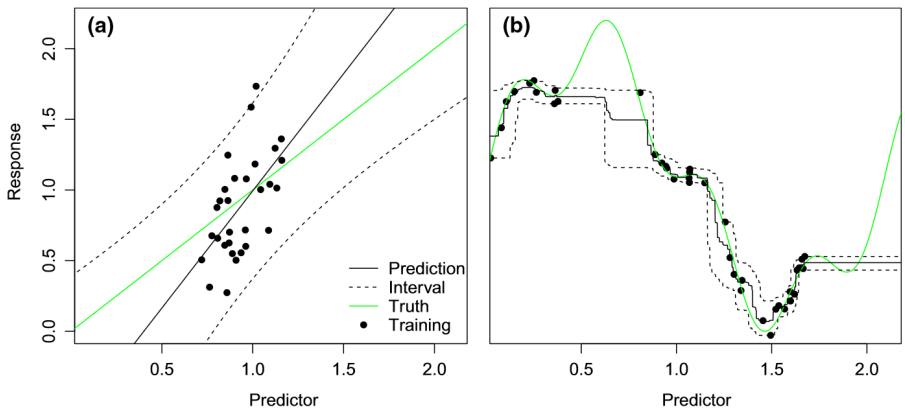

FIGURE 1 Problem of predicting beyond the training data and behaviour of prediction intervals for different models. Left: linear regression prediction interval width increases with distance from the centre of the training data, right: a more complex relationship fitted with Random Forest. Prediction beyond the training data becomes highly unreliable, although prediction interval width outside the data range is constant. Random Forest prediction intervals were obtained by computing quantiles over the predictions from individual trees

decreasing the number of clusters in a stepwise manner, we create multipurpose cross-validation scenarios with increasing dissimilarities (between folds), corresponding to lower performance estimates and larger AOAs, of course up to a point where the training data no longer contain the targeted dissimilarities.

2 | MATERIALS AND METHODS

We suggest a method that provides a unitless measure for expressing how different a new data point is from the training data. We call this the 'dissimilarity index' (DI), and suggest a threshold for it to define the 'area of applicability' (AOA) of a prediction model. The method requires two datasets that are part of any supervised predictive modelling task. The first dataset is the training set that includes the sampling locations that are intended as training data used for model training. The second dataset contains the new locations for which predictions should be made. In the case of spatial mapping, this is the set of spatially continuous data, usually raster data with predictor variable values that are known for the entire area of interest. For a visual explanation of the methodology, please also see the Supporting Information of this paper.

2.1 | Standardization of predictor variables

To ensure that all variables are treated equally, the predictor variables are scaled by dividing mean-centred values by their respective standard deviations,

$$X_{i,j}^{s} = (X_{i,j} - \overline{X}_{.j})/\sigma_j,$$

where $X_{i,j}^{s}$ refers to the scaled value of the jth predictor variable corresponding to the ith observation, $\overline{X}_{.j}$ to the mean and σ_j to the standard deviation of the jth predictor variable, and mean and standard deviation are computed over the training data. If categorical predictor variables are used, dummy variables are created prior to standardization.

2.2 | Weighting of variables

If distances were calculated based on the standardized predictors, all variables would be treated equally important. However, distances are not equally relevant within the predictor space but some variables are more important than others in the machine learning model and hence are mainly responsible for prediction patterns. Most machine learning models provide an estimate of relative variable importance (see e.g. overview in Kuhn, 2008). To reflect the variable importance in the computation of distances in the predictor variable space, we multiply the scaled variables with the non-standardized importance estimate w_j for each variable j before distance calculation takes place, by

$$X_{i,j}^{sw} = w_j X_{i,j}^{s}.$$

As a consequence, distances in the predictor space in the direction of the more important variables have a higher effect on our dissimilarity measure.

2.3 | Multivariate distance calculation

The Euclidean distance between two arbitrary points a and b in the predictor variable space is calculated as

$$d(a,b) = \sqrt{\sum_{j=1}^{p} \left(X_{a,j}^{sw} - X_{b,j}^{sw}\right)^2}.$$

For a new prediction location k, the distance to the nearest training data point i

$$d_k = \underset{i}{arg\,min}\; d(k,i).$$

is used to calculate the DI.

2.4 | Dissimilarity index

To allow for interpretation and comparison between models, we standardize distances in predictor space for new prediction locations k by dividing the minimum distance to the nearest training data point d_k (Figure 2b) by the average of the distances in the training data $\overline{d}$ (Figure 2a), and call this the dissimilarity index DI_k, defined as

$$\mathrm{DI}_k = d_k/\overline{d}$$

with $\overline{d}$ the average of all pairwise distances between the n training data.

Using the standardized weighted distances, the DI can take values ranging from 0 to ∞. If the result is 0, the new data point is identical in its predictor properties to a training data point. With increasing values of the DI, the distance to the nearest training data point increases. If the values are greater than 1, the difference to the nearest training data point is larger than the average dissimilarity (i.e. average distance) between all training data pairs.

2.5 | Deriving the area of applicability

To derive the AOA, a threshold on the DI is required. With regard to the definition of the AOA, we derive the threshold from the DI values of the training data, with the DI calculated based on data points that do not occur in the same cross-validation fold. Taking the cross-validation folds into account is required because cross-validation is based on repeatedly leaving training data out, hence we assume the estimated model performance applies to areas with DI values comparable to those found during cross-validating the training data. Therefore, we calculate the DI for each training data point as described in Section 2.4; however, in line with the cross-validation

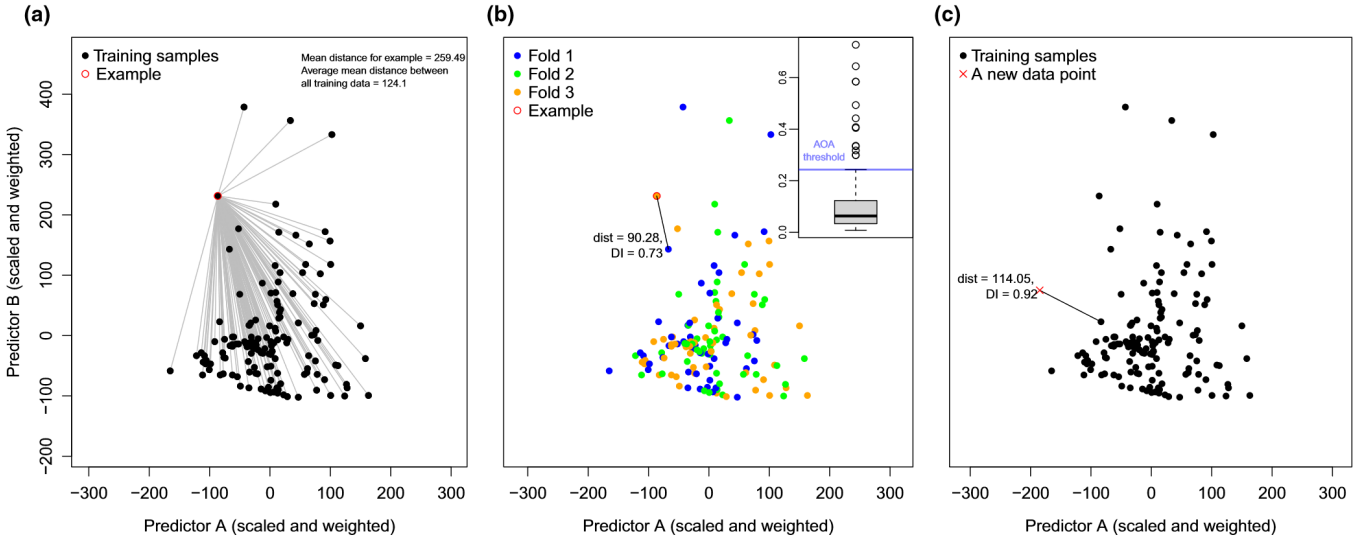

FIGURE 2 Training samples in a multidimensional (here two-dimensional) predictor space that has been scaled and weighted. First, the average of the mean distances between all training data is calculated (a). Next, the dissimilarity index (DI) of the training data is calculated. For each training data point (shown here for one example), the distance to the nearest training data point not located in the same cross-validation fold is calculated (here visualized assuming a threefold cross-validation; b). This distance is divided by the average of the mean distances between all training data (a) to derive the DI. The DI is calculated for each training data point (boxplot in b) and the threshold for the area of applicability (AOA) is then derived from the upper whisker of the DI values. For a new data point, the DI is calculated accordingly (c). In this example, the DI is larger than the DI threshold, indicating that this new data point falls outside the AOA

strategy being used, distance is measured to the nearest training data point (in the predictor space) that is not in the same cross-validation fold (Figure 2b). For example, if a threefold random cross-validation is applied for model validation, data are randomly split into three folds and the cross-validation error is the average over the prediction errors for each of the folds held back for validation. Hence, the predictions for a respective fold are based on a model being trained on the remaining data. In line with that, the DI for each training data point is calculated by using the distance to the nearest training data point that is not in the same fold. The outlier-removed maximum DI of the training data is the one used as threshold for the AOA (boxplot in Figure 2b) where outliers are defined as values greater than the upper whisker (i.e. larger than the 75-percentile plus 1.5 times the IQR of the DI values of the cross-validated training data).

For each new data point, the DI is calculated as described in Section 2.4 (Figure 2c). Applying the threshold described above to every position in the area delineates the AOA.

2.6 | Using DI to quantitatively express prediction uncertainty

The AOA is a binary information about the area that features predictor properties where we enabled the model to learn about relationships, hence it is the area to which the model can be applied with an expected average performance that is comparable to the cross-validation estimate. However, since the assessment of when a prediction is reliable or not depends on its purpose, and hence on the targeted user's requirements, it is hard to argue for a one-fits-all AOA. It would therefore be desirable to have a quantitative measure of uncertainty, for example to limit predictions to an area where a required performance value applies (Petchey et al., 2015). This is especially of relevance as the AOA depends on the cross-validation strategy being used. While there should be agreement that the cross-validation strategy should be designed according to the purpose of the model (e.g. spatial predictions far beyond clustered training samples call for a spatial cross-validation, see Ploton et al., 2020), the actual interests of different users of predictions may vary (e.g. interest in specific locations only, or in predictions made in close distance to training data or in the area where predictions can be made with a required performance).

To solve this, we propose a quantitative uncertainty measure derived from the relationship between the DI and cross-validation performance. Therefore, we use the DI as well as predictions from each cross-validated data point from the training dataset and use a sliding window along DI values to assess performance metrics (e.g. RMSE, Kappa). We suggest to use these data to fit a suitable parametric model that can be used to translate DI values into expected prediction performances.

Since different cross-validation strategies lead to different performance measures with different AOAs, we propose to use this and establish the relationship between DI and performance by adapting the cross-validation strategy. By repeatedly creating cross-validation folds using a clustering of observations in predictor space, and by decreasing the number of clusters in a stepwise manner, we create cross-validation scenarios with increasing dissimilarities (between folds), corresponding to less reliable predictions and larger AOAs (see Supporting Information in the Appendix). We then compile the DIs and predictions from these multiple cross-validation scenarios to establish the relationship between DI values and model

performances. Compared to a single cross-validation, this allows that the relationship between DI and prediction performance can be assessed across potential purposes, ranging from predicting on nearly identical environments, up to clear extrapolation cases.

2.7 | Simulation studies to test the methodology

To test the suitability of the presented methodology to derive the AOA, we simulated nearly 1,000 prediction tasks where the true values were known. As prediction tasks, we used a spatially continuous response variable of Europe that is simulated based on bioclimatic predictor variables (www.worldclim.org/bioclim) and the simulation approach of Leroy et al. (2016). This approach was developed as an example of virtual species suitability but will be used here as an arbitrary response variable based on environmental predictors. The application of an area-wide simulated response variable is important here, as it allows to compare the predictions with true values, and to assess the benefit of the AOA.

As predictor variables, we used the WorldClim dataset (Hijmans et al., 2005) and chose the 19 bioclimatic variables in 10-min spatial resolution. The response variable was generated by a principal component analysis (PCA) of a subset of the bioclimatic variables. The variables used to simulate the response were the mean diurnal range ('bio2'), maximum temperature of the warmest month ('bio5'), mean temperature of the warmest quarter ('bio10'), precipitation of the wettest month ('bio13'), precipitation of the driest month ('bio14') and precipitation of the coldest quarter ('bio19'). For the PCA, the response to each of the first two principal components (axes) is defined and combined to create the final response variable. Therefore, the response to the two first axes of the PCA is determined with Gaussian functions as described in Leroy et al. (2016). The means of the Gaussian response functions to the axes of the PCA were varied here between 1 and 3 (first axis) and −1 and 1 (second axis), and the standard deviations were varied between 1 and 3 for both axes, resulting in 81 different response variables.

The simulated response variables are available in a spatially continuous way; however, to simulate typical prediction tasks, we simulated field sampling locations. We selected sample point locations randomly from the target area with varying sample sizes (n = 25, 50, 75, 100). Each combination of response variable and sample size was tested with three independent replicates of the random sampling design, resulting in a total of $81 \times 4 \times 3 = 972$ different simulated 'realities'.

We used Random Forests (Breiman, 2001) as machine learning algorithm because it is one of the most frequently used algorithms in the context of environmental mapping (e.g. used in the context of global mapping in Bastin et al., 2019; Hengl et al., 2017; van den Hoogen et al., 2019). To prepare model training, the 19 predictors and the response variables were extracted for the locations of the sampling data points. For model training, the Random Forest implementation of Liaw and Wiener (2002) was used and accessed via the CARET package (Kuhn, 2019) in R (R Core Team, 2020). Each forest consisted of 500 trees and the number of randomly selected variables at each split (mtry) was tuned between 2 and 19 (the number of predictor variables). The minimal size of terminal nodes was 5. Tuning and performance estimation was done using random 10-fold cross-validation. The trained models were applied to the complete set of predictor variables to make spatial predictions over the entire area (81,796 valid pixels). To assess the relative variable importance required for the estimation of the DI, the approach of Liaw and Wiener (2002) was used to estimate w_j. Importance is indicated as the increase in the mean squared error when a variable is randomly permuted. Hence, the higher the decrease, the higher the importance in the model.

From the catalogue of simulations, we computed prediction errors by subtracting predicted from true values, for all prediction locations. Using the derived threshold for the AOA, we compared the root mean square prediction error (RMSPE) which reflects the differences between predicted and true values with the root mean square error (RMSE) which represents the cross-validation error of the model. We tested whether the RMSPE and RMSE correspond, on average, over the set of simulations.

To derive a quantitative performance measure from the DI, we retrained each model using 10 different cross-validations designed to test the ability of the model for interpolation as well as extrapolation. Therefore, we defined the cross-validation folds by k-means clustering in the predictor space with 10 different cluster numbers ranging from 3 to N, where N is the number of training data points (i.e. leave-one-out cross-validation). We then analysed the relationship between all cross-validated predictions and their DIs by calculating the RMSE for a sliding window of DI values (window size = 10). The relationship was used to calibrate the DI, hence to derive quantitative performance estimates. We used a shape-constrained additive model (Pya, 2020; Pya & Wood, 2015) with a monotone increasing constraint to model the relationship.

2.7.1 | Case study

To motivate our proposal, we show the application of the approach in an illustrated case study using a single simulation from the scenarios described above. Therefore, we used a single setting from the 972 simulations where the response was developed from the bioclimatic predictors (presented in Figure 3a) with means of the Gaussian response functions of 3 (first axis) and −1 (second axis) and standard deviations of 2 for both axes. The simulated response had values between 0 and 1 with a mean of 0.31 (Figure 3b). As training data, we randomly selected 50 sample point locations from the target area (red markers in Figure 3b). To further illustrate the suitability of the presented methodology across sampling designs (and hence across suitable cross-validation strategies), as a second example, we simulated a spatially clustered sampling design—instead of 50 randomly selected sampling locations as described before, we simulated 10 sampling points clustered around each of the 50 locations, resulting in 500 training points across the 50 independent locations (Figure 7a).

Model performance estimation was done using a random 10-fold cross-validation for the randomly distributed training data and a leave-cluster-out spatial cross-validation for the clustered training data. DI and AOA were derived as explained above. To arrive at a quantitative uncertainty estimate, the relationship between RMSE and DI was estimated using the data from a single cross-validation as well as from multipurpose cross-validation as described in Section 2.7.

To highlight the advantage of the newly developed methodology, we further compared the DI to the commonly applied standard deviation of the Random Forest ensemble. Therefore, the standard deviations of the individual predictions made by the 500 trees were calculated for a respective pixel.

3 | RESULTS

Using the 972 scenarios, Figure 4a shows that the prediction error within the AOA is in high agreement with the estimated cross-validation error of the model. The model error was not valid outside the AOA, indicated by considerably higher RMSE values for the prediction compared to the cross-validation error (Figure 4b).

As described above, for each scenario, the DI values on a pixel level were calibrated using the respective relationship between the cross-validated prediction errors (i.e. RMSE) and the DI in a sliding window. Using this relationship, the estimated RMSE of predictions within the AOA could reflect the true prediction error with an average R^2 of 0.21 across the model scenarios.

3.1 | Case study

The case study model had a high ability to predict the response variable, indicated by a random cross-validation R^2 of 0.95 and a RMSE of 0.08 for the prediction task using randomly distributed data. The importance of the different predictor variables ranged from 1.5 to 12 (Figure 5) which represented the baseline for variable weighting used to estimate the DI.

The DI (Figure 6e) shows clear spatial patterns across Europe. Values range from 0 to 2.89 with an average of 0.25. Noticeable are high values (low applicability) in the Alps and at the west coast of Norway. This means that these areas feature very distinct environments compared to the environments covered by the training data. The standard deviations of the Random Forest predictions feature very different spatial patterns (Figure 6c) that are not in agreement with the true absolute prediction error (Figure 6d). In contrast, the DI (Figure 6e) reflects the spatial patterns in the true error (Figure 6d), with a correlation coefficient of $r = 0.71$. If variables were not weighted according to their relevance in the model (Figure 5), the DI was found to be less in accordance with the true absolute error ($r = 0.62$).

The threshold for the AOA as derived from the DI of the cross-validated training data was 0.64. Figure 6f shows the predictions made by the model (Figure 6b) but masked by the AOA. The average agreement between the reference and the prediction was higher within the AOA ($r = 0.97$, RMSE = 0.07) compared to the entire study area ($r = 0.93$, RMSE = 0.10). Outside the AOA, the agreement was considerably lower ($r < 0.00$, RMSE = 0.44). Note that the prediction error within the AOA was in high agreement with the random cross-validation error of the model (RMSE = 0.08).

Using the scenario of spatially clustered data points for model training (Figure 7a), the random cross-validation RMSE was 0.019. When testing the ability of the model to make predictions beyond clusters, hence when validated with a leave-cluster-out spatial cross-validation, the RMSE increased to 0.036. Using the threshold on the DI estimated by taking into account distances to data points not located in the same spatial cluster, the AOA for which the

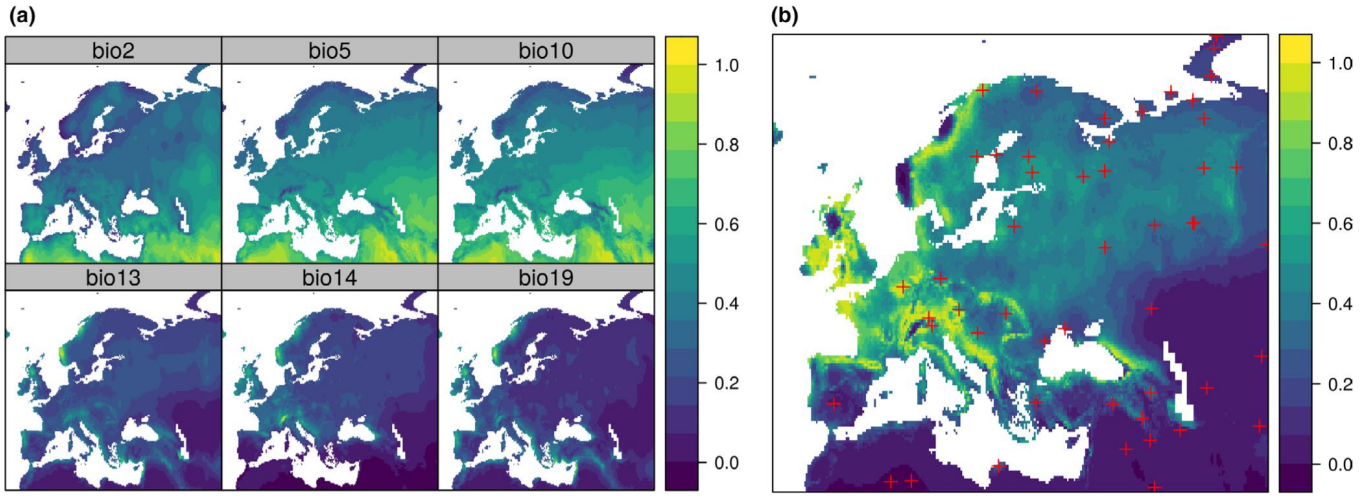

FIGURE 3 Bioclimatic variables used to simulate the response variable for the case study. The variables are stretched here between 0 and 1 for visualization purposes (a). (b) Simulated response variable for the case study and the location of the 50 randomly selected sampling points (red markers) used as training data

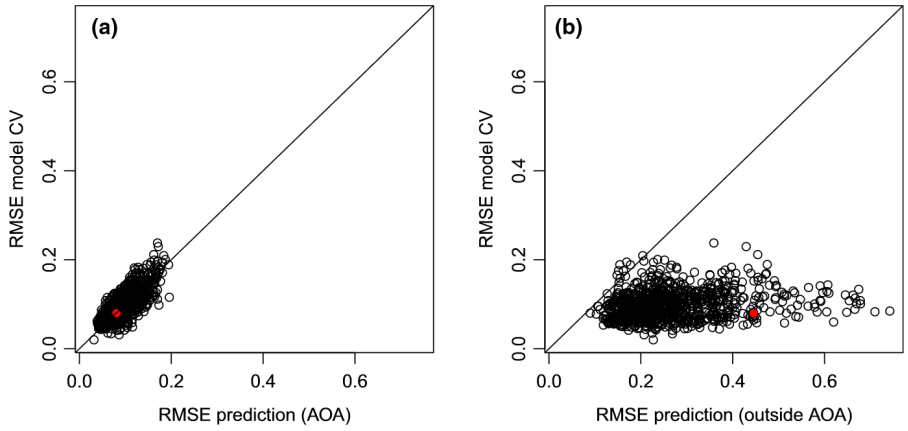

FIGURE 4 RMSE of the model (cross-validation, *y*-axis) against RMSPE (of true prediction errors outside training data, *x*-axis) for the 972 simulations, inside the area of applicability (AOA; left) and outside the AOA (right); inside the AOA both errors correspond on average, outside this area the RMSPE is much larger. The red dot shows the results for the case study scenario

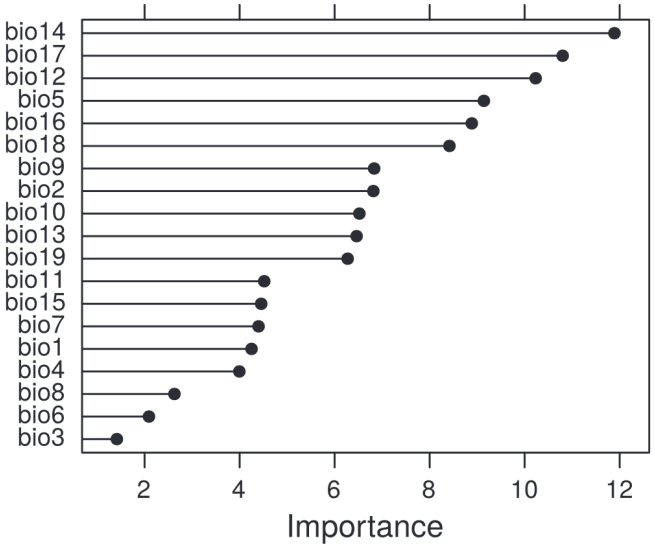

FIGURE 5 Importance of the predictor variables within the Random Forest model based on the 50 randomly distributed sample data. Importance is indicated as the increase in the mean squared error when a variable is randomly permuted

spatial cross-validation applies (Figure 7b) was considerably larger compared to the AOA for which the random cross-validation error applies (Figure 7c). The true prediction RMSE within the AOA was in both cases comparable to the respective cross-validation RMSE—0.046 for the AOA of the spatial model and 0.022 for the AOA of the random model.

Obviously, the AOA depends on the cross-validation strategy that was agreed on according to the intended purpose of the model. Using the relationship between the DI and the performance allows limiting the AOA to a user-defined performance threshold. The results showed a strong relationship between the DI and the RMSE within the AOA, modelled with the shape-constrained additive model (R^2 = 0.82). Figure 8a shows this for the case study using the clustered design and where a spatial cross-validation was applied. The RMSE was low for data points with a small DI and increased with increasing DI values up to the AOA threshold at DI = 0.50. The true relationship between the RMSE and the DI (red points in Figure 8) was generally comparable to the relationship estimated based on the cross-validation. Using multiple cross-validation strategies allowed for a larger AOA (up to a DI of 0.79) and hence for a more comprehensive assessment of the DI-dependent RMSE (Figure 8b; R^2 = 0.83). Using the model allowed mapping the estimated performance on a pixel level as a baseline to limit predictions to areas where user-defined performance applies—of course only within the maximum possible AOA (Figure 9).

4 | DISCUSSION

We propose a method to estimate the AOA of predictive models, by which we mean the area where we enabled models to learn about relationships and where, as a consequence, predictions are expected to have an average error that is comparable to the model error estimated using cross-validation.

The AOA is derived by thresholding the DI, a standardized distance in the multidimensional predictor space, using the outlier-removed maximum DI of the training data encountered during cross-validation. A new data point is outside the AOA when its DI exceeds this threshold. Based on a catalogue of 972 simulations, we found that prediction errors within the AOA are on average similar to the cross-validation error. The cross-validation error of the model should not be considered valid outside the AOA because the DI (i.e. dissimilarity) is greater than the DI values encountered during cross-validation.

Knowledge on the AOA is relevant when predictions are made for heterogeneous areas but based on limited field data, or are made across study areas where it is unclear whether the model can be applied to the new environment. Yates et al. (2018) raised the need for assessing the transferability of prediction models as an 'outstanding

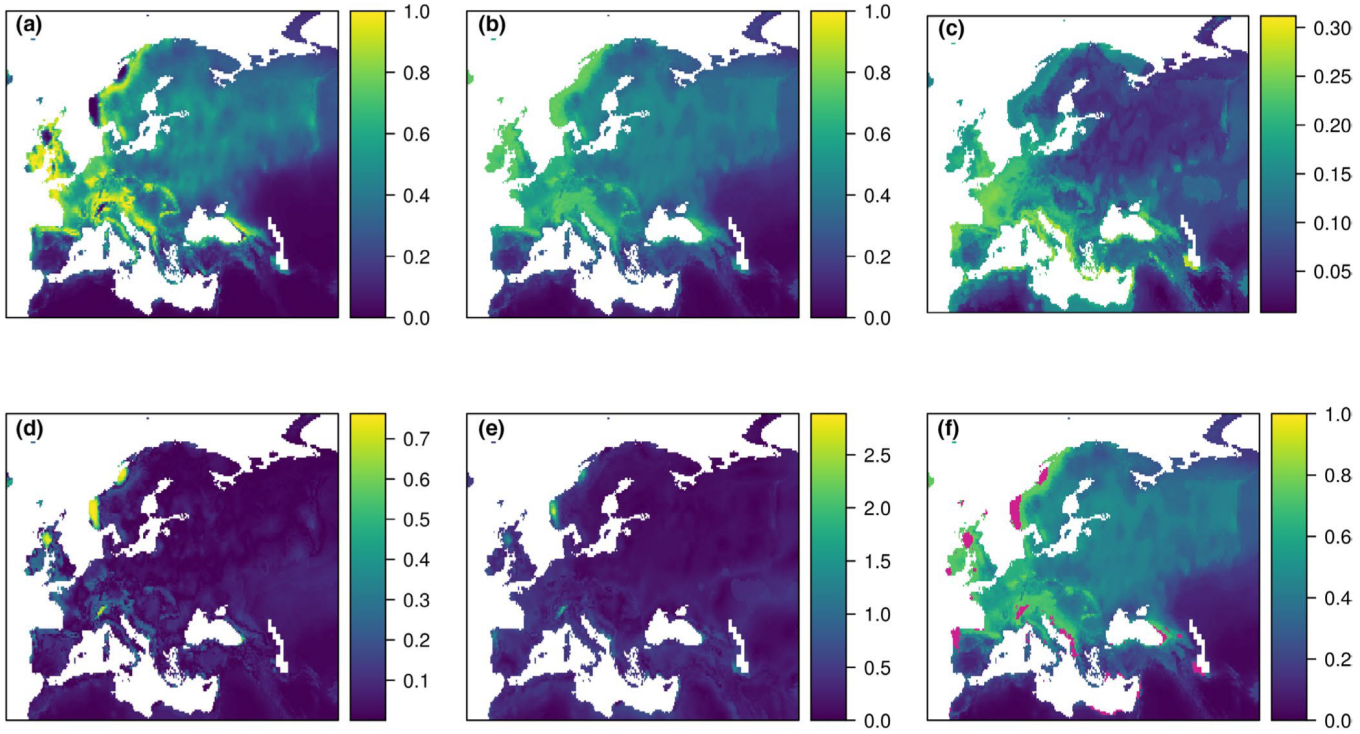

FIGURE 6 Comparison between reference (a), prediction (b), standard deviation of predictions (c), the true absolute prediction error (d), the newly suggested dissimilarity index (e) and the predictions masked by the derived area of applicability, where the areas outside the area of applicability are shown in pink (f)

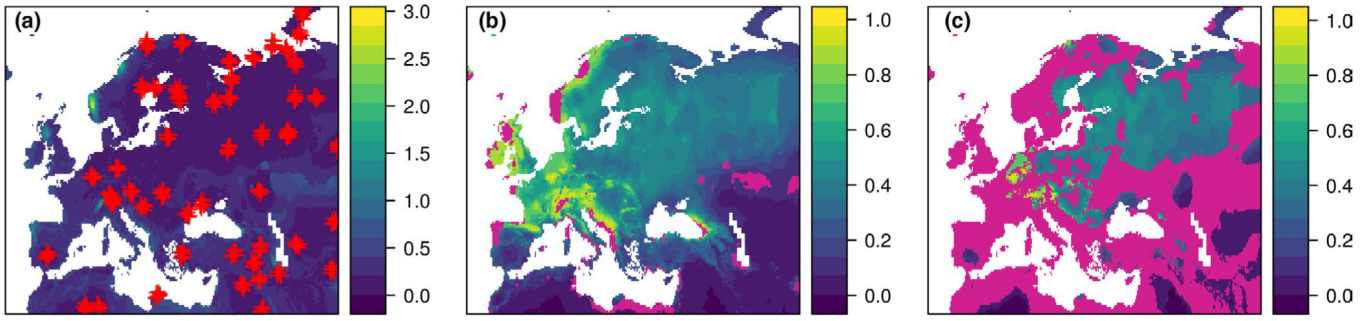

FIGURE 7 Example of an extremely clustered sampling design and the consequences for the estimation of the area of applicability: the dissimilarity index overlaid by 500 training data points clustered around 50 locations (a), predictions for the derived area of applicability for which the spatial cross-validation error applies (b), as well as predictions for the area of applicability for which the (lower, but in this case inappropriate) random cross-validation error applies (c). Areas outside the area of applicability are shown in pink

challenge'. The methodology to estimate the AOA as presented here provides one suggestion to assess the transferability by quantifying the differences in the environmental conditions between training data and target area and identifying the area for which the model can be expected to make predictions with an error comparable to the communicated model performance. Limiting predictions to the AOA is especially relevant when predictions, along with cross-validation-based error estimates, are used as a baseline for decision-making (e.g. in the context of nature conservation). It is further of high relevance when a prediction map is used for subsequent modelling to limit the propagation of massive errors. The frequently used global soil maps of Hengl et al. (2017), for example, represent a basis for many subsequent environmental prediction models (e.g. for mapping soil organisms in van den Delgado-Baquerizo et al., 2018; Hoogen et al., 2019). The presented approach would allow assessing the spatial suitability of these products, and allow for constraining further application to the estimated AOA or to the area where a required performance applies within the AOA.

The AOA adds an important information to the validation metrics based on test data taken from a biased sample or via cross-validation. First, the model validation usually provides a global estimate that does not allow for representing the varying performance of the model in a spatial way. Second, (cross-)validation estimates are based on the sample data only, but spatial sampling is usually biased (e.g. Bystriakova et al., 2012; Kadmon et al., 2004) and is unlikely to cover the entire environmental conditions of heterogeneous environments even when a sampling design is planned with the aim to cover the range of the predictor space (Hengl et al., 2003). Subsequently the validation that

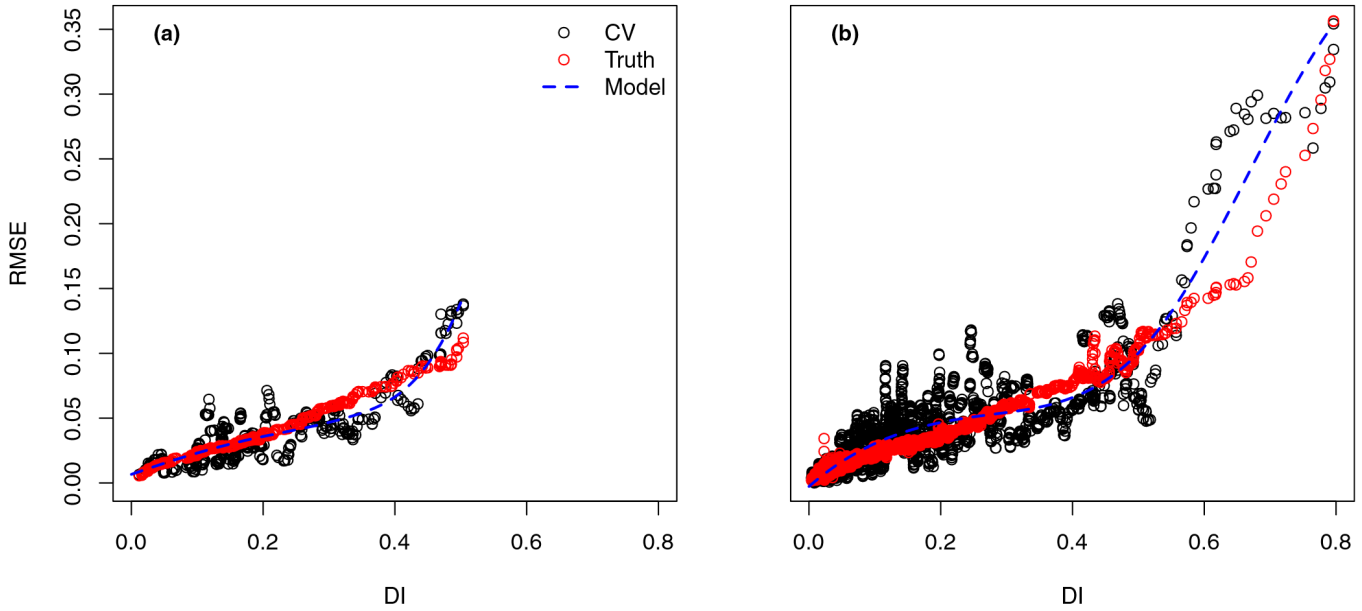

FIGURE 8 Relationship between RMSE and dissimilarity index (DI), computed up to DI threshold values, based on the case study scenario using the 500 spatially clustered training data and single spatial cross-validation (a) as well as the results for 10 different cross-validations, each using different clustering in predictor space (b). Each data point corresponds to the RMSE calculated in a sliding window of size 10 along the DI axis. The fitted model that is used to estimate model performance based on the DI is shown in blue (see also Figure 9). The true RMSE which was calculated using the reference map and corresponding predictions within the identical windows of DI values is shown in red

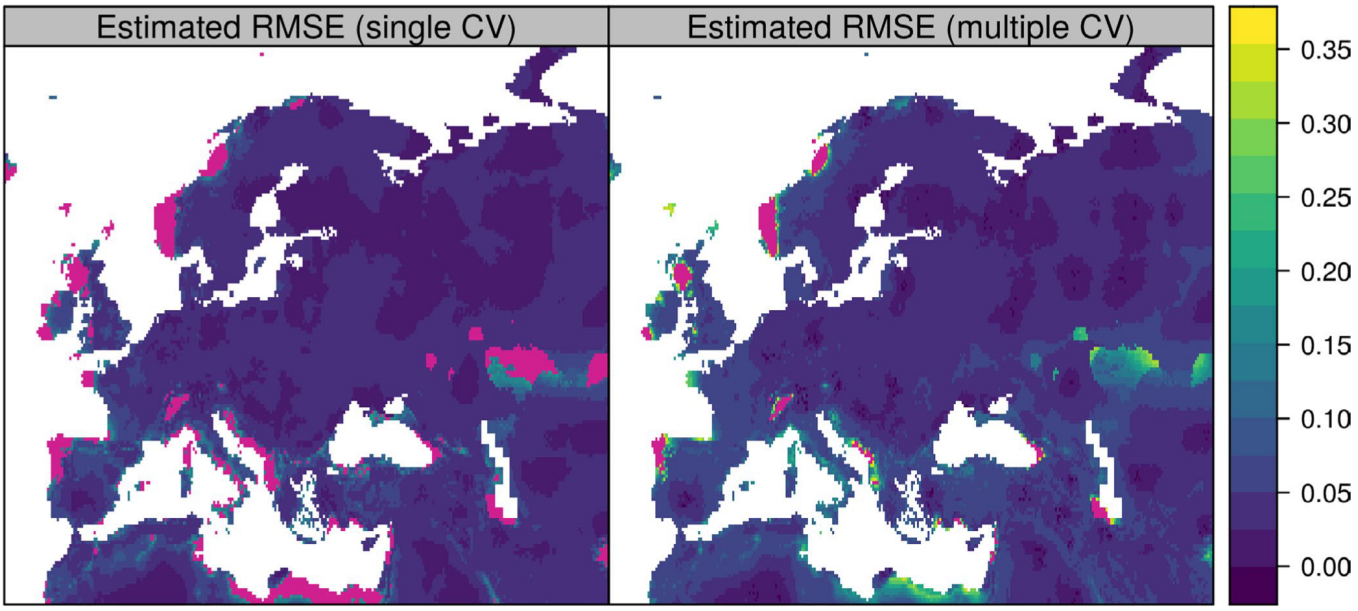

FIGURE 9 Spatial patterns of the estimated RMSE based on the relationship between dissimilarity index (Figure 7a) and RMSE from single or multiple cross-validations as shown in Figure 8. Areas outside the area of applicability are shown in pink

is based on independent subsets of the samples will be biased as well. One might argue that this is rather a problem of sampling strategies that need to be improved in the first place rather than addressing this issue by mapping the AOA of the model. However, the idea of machine learning for spatial mapping is that we are able to deal with complex relationships and a large variety of potential predictor variables. This challenges sampling by expert knowledge because gaps in the predictor space are hard to identify in high-dimensional predictor spaces. Also, in many prediction tasks, data from large composite databases are used that lack a common or shared sampling design, or for which the sampling design is unknown.

The uncertainty originating from missing knowledge about environments is also not reflected by standard deviations of predictions made by individual predictions of an ensemble. They give valuable

insights into the model by indicating areas where the model is very sensitive to changes in the data or randomness, hence may also reflect low training data point densities, but do not provide information on the AOA (see again Figure 1b). The results showed that it is required to account for incomplete coverage of environmental properties and to limit predictions to areas that are similar in their predictor properties compared to the training data and are therefore within the AOA.

As it is derived from distances in predictor space encountered during cross-validation, the AOA depends on distances considered as a result of the folding strategy followed during cross-validation. The distances are strongly influenced by the spatial pattern of sampling. In cases where training data are spatially clustered (e.g. when pixels are obtained from training polygons for land cover classifications, see Meyer et al., 2019), a naive strategy using random cross-validation folds effectively evaluates how well values in a cluster can be predicted based on other observations from the same cluster, and performance measures obtained this way do not reflect prediction beyond the cluster (e.g. Meyer et al., 2018, 2019; Pohjankukka et al., 2017; Roberts et al., 2017; Schratz et al., 2019; Valavi et al., 2018). In such cases, the AOA reflects this and renders the model 'not applicable' to most of the target area (Figure 7c). Alternative cross-validation strategies, for example using spatial blocks for folding, result in larger model error estimates corresponding to a much larger AOA (Figure 7b). Choosing the cross-validation strategy can be a tool to arrive at an AOA that corresponds to a user-specified average prediction error. If folds are chosen in such a way that the distances considered are larger, for example by choosing spatial blocks (Figure 8a) or by defining folds based on clustering points in prediction space for a varying number of clusters (Figure 8b), we find a relationship between DI and model error. This can be used to choose a DI threshold for the AOA such that a user-specified error level is attained. This approach is limited to the range of DI values with sufficient replicates in the training data (maximum AOA).

The application of simulated response variables was required to validate the proposed methodology. Note that the simulations applied here lead to very strong prediction models where the simulated response is a clear function of the predictors. Therefore, prediction errors can, to a large degree, be traced back to missing coverage in the environmental predictors. The relationship between the DI and a true error will be less strong for weak prediction tasks because missing knowledge of the environment will not be the major source of uncertainty. Other factors, especially a poor ability of the predictors to model the response, also influence uncertainty. This is not considered in the DI calculation, but reflected by the (cross-)validation error. Also note that high differences between training data and new data do not necessarily lead to a high prediction error. Instead, locations with a high DI, falling outside the AOA, are associated with a high uncertainty because the environment, and hence the prediction success, is unknown (see also the Supporting Information in the Appendix). Similarly, the method provides no guarantee that predictions within the AOA are reliable. Other factors that have not been considered as predictors might influence the response in certain environments. If these have not been considered in the model, they are likewise not considered in its AOA. Further, certain environments within the AOA might not be sufficiently covered in the training data to derive reliable predictions, which additionally requires additional uncertainty measures, for example standard deviations of predictions in an ensemble. Therefore, we see the AOA as a precondition for reliable predictions, but it should in no way be regarded as a guarantee for this.

The DI and the derived AOA do not only provide relevant information for estimating the reliability of predictions but can also serve model improvement. The uncertainty originating from missing knowledge of the model represents a reducible part of the total prediction uncertainties, because it is based on the training data availability. Knowledge of the AOA allows to improve the model quality by targeting subsequent sampling effort to improve the data basis. Therefore, the suggested DI can be used to identify the environments that are not covered by training data and hence can be used as a baseline for further sampling campaigns with the aim to increase the AOA of a model. Since the estimation of the AOA requires the predictive model for variable weighting only, the effect of new samples on the AOA can be assessed without high computation times. It can also be an option that variable weighting is done by expert elicitation prior to a modelling procedure so that the approach can be deployed in the early stages of a research project starting with the selection of sampling locations.

The method to estimate the AOA as presented here should be considered a first attempt and contains a number of aspects that are up for discussion. These include:

1. The use of distances in a weighted predictor space; weighting effectively alleviates the curse of dimensionality, but lacks a formal statistical argument,
2. The estimation of variable importance used for weighting; there are different strategies that will lead to different results in the estimation of the AOA. This is a very general issue in machine learning applications that goes beyond the scope of this paper. Here, we recommend that users of the AOA should use the method that is regarded as the most accurate for the respective algorithm being used,
3. The use of Euclidean distance; monotone transformation (e.g. log or power transforms) of predictors would not affect a Random Forest model fit or prediction, but would strongly affect the AOA,
4. The use of the nearest training data point d_k; this does not discriminate between cases where one isolated, remote training point is nearest, or a predictor space location is surrounded by training points at this same distance; as an alternative, distances to multiple points (k-NN) could be used (e.g. Sahigara et al., 2013, in the context of chemical modelling), or local training data point densities (e.g. Aniceto et al., 2016). Though the uncertainty caused by data point density is already reflected in ensemble-based prediction intervals (see Figure 1b), which should also be presented

alongside predictions, it remains to be tested if considering data point densities can be included in the delineation of the AOA,

5. The suitability of the DI as a quantitative uncertainty measure; we assume that a universal formal relationship between DI and performance cannot be simply provided because it depends in large parts on the respective model complexity among other factors. Here we suggest fitting the relationship with shape-constrained additive models in moving windows of DI values but suggest more research using diverse datasets. We also suggest that (if computation times allow) multipurpose cross-validation is applied to model the relationship between DI and performance which to our knowledge has not been suggested before in the context of model validation and uncertainty assessment. We are not suggesting to replace (ensemble-based) uncertainty estimates such as prediction intervals or standard deviations of predictions. Instead, we see DI and AOA as a useful addition to existing measures to account for a relevant source of uncertainty that has not been considered in the more commonly applied measures,
6. The applicability across machine learning algorithms; here, the Random Forest algorithm was used inspired by its multiple applications in the context of global mapping. However, the problem of predicting beyond the data applies to other algorithms as well. Though not explicitly studied here, the approach should be applicable to other machine learning algorithms in the same way (if variable importance can be estimated) and is also not restricted to spatial data. Further studies are needed here to confirm this.

Hence, the results shown here should be considered as a baseline for ongoing discussions on this topic. The methodology to estimate the AOA has been implemented and published in the R package CAST (Meyer, 2021a). The simulation studies are available as open source R scripts, and can be easily modified to other simulation models and/or other spatial sampling designs.

5 | CONCLUSIONS

We proposed a simple approach to map the AOA of spatial prediction models, which is the area that features predictor properties the model was enabled to learn about. The AOA is, as a consequence, the area where the model is expected to make predictions with an expected error that is comparable to the cross-validation error of the model. Predictions outside the AOA should be handled with care or be left out from further consideration because the environmental properties differ too strongly from those observed in the training data. Communicating the AOA is important to avoid misplanning when predictive mapping is used as a tool for decision-making (e.g. in the context of nature conservation), as well as to avoid propagation of massive errors when spatial predictions are used as input for subsequent modelling. We believe that the method proposed in this study will support critical assessment of overly optimistic data-driven prediction maps. We therefore suggest that the AOA should be provided alongside the prediction map and complementary to the communication of (cross-)validation performance measures and commonly applied (ensemble-based) prediction errors or intervals.

ACKNOWLEDGEMENT

Open Access funding enabled and organized by Projekt DEAL.

AUTHORS' CONTRIBUTIONS

H.M. and E.P. conceived the ideas, conducted the study and wrote the manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13650.

DATA AVAILABILITY STATEMENT

The methodology to estimate the AOA has been implemented and published in the R package CAST (Meyer, 2021a) which is available on CRAN. The simulation studies and all figures of this paper can be reproduced using the R-markdown scripts available from https://doi.org/10.5281/zenodo.4764404 (Meyer, 2021b; or use the developer version: https://github.com/HannaMeyer/MEE_AOA). This repository further includes more detailed figures on the methodology.

ORCID

Hanna Meyer https://orcid.org/0000-0003-0556-0210
Edzer Pebesma https://orcid.org/0000-0001-8049-7069

REFERENCES

Aniceto, N., Freitas, A. A., Bender, A., & Ghafourian, T. (2016). A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: Reliability-density neighbourhood. *Journal of Cheminformatics*, *8*(1), 69. https://doi.org/10.1186/s13321-016-0182-y

Bastin, J.-F., Finegold, Y., Garcia, C., Mollicone, D., Rezende, M., Routh, D., Zohner, C. M., & Crowther, T. W. (2019). The global tree restoration potential. *Science*, *365*(6448), 76–79. Retrieved from https://science.sciencemag.org/content/365/6448/76

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

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In *2012 IEEE International Geoscience and Remote Sensing Symposium* (pp. 5372–5375). IEEE. https://doi.org/10.1109/IGARSS.2012.6352393

Bystriakova, N., Peregrym, M., Erkens, R. H., Bezsmertna, O., & Schneider, H. (2012). Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. *Systematics and Biodiversity*, *10*(3), 305–315. https://doi.org/10.1080/14772000.2012.705357

Coulston, J. W., Blinn, C. E., Thomas, V. A., & Wynne, R. H. (2016). Approximating prediction uncertainty for random forest regression models. *Photogrammetric Engineering & Remote Sensing*, *82*(3), 189–197. https://doi.org/10.14358/PERS.82.3.189

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., Maestre, F. T., Singh, B. K., & Fierer, N. (2018). A global atlas of the dominant bacteria found

in soil. *Science*, *359*(6373), 320–325. https://doi.org/10.1126/science.aap9516

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, *1*(4), 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Gadaleta, D., Mangiatordi, G. F., Catto, M., Carotti, A., & Nicolotti, O. (2016). Applicability domain for QSAR models: Where theory meets reality. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, *1*(1), 45–63. https://doi.org/10.4018/IJQSPR.2016010102

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). Soilgrids250m: Global gridded soil information based on machine learning. *PLoS ONE*, *12*, e0169748. https://doi.org/10.1371/journal.pone.0169748

Hengl, T., Nussbaum, M., Wright, M., Heuvelink, G., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518.

Hengl, T., Rossiter, D., & Stein, A. (2003). Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Australian Journal of Soil Research*, *41*(8). https://doi.org/10.1071/SR03005

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Janet, J. P., Duan, C., Yang, T., Nandy, A., & Kulik, H. J. (2019). A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chemical Science*, *10*, 7913–7922. https://doi.org/10.1039/C9SC02298H

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, *14*(2), 401–413. https://doi.org/10.1890/02-5364

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. Retrieved from https://www.jstatsoft.org/v028/i05

Kuhn, M. (2019). *caret: Classification and regression training*. R package version 6.0-84. Retrieved from https://CRAN.R-project.org/package=caret

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*(6), 599–607. https://doi.org/10.1111/ecog.01388

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*, *2*(3), 18–22.

Mathea, M., Klingspohn, W., & Baumann, K. (2016). Chemoinformatic classification methods and their applicability domain. *Molecular Informatics*, *35*(5), 160–180. https://doi.org/10.1002/minf.201501019

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999.

Mesgaran, M. B., Cousens, R. D., & Webber, B. L. (2014). Here be dragons: A tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions*, *20*(10), 1147–1159. https://doi.org/10.1111/ddi.12209

Meyer, H. (2021a). *CAST: 'caret' applications for spatial-temporal models*. R package version 0.5.0. Retrieved from https://CRAN.R-project.org/package=CAST

Meyer, H. (2021b). *MEE_AOA: code to run the case study for the estimation of the area of applicability published in MEE*. https://doi.org/10.5281/zenodo.4764404

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, *101*, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*. https://doi.org/10.1016/j.ecolmodel.2019.108815

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.

Netzeva, T. I., Worth, A., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D., Schultz, T., Stanton, D. W., van de Sandt, J. J. M., Tong, W., … Yang, C. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM workshop 52. *Alternatives to Laboratory Animals*, *33*(2), 155–173. https://doi.org/10.1177/026119290503300209

Pelletier, T. A., Carstens, B. C., Tank, D. C., Sullivan, J., & Espndola, A. (2018). Predicting plant conservation priorities on a global scale. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(51), 13027–13032. https://doi.org/10.1073/pnas.1804098115

Petchey, O. L., Pontarp, M., Massie, T. M., Kéfi, S., Ozgul, A., Weilenmann, M., Palamara, G. M., Altermatt, F., Matthews, B., Levine, J. M., Childs, D. Z., McGill, B. J., Schaepman, M. E., Schmid, B., Spaak, P., Beckerman, A. P., Pennekamp, F., & Pearse, I. S. (2015). The ecological forecast horizon, and examples of its uses and determinants. *Ecology Letters*, *18*(7), 597–611. https://doi.org/10.1111/ele.12443

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., & Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, *11*(1), 4540. https://doi.org/10.1038/s41467-020-18321-y

Pohjankukka, J., Pahikkala, T., Nevalainen, P., & Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial *k*-fold cross validation. *International Journal of Geographical Information Science*, *31*(10), 2001–2019. https://doi.org/10.1080/13658816.2017.1346255

Pya, N. (2020). *scam: Shape constrained additive models*. R package version 1.2-9. Retrieved from https://CRAN.R-project.org/package=scam

Pya, N., & Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing*, *25*(3), 543–559. https://doi.org/10.1007/s11222-013-9448-7

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2013). Defining a novel *k*-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *Journal of Cheminformatics*, *5*(1), 27. https://doi.org/10.1186/1758-2946-5-27

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Sheridan, R. P., Feuston, B. P., Maiorov, V. N., & Kearsley, S. K. (2004). Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and*

Computer Sciences, *44*(6), 1912–1928. https://doi.org/10.1021/ci049782w

Toplak, M., Močnik, R., Polajnar, M., Bosnić, Z., Carlsson, L., Hasselgren, C., Demšar, J., Boyer, S., Zupan, B., & Stålring, J. (2014). Assessment of machine learning reliability methods for quantifying the applicability domain of QSAR regression models. *Journal of Chemical Information and Modeling*, *54*(2), 431–441. https://doi.org/10.1021/ci4006595

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). blockcv: An R package for generating spatially or environmentally separated folds for *k*-fold cross-validation of species distribution models. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2018/06/28/357798

van den Hoogen, J., Geisen, S., Routh, D., Ferris, H., Traunspurger, W., Wardle, D. A., de Goede, R. G. M., Adams, B. J., Ahmad, W., Andriuzzi, W. S., Bardgett, R. D., Bonkowski, M., Campos-Herrera, R., Cares, J. E., Caruso, T., de Brito Caixeta, L., Chen, X., Costa, S. R., Creamer, R., … Crowther, T. W. (2019). Soil nematode abundance and functional group composition at a global scale. *Nature*, *572*(7768), 194–198. https://doi.org/10.1038/s41586-019-1418-6

Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, *291*, 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., … Sequeira, A. M. M. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001

Zhu, A. X., Liu, J., Du, F., Zhang, S. J., Qin, C. Z., Burt, J., Behrens, T., & Scholten, T. (2015). Predictive soil mapping with limited sample data. *European Journal of Soil Science*, *66*(3), 535–547. https://doi.org/10.1111/ejss.12244

Zohner, C. M., Mo, L., Renner, S. S., Svenning, J.-C., Vitasse, Y., Benito, B. M., Ordonez, A., Baumgarten, F., Bastin, J.-F., Sebald, V., Reich, P. B., Liang, J., Nabuurs, G.-J., de Miguel, S., Alberti, G., Antón-Fernández, C., Balazy, R., Brändli, U.-B., Chen, H. Y. H., … Crowther, T. W. (2020). Late-spring frost risk between 1959 and 2017 decreased in North America but increased in Europe and Asia. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(22), 12192–12200. https://doi.org/10.1073/pnas.1920816117

Zurell, D., Elith, J., & Schröder, B. (2012). Predicting to new environments: Tools for visualizing model behaviour and impacts on mapped distributions. *Diversity and Distributions*, *18*(6), 628–634. https://doi.org/10.1111/j.1472-4642.2012.00887.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12, 1620–1633. https://doi.org/10.1111/2041-210X.13650