Received: 8 September 2022 | Revised: 25 November 2022 | Accepted: 13 December 2022

DOI: 10.1111/geb.13635

Global Ecology and Biogeography | A Journal of Macroecology

WILEY

METHOD

Assessing and improving the transferability of current global spatial prediction models

Marvin Ludwig[1,2] | Alvaro Moreno-Martinez[3] | Norbert Hölzel[1] | Edzer Pebesma[2] | Hanna Meyer[1]

[1]Institute for Landscape Ecology, University of Münster, Münster, Germany

[2]Institute of Geoinformatics, University of Münster, Münster, Germany

[3]Image Processing Laboratory (IPL), Universitat de Valéncia, Paterna, Spain

Correspondence
Marvin Ludwig, Institute for Landscape Ecology, University of Münster, Münster, Germany.
Email: marvin.ludwig@uni-muenster.de

Funding information
ERC-SyG-2019 USMILE, Grant/Award Number: 855187; Federal Ministry for Economic Affairs and Climate Action of Germany, Grant/Award Number: 50EE2009

Handling Editor: Lenore Fahrig

Abstract

Aim: Global-scale maps of the environment are an important source of information for researchers and decision makers. Often, these maps are created by training machine learning algorithms on field-sampled reference data using remote sensing information as predictors. Since field samples are often sparse and clustered in geographic space, model prediction requires a transfer of the trained model to regions where no reference data are available. However, recent studies question the feasibility of predictions far beyond the location of training data.

Innovation: We propose a novel workflow for spatial predictive mapping that leverages recent developments in this field and combines them in innovative ways with the aim of improved model transferability and performance assessment. We demonstrate, evaluate and discuss the workflow with data from recently published global environmental maps.

Main conclusions: Reducing predictors to those relevant for spatial prediction leads to an increase of model transferability and map accuracy without a decrease of prediction quality in areas with high sampling density. Still, reliable gap-free global predictions were not possible, highlighting that global maps and their evaluation are hampered by limited availability of reference data.

KEYWORDS
area of applicability, global maps, machine learning, model simplification, model transferability, spatial prediction

1 | INTRODUCTION

Global-scale maps are a relevant source of information to gain knowledge about the environment, monitor changes and support decisions (e.g., Schmidt-Traub, 2021; Wyborn & Evans, 2021). In the past few years there has been a considerable increase in studies that produced global maps of environmental properties on the basis of machine learning algorithms, such as maps of land cover (Buchhorn et al., 2020; Venter & Sydenham, 2021), canopy height (Lang et al., 2022), soil properties (Hengl et al., 2017), biomass (Ma et al., 2021; Rodríguez-Veiga et al., 2017), abundances of species (e.g., nematodes, see van den Hoogen et al., 2019), and potential tree cover (Bastin et al., 2019). The current popularity of machine learning-based global mapping studies can be attributed to

1. an increase in open-access global databases that provide a large amount of reference samples, for example, on plant traits (TRY, Kattge et al., 2020), soil properties (WoSIS, Batjes et al., 2020), and species occurrence from citizen science (https://www.gbif.org);
2. the increasing availability of global environmental base layers, for example, on climate (Fick & Hijmans, 2017) and soil properties (Hengl et al., 2017) as well as long-term satellite imagery;
3. the documented success of machine learning in remote sensing and related fields (e.g., reviews from Holloway & Mengersen, 2018; Lary et al., 2016; Scowen et al., 2021);
4. increasingly easy access to computational resources via platforms like Google Earth Engine (Amani et al., 2020) or openEO (Schramm et al., 2021) and promotion of fully automated pipelines for global mapping in the cloud (van den Hoogen et al., 2021).

Most of the global maps mentioned above were created following a similar logic and workflow (in its basic components comparable to the workflow described in van den Hoogen et al., 2021). Reference data of a target variable are acquired at spatially explicit locations during field sampling campaigns or compiled from collaborative databases. The reference data get spatially matched with globally available environmental variables – the predictors. The relationship between the target variable and the predictors is then trained with a machine learning algorithm where random forest has prevailed as the most commonly used algorithm (e.g., Bastin et al., 2019; Buchhorn et al., 2020; Ma et al., 2021; Moreno-Martinez et al., 2018; van den Hoogen et al., 2019; Venter & Sydenham, 2021). Once the model is trained, it is applied to the global predictor set to map the target variable.

This approach is based on the assumption that the predictors are a representation of the environment and that we train the statistical relationships with the machine learning model between the represented environment and the target variable from the reference data. Global predictions therefore require that the reference samples provide sufficient opportunities to train globally applicable relationships. However, field surveys are often conducted in opportunistic locations (Yates et al., 2018) and are biased towards populated or protected areas (Martin et al., 2012), leading to strong clustering of reference data in areas that received intensive research (especially western countries) but extremely poor coverage in other regions (especially global south, former USSR). These knowledge gaps in parts of the global geographic domain and consequently in the environmental predictor space might also hamper spatial predictions. While technically a trained model can be applied globally as long as global predictors are available, machine learning models typically produce meaningless predictions when provided with predictor values that do not resemble the reference data. This applies to extrapolation situations outside the value range of predictors (Barbiero et al., 2020) as well as to gaps within the multivariate predictor space (Meyer & Pebesma, 2021).

Clustered reference data also imply challenges for quality assessment of the spatial predictions. Ideally, map accuracy is estimated by a probability sample and design-based inference (Wadoux et al., 2021). In most cases, however, it is not possible to derive a suitable probability sample for the prediction domain from the clustered and sparse reference data. Instead, the model accuracy is estimated and often equated with the map accuracy. Model accuracy is typically assessed by simple train–test splits of the data or by cross-validation where in both cases, prediction situations are created artificially by leaving out data points and predicting their value from a model trained on the remaining points. The derived model accuracy, however, heavily depends on how the artificial prediction situations are created – that is, how the training data are divided into folds. There is an ongoing discussion about which cross-validation strategy is most suitable for the estimation of the map accuracy. While Ploton et al. (2020) shows that the commonly applied random cross-validation is not appropriate in the presence of clustered reference data, Wadoux et al. (2021) argue that spatial cross-validation might result in overly pessimistic estimates of map accuracy. Milà et al. (2022) resolve the discussion insofar as they suggest that the defined cross-validation folds should resemble the required prediction conditions in order to derive a model accuracy that can be used as an approximation of map accuracy. Still, Meyer and Pebesma (2021) show that the derived model accuracy can only be a proxy for map accuracy in regions where predictor values resemble the predictor values of the reference data. Therefore, Meyer and Pebesma (2021, 2022) suggest that predictions should be limited to the area of applicability of the model, which is the area where the model was enabled to learn about relationships and where the estimated cross-validation error holds. Predictions outside the area of applicability should be avoided, since we cannot expect reasonable predictions due to the inability of machine learning models to deal with extrapolation (in the predictor space) and since we have no certainty on how accurate the model predictions are at those locations.

A high spatial model error and/or a small area of applicability is an indication for poor model generalization. The model might well reproduce the clustered training locations (which can be tested by random cross-validation) but cannot make reliable predictions beyond them. This can have obvious reasons like reference data that are too sparse to cover a sufficient range of environments, which is reflected by a small area of applicability. Poor generalizability can also originate from overly complex models that are overfitted to the specific environments they were trained on (Barbiero et al., 2020). The greater the number of predictors in the model, the more complex and specialized the trained relationships will be (Hassine et al., 2019; Roe et al., 2020). This leads to an increased risk of overfitting and a high probability that new locations contain predictor combinations that are not reflected in the model (i.e., outside the area of applicability). A well-established strategy to prevent overfitting is to simplify the model by removing predictors that have little effect on the outcome (Kuhn & Johnson, 2013; Merow et al., 2014) – an approach that is also utilized in global mapping studies (e.g., Bastin et al., 2019; Moreno-Martinez et al., 2018). In the context of spatial mapping, an effective variable reduction strategy limits predictors

to those that are best suited to make predictions for new locations, that is, beyond clustered reference data (Le Rest et al., 2014; Meyer et al., 2018, 2019). In order to select predictor variables suitable for spatial prediction, Meyer et al. (2018) suggests to use forward variable selection in conjunction with a cross-validation strategy that is appropriate for the respective study. While variable selection with random cross-validation leads to variables that can well reproduce clustered training data but might not be suitable for spatial prediction due to the previously mentioned reasons, in contrast, 'spatial variable selection' that uses spatial cross-validation for selection will favour variables that lead to a high ability to make predictions for new areas.

Here we build upon the recently published method of the area of applicability (Meyer & Pebesma, 2021) and novel cross-validation designs (Milà et al., 2022) to address the concerns about adequate map validation stated in Meyer and Pebesma (2022) using real world examples. We specifically test whether spatial variable selection can be used to increase the applicability and accuracy of global prediction models. In order to test our approach, we utilize data from three recent global mapping studies: soil nematode abundances as modelled by van den Hoogen et al. (2019), potential tree cover presented in Bastin et al. (2019) and specific leaf area of plants based on Moreno-Martinez et al. (2018). We train random forest models, estimate their performance using random and spatial cross-validation, and we estimate the area of applicability. We then test if the accuracy and the area of applicability can be increased when models are trained on a reduced set of predictors. Therefore we conduct spatial variable selection to limit predictors to those that improve predictions beyond the location of the reference data. We compare the area of applicability as well as the estimated map accuracy of the simplified models to the ones that utilize all predictors.

2 | METHODS

To test the potential of model simplification, we train models based on data used in three previously published global prediction studies: soil nematode abundances by van den Hoogen et al. (2019), specific leaf area of plants by Moreno-Martinez et al. (2018) and potential tree cover by Bastin et al. (2019). The following sections overview the data used and denote differences from the original studies. We then describe the applied cross-validation strategies, the spatial variable selection, the estimation of the area of applicability, and map accuracy.

2.1 | Data and description of the three global prediction studies

2.1.1 | Soil nematodes

The original soil nematodes study is described in van den Hoogen et al. (2019). The authors provided their reference samples, the source of each predictor, as well as the scripts for Google Earth Engine and R alongside their paper. The original model is based on 1,876 reference samples and 73 global predictor variables, such as information of terrain properties, bio-climate (Fick & Hijmans, 2017), soil maps (Hengl et al., 2017) and surface reflectance in several spectral wavelengths as acquired by the Moderate Resolution Imaging Spectroradiometer (MODIS, see van den Hoogen et al., 2019, for the full list of predictor variables). We acquired all predictor layers according to the documented download source. In the original study, a gap filling algorithm was used in Google Earth Engine to fill missing values in the predictors. Since this could not be reproduced outside Google Earth Engine, we neglected the gap filling and restricted all analysis to the valid data range. Consequently, we ignored the few reference sample points that were located in the gaps, leading to a slightly reduced set of 1,523 reference samples. More than half of the reference samples were located in Europe (roughly 20% in the Netherlands alone) leading to severe spatial clustering of training locations (Figure 1).

2.1.2 | Specific leaf area

The original model of specific leaf area is presented in Moreno-Martinez et al. (2018). Since the publication, the authors have reprocessed and improved the map to minimize extrapolation and provide more robust estimates globally using random forests with a surrogate splits approach (Hothorn et al., 2006). Hence, the results of our study do not refer to the updated map presented here (https://isp.uv.es/code/try) but only the original publication. The authors used a multi-stage modelling workflow to provide community-weighted means of plant traits on a 250-m scale based on the integration of Landsat and MODIS as well as records in the TRY database (Kattge et al., 2020). The authors provided their reference data containing 5,867 samples of community-weighted means of specific leaf area and 15 global predictor layers. These 15 predictors were the best ranking variables out of 36 predictors (Moreno-Martinez et al., 2018). Reference samples are heavily clustered in Catalonia, Spain (over 4,200 points, 74% of all reference samples, see Figure 2). Predictors consisted of five bio-climate variables (Fick & Hijmans, 2017), Shuttle Radar Topography Mission (SRTM)-derived elevation, six surface reflectance bands from MODIS and three derived vegetation indices (two versions of the enhanced vegetation index and the normalized difference water index). In accordance with Moreno-Martinez et al. (2018), we excluded urban areas, water bodies, and vegetation-free areas from all analyses.

2.1.3 | Potential tree cover

The potential tree cover model from Bastin et al. (2019) was trained on 78,744 globally distributed samples (Figure 3) and 10 predictor variables. The authors assumed that nature conservation areas represent areas where the actual tree cover equals the potential tree cover. The

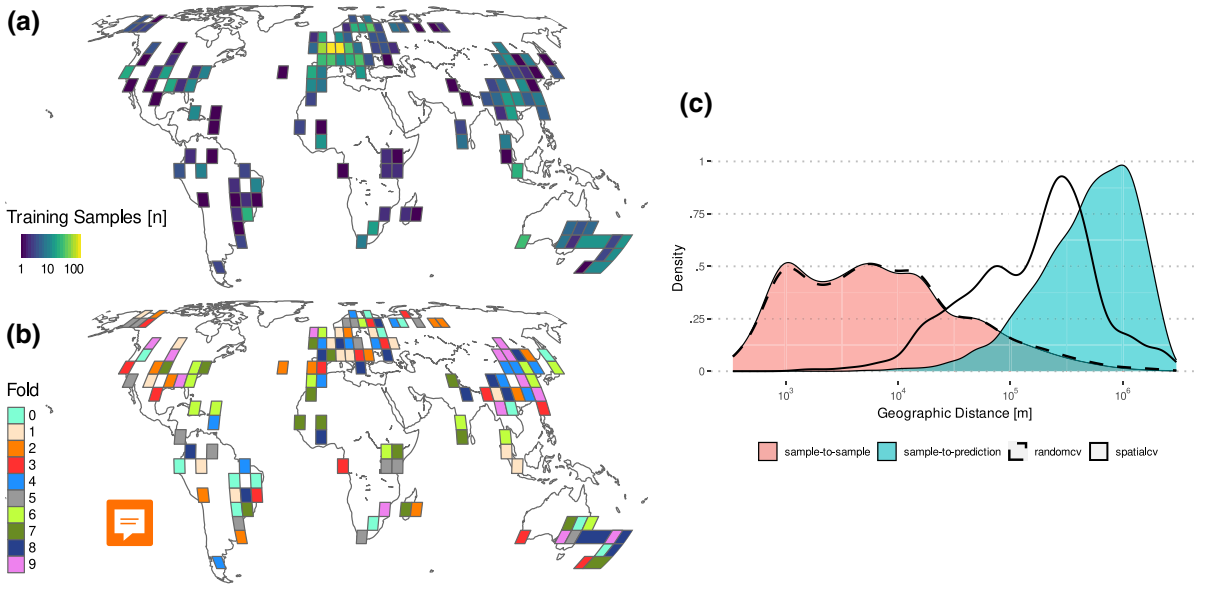

FIGURE 1 Distribution of reference data for soil nematodes used in van den Hoogen et al. (2019). Colours in (a) represent the density of training samples. (b) Shows information of the spatial cross-validation applied in this study: the global domain was divided into spatial blocks of 6° and randomly assigned to one of 10 folds (shown here in different colours) in order to resemble the prediction-to-sample distance (see Section 2.3 for details). (c) Compares the geographic distance between folds of the random (randomcv) and spatial cross-validation (spatialcv): while random folds reproduce the distances between samples (red), the spatial folds better resemble the distance from prediction locations to training samples (blue).

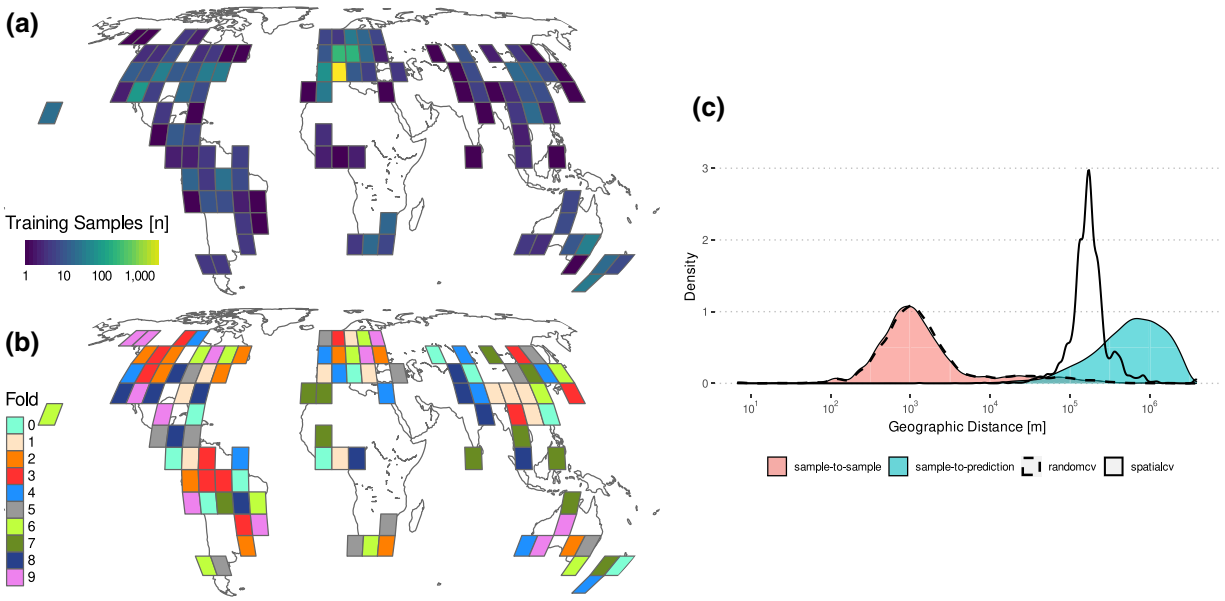

FIGURE 2 Distribution of reference data for specific lead area (SLA) used in Moreno-Martinez et al. (2018). Colours in (a) represent the density of training samples. (b) Depicts information of the spatial cross-validation applied in this study: the global domain was divided into spatial blocks of 9° and randomly assigned to one of 10 folds (shown here in different colours) in order to resemble the prediction-to-sample distance (see Section 2.3 for details). (c) Compares the geographic distance between folds of the random (randomcv) and spatial cross-validation (spatialcv): while random folds reproduce the distances between samples (red), the spatial folds better resemble the distance from prediction locations to training samples (blue).

reference samples hence provide actual tree cover from conservation areas that was derived via high resolution image classification. The authors provided their sample coordinates with the corresponding classified tree cover alongside the paper (Bastin et al., 2019). Out of 58 predictor variables, cluster-based variable selection resulted in five bio-climate variables (Fick & Hijmans, 2017), three soil parameters (Hengl et al., 2017) and SRTM-derived elevation and hillshade for the modelling. In contrast to the other two studies used here, the reference data of Bastin et al. (2019) are less clustered and cover a large part of the global modelling domain (Figure 3).

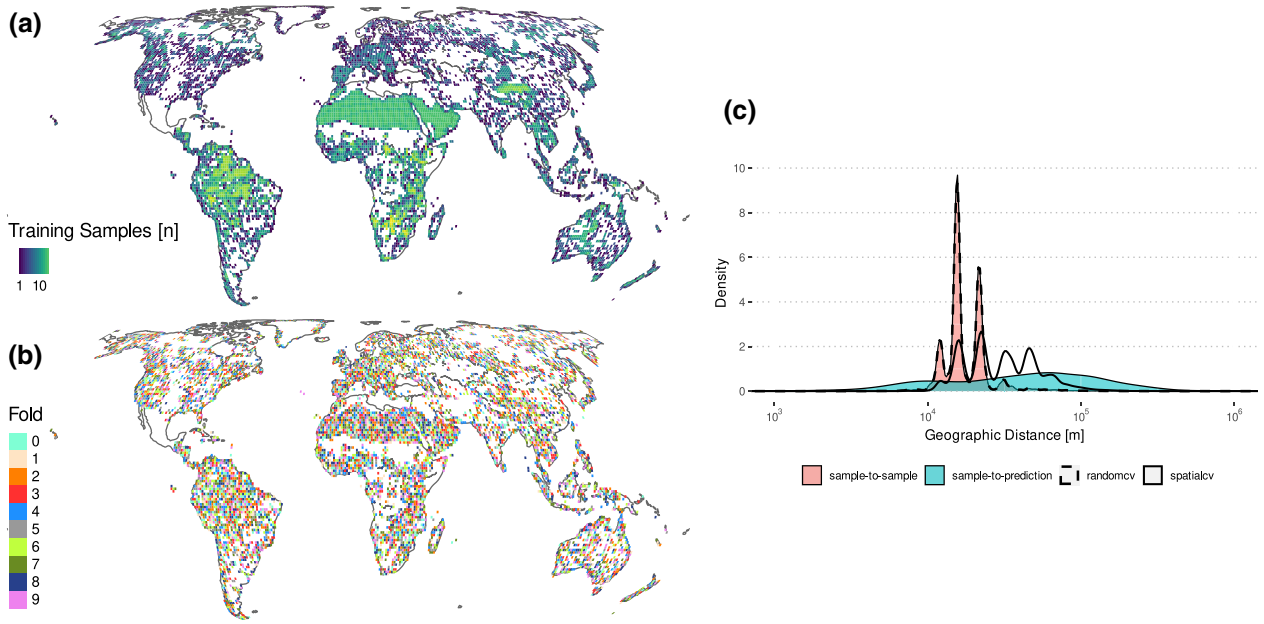

FIGURE 3 Distribution of reference data for potential tree cover used in Bastin et al. (2019). Colours in (a) represent the density of training samples. (b) Shows information of the spatial cross-validation applied in this study: the global domain was divided into spatial blocks of 1° and randomly assigned to one of 10 folds (shown here in different colours) in order to resemble the prediction-to-sample distance (see Section 2.3 for details). (c) Compares the geographic distance between folds of the random (randomcv) and spatial cross-validation (spatialcv): while random folds reproduce the distances between samples (red), the spatial folds better resemble the distance from prediction locations to training samples (blue). The three distinct density peaks stem from regular sampling strategies over larger areas used in Bastin et al. (2019).

2.2 | Model training

All three studies used random forest as a machine learning algorithm to learn relationships between predictors and the response variable. Here, we used the random forest implementation from the ranger package (Wright & Ziegler, 2017) and the caret framework (Kuhn, 2008) within R (v4.1.0, R Core Team, 2021). We computed 300 trees for soil nematodes and specific leaf area and 20 trees in the model of potential tree cover (as specified in Bastin et al., 2019). We neglected hyperparameter tuning since, for random forests, it commonly has a marginal impact on the model outcome (e.g., Schratz et al., 2019). The number of potential predictors per split (mtry) was set to 3 and the minimum number of samples per node was set to 5. All models were evaluated with three cross-validation strategies that are described in Section 2.3 below. R^2 and Root Mean Square Error (RMSE) values were calculated from all pairs of observed and predicted response values when held back from model training.

2.3 | Cross-validation strategies

Cross-validation is a key element of this study as it is used during spatial variable selection, the estimation of the model performance and the delineation of the area of applicability. We utilized three different strategies: (a) a random cross-validation as it was originally used in the three studies, (b) a spatial cross-validation which tests the ability of the model to predict unknown geographic locations and (c) a cross-validation based on clusters in the feature space to specifically evaluate the ability of the model for situations of extrapolation (in predictor space).

2.3.1 | Random cross-validation

For random cross-validation, the training data were randomly split into 10 folds (note that the original tree cover study used five folds; however, we expect that this deviation has negligible impact on the outcome and we preferred a consistent strategy over the three studies). In the presence of spatially clustered training data, random cross-validation only indicates the ability of the model to make prediction within the clusters (e.g., Meyer et al., 2019; Ploton et al., 2020; Roberts et al., 2017; Wenger & Olden, 2012). In contrast, spatial cross-validation strategies, where spatial units are held back for validation (e.g., Brenning, 2012; Ploton et al., 2020; Roberts et al., 2017), assess the ability of the model to make predictions beyond clustered training data.

2.3.2 | Spatial cross-validation

While the suitability of different spatial cross-validation methods has recently been under discussion (e.g., Wadoux et al., 2021), Milà et al. (2022) and Meyer and Pebesma (2022) argued that, in order to be a suitable cross-validation strategy, prediction situations during cross-validation need to resemble those encountered while predicting the global map

from the reference data. To approach this, we follow the ideas of Milà et al. (2022) and suggest to consider the Euclidean geographic distance distribution between training data and all prediction locations (pixels in the global domain), that is, how far away are the prediction locations from the training data? We then try to emulate this distribution during cross-validation with the aim that each fold consists of similar distances to the remaining reference data. Since the originally suggested method from Milà et al. (2022) is a variant of leave-one-out cross-validation and hence computation time expensive, we kept the idea of Milà et al. (2022) in a *k*-fold approach. To do this we applied a spatial block strategy (Valavi et al., 2019) where we divided the reference data into spatial blocks, which are then randomly grouped into 10 folds. We calculated the minimum spatial distance of each reference sample to the reference samples not included in the same fold ('between-fold distance') and compared it to the spatial distance between prediction locations and sampling locations ('sample-to-prediction'). We tested different block sizes and used the size where the between-fold distance best resembled the sample-to-prediction distance (Figures 1–3). This approach resulted in block sizes of 6° for the soil nematode samples, 9° for the specific leaf area and 1° for the potential tree cover.

2.3.3 | Feature-space cross-validation

We consider the described spatial cross-validation strategy as the most suitable with which to estimate map accuracy. To further test the ability of the models to extrapolate in the feature space, we additionally performed a feature-space cross-validation where entire parts of the multidimensional predictor space are held back. For this, we divided the reference data into three folds using a *k*-means clustering algorithm on the multivariate predictor space. Each fold therefore is highly dissimilar to the other two, which results in a performance estimate that applies to areas with new, unseen predictor values. While this cross-validation strategy is not suitable to estimate the overall map accuracy, it will be used as described in Section 2.6 to derive the maximum area for which an accuracy assessment is possible based on the available data.

2.4 | Spatial variable selection

We outlined above that models based on highly clustered data have a high risk of spatial overfitting especially when a large number of predictors is used. The three studies regarded here accounted for this problem with the inclusion of variable selection, although the presented soil nematode prediction in van den Hoogen et al. (2019) is based on all available predictors. We therefore test whether more accurate models can be developed when predictor variables are selected with regard to their potential for spatial transferability. To do so, we use forward variable selection in conjunction with the spatial cross-validation described in Section 2.3 to identify the set of predictors that performs best in foreign geographic space, following the methodology described in Meyer et al. (2018, 2019) and implemented in the R package 'CAST' (Meyer & Ludwig, 2022). The variable selection starts by selecting the best performing combination of two predictor variables, where performance is assessed by spatial cross-validation. It then gradually adds more predictors to the model. The algorithm stops when the spatial cross-validation model performance does not increase with the addition of any further predictor.

2.5 | Area of applicability

Most machine learning models, such as standard random forests, produce meaningless predictions and uncertainty estimates when provided with predictor values that do not resemble the reference data (Meyer & Pebesma, 2021). This applies to extrapolation situations outside the value range of predictors (Barbiero et al., 2020) as well as to gaps in the multivariate predictor space. Meyer and Pebesma (2021) therefore suggested that predictions should be limited to the area where the model was enabled to learn about relationships and where the estimated cross-validation performance holds – the area of applicability of the model.

The method to derive the area of applicability is based on distances to training data in the multivariate predictor space. Unlike other methods that assess extrapolation conditions by accounting for the range of observed predictor values (Moreno-Martinez et al., 2018; van den Hoogen et al., 2019), this method also accounts for gaps in the multivariate predictor space. To estimate the area of applicability a dissimilarity index is calculated for each pixel of the global prediction domain. The dissimilarity index is the normalized Euclidean distance to the nearest training data point in the multivariate predictor space, with predictors being scaled and weighted by their respective importance in the model (see Meyer & Pebesma, 2021, for more details on the calculation of the dissimilarity index). The area of applicability is then derived by applying a threshold to the dissimilarity index. The threshold is the (outlier-removed) maximum dissimilarity index of the training data derived via cross-validation. Hence, different cross-validation strategies can result in different thresholds – depending on how dissimilar the individual folds are from one another. A new data point (i.e., here pixel in the global prediction domain) is outside of the area of applicability if it is more dissimilar in its predictor properties than the dissimilarity observed between training data of different folds. Only inside the area of applicability the respective model accuracy from cross-validation can be regarded as a proxy for map accuracy. We assessed the area of applicability of all prediction models as described in Meyer and Pebesma (2021) and as implemented in the R package 'CAST' (Meyer & Ludwig, 2022).

2.6 | Pixel-wise accuracy estimation

Different cross-validation strategies lead to different performance estimates that may serve as a proxy for map accuracy in different areas (i.e., random cross-validation performance applies only to highly similar areas while in the extreme case of the threefold feature-space cross-validation, the estimated accuracy applies to areas that are

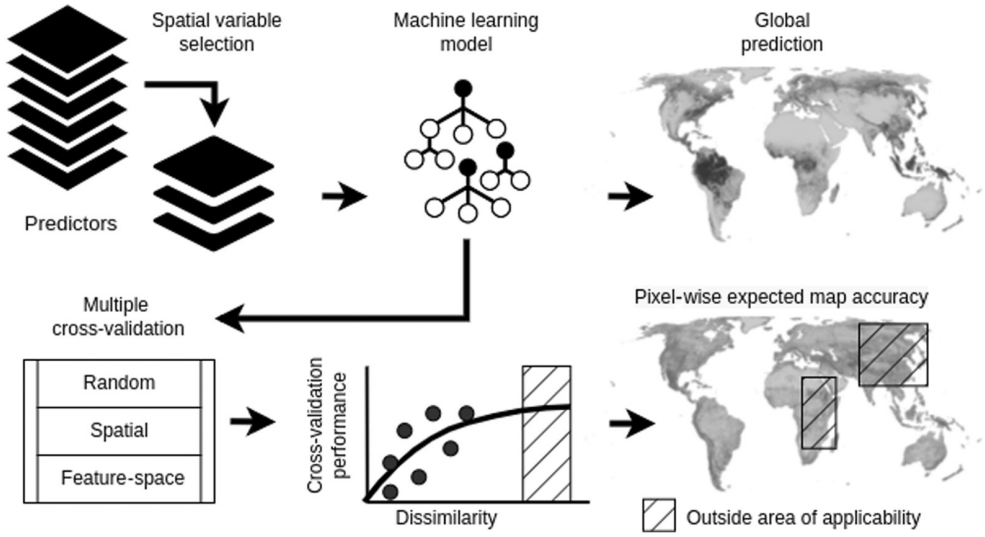

FIGURE 4 Overview of the modelling workflow and map accuracy assessment. Potential predictors are reduced by spatial variable selection for a simplified machine learning model used for the global prediction. The relationship between the dissimilarity index and the performance from multiple cross-validation strategies is modelled for a pixel-wise expected map accuracy. Pixels with predictors that are too dissimilar from training data are considered outside the area of applicability.

highly dissimilar). Since usually only one strategy is applied, the accuracy of the entire map is summarized by a single performance estimate. Instead, a spatially continuous, (i.e., pixel-wise) estimation of the expected error might be more favourable. For a spatially continuous representation of the expected accuracy, we assess the relationship between the dissimilarity index and the prediction performance as described in Meyer and Pebesma (2021). We combined the values of the dissimilarity index with the prediction errors of the left out folds as derived from the three cross-validation strategies (see Section 2.3) to include a wide range of dissimilarities. We then calculated the RMSE of predictions in a sliding window (25 observations) over the dissimilarity index. To model the relationship between the dissimilarity index and RMSE we used shape constrained additive models (see Figures S1–S3 and Pya, 2021). Applying this model to the dissimilarity index of every pixel in the global domain results in a spatially continuous and differentiated map of estimated accuracy values that can be expected based on the dissimilarity to the training data.

2.7 | Model comparison

For all three case studies we

1. trained models with the original set of predictor variables and assessed the success of our model training by comparing the random cross-validation R^2 to those communicated in the original studies;
2. compared the model performances of random and spatial cross-validation (difference between R^2) to assess the degree of spatial overfitting;
3. estimated the area of applicability of the random cross-validation performance to analyse for which areas the originally communicated model accuracy applies;
4. mapped accuracy estimates based on the relationship between the dissimilarity index and the model performance up to the maximum possible area of applicability (Figure 4);
5. reduced the number of predictors by spatial variable selection and compared the map accuracy and the area of applicability between the models using the full predictor set to models using the reduced predictor set in order to measure the benefits of spatial variable selection (Figure 4).

2.8 | Code availability

We want to highlight that this research was only possible because Bastin et al. (2019); Moreno-Martinez et al. (2018); van den Hoogen et al. (2019) made code and data available alongside their manuscripts. The code to reproduce the results presented in this study can be accessed at: https://github.com/LOEK-RS/global_applicability. The methods for the forward variable selection, area of applicability and dissimilarity index calibration are implemented in the R package 'CAST' (Meyer & Ludwig, 2022): https://cran.r-project.org/web/packages/CAST/index.html.

3 | RESULTS

3.1 | Performance and applicability of global prediction models

We trained random forest models based on the information given in the three publications and their supplemented code and data. Although an identical reproduction of the results was not possible (and not intended) due to different technical implementations of the algorithms and randomness involved in data splitting and model

training, we obtained similar random cross-validation R^2 values and prediction patterns for all three studies (Table 1).

We found large differences between random and spatial cross-validation R^2 values for the models of soil nematodes and specific leaf area. With spatial cross-validation R^2 values of .19 and .30, respectively, both models show clear indications of spatial overfitting when compared to their random cross-validation values (R^2 of .46 and .62, respectively). When cross-validated with random folds, both models had a very narrow area of applicability that was limited mainly to the geographic area around the available sample points (Figures 5a and 6a). The model of potential tree cover showed a better model performance in general and less discrepancy between the random (R^2 = .67) and spatial cross-validation (R^2 = .65). Further, the random cross-validation error of the model can be assumed to apply widely, with 92% of all prediction locations falling inside the area of applicability of the model (Figure 7a). Remaining non-applicable areas are mainly located in alpine areas and deserts where reference samples were sparse (Figure 7a).

Calibrating the dissimilarity index of the different cross-validation strategies allowed for a continuous representation of the

TABLE 1 Data and modelling information about the three global prediction studies on soil nematodes (van den Hoogen et al., 2019), potential tree cover (Bastin et al., 2019) and specific leaf area (Moreno-Martinez et al., 2018).

	Soil nematodes	Specific leaf area	Potential tree cover
Number of reference samples	1,523	5,867	78,774
Number of predictors	73	15	10
Published random CV R^2	.43	.59	.71
Reproduced random CV R^2	.46	.62	.67

Note: Comparison of model accuracies (R^2) based on a 10-fold random cross-validation strategy.
Abbreviation: CV, cross-validation.

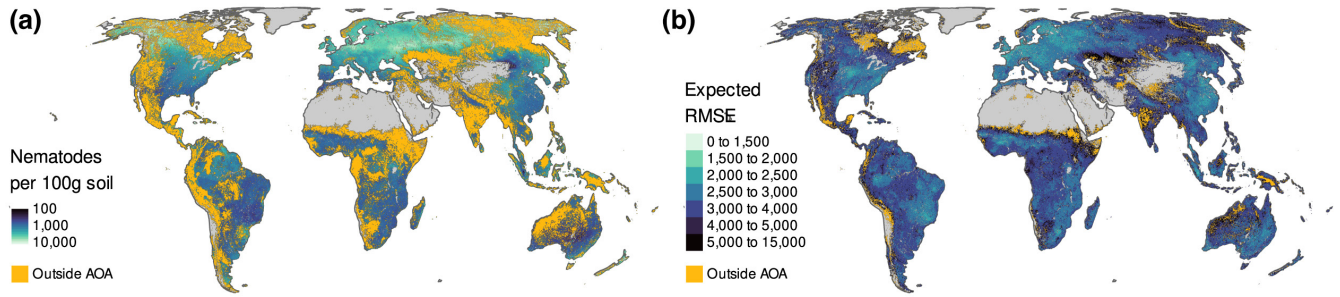

FIGURE 5 (a) Global soil nematode prediction with the area for which the model's random cross-validation accuracy [R^2 = .46, Root Mean Square Error (RMSE) = 2,938 (n/100 g)] may hold. Predictions are based on the trained model using 73 predictors; (b) the expected mapping error is based on the relationship between dissimilarity index and RMSE from multiple cross-validation strategies (see Figure S1). Locations that are considered outside the area of applicability (AOA) are more dissimilar than what could be observed based on the reference data, even when using the strictest cross-validation strategy (threefold feature-space). Grey areas (no data) indicate missing predictor data that were excluded from all analyses.

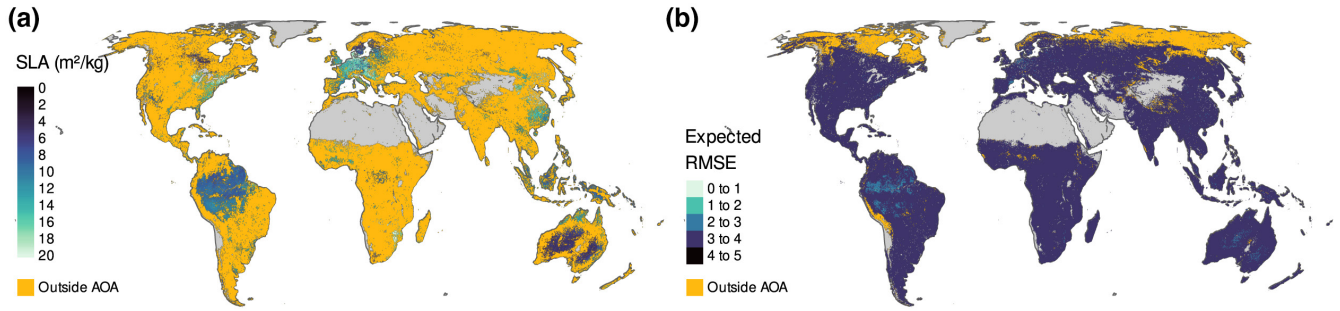

FIGURE 6 (a) Global specific leaf area (SLA) prediction with the area for which the model's random cross-validation accuracy [R^2 = .62, Root Mean Square Error (RMSE) = 2.91 (m^2/kg)] of our trained model with 15 predictors may hold; (b) the expected mapping error based on the relationship between RMSE and dissimilarity index from multiple strategies for our modelling approach (Figure S2). Locations that are considered outside the area of applicability (AOA) are more dissimilar than what could be observed based on the reference data, even when using the strictest cross-validation strategy (threefold feature-space). Vegetation-free areas were masked out (coloured in grey) in accordance with Moreno-Martinez et al. (2018).

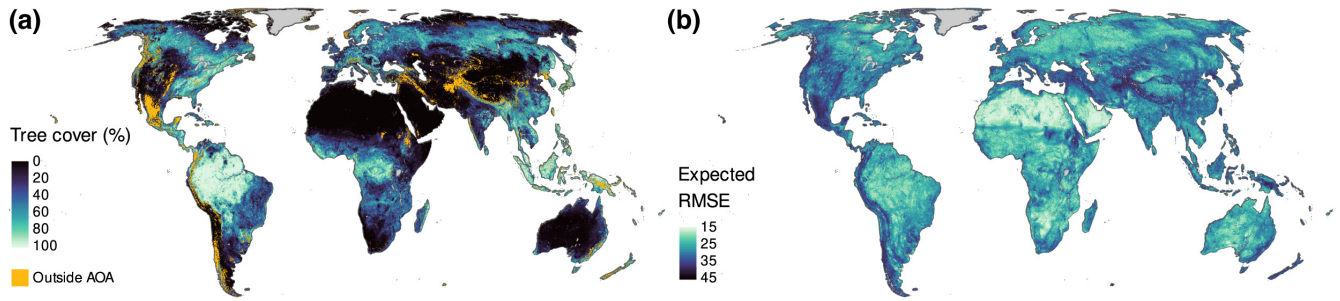

FIGURE 7 (a) Global tree cover prediction with the area for which the random cross-validation accuracy (R^2 = .67, Root Mean Square Error (RMSE) = 25.7%) of our modelling approach with 10 predictors may hold; (b) the expected mapping error based on the relationship between RMSE and dissimilarity index from multiple strategies (Figure S3). Note that no pixels are outside the area of applicability (AOA) anymore because the additional use of spatial and feature-space cross-validation allowed us to test the ability of the model to make predictions for the entire range of dissimilarities that occur globally.

TABLE 2 Comparison of model performances (Root Mean Square Error, RMSE) estimated using random, spatial and feature-space cross-validation (CV) strategies as well as the percentage of the global prediction domain that is inside the area of applicability (AOA).

Study	CV strategy	Full model RMSE | AOA	Simplified RMSE | AOA
Nematodes	Random	2,938 | 51%	2,881 | 45%
Nematodes	Spatial	3,599 | 92%	3,430 | 83%
Nematodes	Feature	3,414 | 91%	3,830 | 99%
Specific leaf area	Random	2.91 | 13%	2.86 | 11%
Specific leaf area	Spatial	4.16 | 68%	4.13 | 69%
Specific leaf area	Feature	4.05 | 82%	3.97 | 82%
Tree cover	Random	25.7 | 92%	NA
Tree cover	Spatial	26.5 | 95%	NA
Tree cover	Feature	38.9 | 100%	NA

Note: Our 'full models' utilized all predictors of the original studies (nematodes: 73; specific leaf area: 15; tree cover: 10) while simplified models utilized predictors selected by spatial variable selection only (nematodes: 8; specific leaf area: 11). Spatial variable selection did not reduce the number of predictors for the tree cover model; hence, no simplified model is shown.

expected accuracies based on the relationship between dissimilarity to the training data and prediction performance. The patterns of the expected error are similar for all three studies. In regions with extensive sampling, the estimated accuracy is higher and comparable to the random cross-validation performance. The inclusion of the spatial and feature-space strategies in the calibration led to an increased area of applicability of the models (Table 2) since we also evaluated the model performance on more dissimilar folds. Hence, more dissimilar locations can be considered inside the area of applicability. For the soil nematode and specific leaf area models, regions with lower RMSE correspond well to the regions that are considered inside the area for which the estimated random cross-validation performance holds (Figures 5 and 6). Outside the area where the random cross-validation performance holds, the expected RMSE increased and is more comparable to the spatial or feature-space cross-validation performances (Figures 5 and 6b compared to Table 2). Further, in both cases there are still regions that are too dissimilar from the training samples to be considered inside the area of applicability – even when specifically tested for extrapolation situations with the feature-space cross-validation. Especially the map of the expected RMSE of the specific leaf area (Figure 6b) shows that the available training samples are not suitable to validly predict in boreal regions. It is worth noting that we have not considered alternative surrogate splits in the random forest as the original authors did to minimize extrapolation problems. These methodological differences could yield significant deviations in the area of applicability, statistics, and conclusion reported in this work, making it hard to draw similar conclusions for both approaches. The expected RMSE of the potential tree cover model is the lowest in regions with high sampling densities (e.g., South America and the Sahara Desert). Here, the expected RMSE based on the dissimilarity index is even lower than the estimation with random cross-validation (Figure 7b compared to Table 2). In regions where the model needs to extrapolate (e.g., high altitudes in Central Asia) the expected RMSE is the highest and closely resembles the performance estimation based on the feature-space cross-validation.

3.2 | Improving the transferability of global mapping models

The spatial variable selection reduced the number of predictor variables in the models of soil nematodes and specific leaf area. A list of selected predictors is provided in Tables S1–S3. In both cases, the model simplification led to similar random cross-validation RMSE values when compared to the original models that utilized all available predictors (Table 2). This indicates that the simplified models were still capable of learning the patterns in the training data. The decrease in predictor variables was especially large for the soil nematode model where the original 73 predictors were reduced to 8. This model simplification led to a global applicability (Figure 8) and generally to lower expected prediction errors. The improvement of the model performance was especially large for regions that were

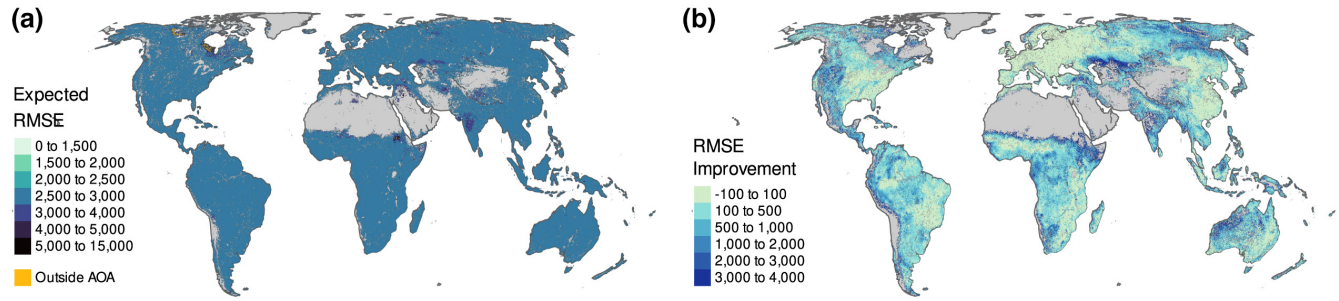

FIGURE 8 (a) Expected mapping error and area of applicability (AOA) of the global soil nematode prediction after spatial variable selection (eight predictors). (b) Effect of the spatial variable selection on the expected mapping error (Root Mean Square Error, RMSE) in comparison to the model that utilizes all 73 predictors. Grey areas (no data) indicate missing predictor data that were excluded from all analyses.

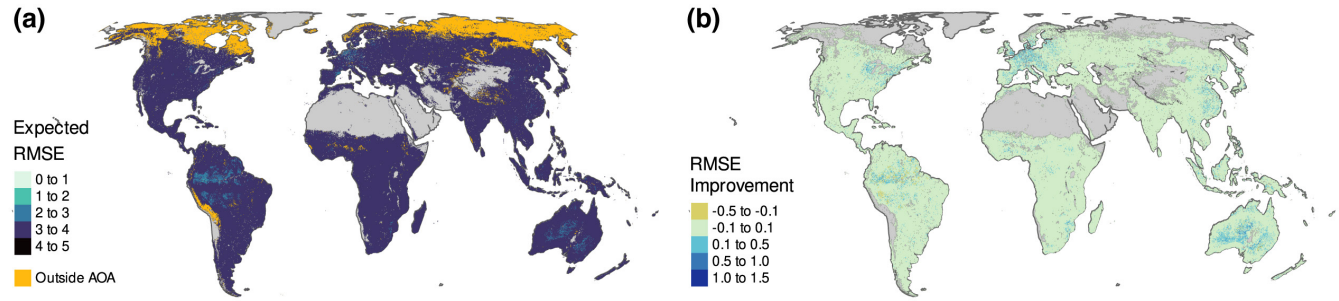

FIGURE 9 (a) Expected mapping error and area of applicability (AOA) of the specific leaf area prediction after spatial variable selection (11 predictors). (b) Effect of the spatial variable selection on the expected mapping error (Root Mean Square Error, RMSE) in comparison to the model that utilizes all 15 predictors. Vegetation-free areas were masked out (coloured in grey) Moreno-Martinez et al. (2018).

considered not applicable with a random cross-validation alone (Figure 8b). This indicates an improved ability of the model to extrapolate to unseen locations.

For specific leaf area, a set of 11 predictors led to the best results (Tables S1–S3). Since the original study also carried out variable selection based on ranked variable importance, we found little to no improvements in terms of cross-validation performance or increased area of applicability as a result of spatial variable selection (Table 2). Predictions in boreal regions are still not considered to be valid with our approach (Figure 9).

For the potential tree cover, spatial variable selection did not reduce the number of predictors, hence no further results are shown here.

4 | DISCUSSION

The recent increase in published global maps has also raised discussions on the limits of large-scale prediction models and their assessment (Meyer & Pebesma, 2022). In this study we suggested a workflow that relies on very recent developments in this field and, for the first time, combines new strategies of cross-validation, spatial variable selection and an assessment of the area of applicability with the aim of improved and more differentiated accuracy assessment of machine learning-based (global) maps of the environment. We tested the workflow with data from recently published and representative large-scale ecological prediction models to discuss the opportunities and limitations of the suggested approach.

Producing global prediction models is nowadays relatively simple following workflows such as that presented in van den Hoogen et al. (2021). While technically, predictions can be made globally as long as global predictor variables are available, we have shown that models intended for global mapping may not necessarily allow for meaningful predictions with full global coverage. This can be attributed to the spatial distribution of the reference data, which rarely cover information from all environments (reflected by the predictors). The inability of machine learning models to make predictions for predictor values different from those encountered during model training calls for applying strategies to prevent extrapolation and masking predictions outside the area of applicability of the model, especially because an assessment is impossible for such areas as outlined in Meyer and Pebesma (2022).

This is especially relevant when reference data are sparse and clustered in geographic space as seen for the nematode and specific leaf area examples in this study. The models heavily overfit the sampled environments, which results in low model performances in regions beyond the reference samples and a narrow area for which the random cross-validation performance applies (Figures 5 and 6). Well-distributed, large sets of reference data like the tree cover example diminish this problem (Figure 7). However, such samples are rarely available for many ecological target variables (see distribution

of data, e.g., in the databases like TRY, WoSIS or GBIF, and see Beck et al., 2014), unless they can be derived from high resolution remote sensing imagery such as in Bastin et al. (2019). Hence, we can expect that also beyond the case studies shown here, many global maps suffer from a limited area of applicability. The purpose of global maps, however, is to spotlight particularly urgent phenomena to an international audience (Wyborn & Evans, 2021). If the area of applicability of a model is ignored, misinterpretations and/or propagation of large errors might be the consequence. We therefore think that limiting predictions to the area of applicability or similar approaches, for example, masking areas without samples (Sabatini et al., 2022), is required.

The trained models of soil nematodes and specific leaf area clearly highlight that performances assessed by random cross-validation are not a suitable proxy for map accuracy if reference samples are sparse and clustered. As an alternative, we computed the expected accuracy as a function of the dissimilarity of new locations as suggested in Meyer and Pebesma (2021). We assessed the model accuracy with different dissimilarities between the folds with the combination of random, spatial, and feature-space cross-validation strategies. Hence, a large variety of possible prediction scenarios were included in the model validation. The resulting models are therefore evaluated and applicable for large parts or even all of the global domain and can be accompanied by a differentiated accuracy assessment (Figure 7b).

It should, however, be clarified that areas inside the area of applicability do not guarantee a meaningful model applicability (also discussed in the responses to Bastin et al., 2019, e.g., Skidmore et al., 2019; Veldman et al., 2019). Estimating the area of applicability allows predictions to be limited to the area that is comparable to the training data considering the predictors used in the model but it assumes that modelled relationships are not influenced by further unconsidered drivers. The same applies to the presented pixel-wise estimated performance: this is based on the dissimilarity index only, but other factors that this approach could not account for could also influence local performance. Still, the presented approach overcomes the limitations of one single accuracy value that highly depends on the chosen cross-validation strategy.

Limiting predictions to the area of applicability allows for more reliable machine learning-based prediction products. The dilemma now lies in the fact that neither inaccurate nor incomplete maps are particularly useful for purposes of nature conservation, planning, risk assessment or subsequent modelling. This motivated us to test whether the expected mapping error can be decreased and the area of applicability can be increased by training more simplified models. This was achieved here by spatial variable selection that reduces the predictor variables to those suitable for predicting values at new spatial locations. A similar strategy was used in a reprocessing of the map in Moreno-Martinez et al. (2018, random forest with surrogates) in which different predictors are selected for pixels that were originally extrapolated. Our results indicate that an optimal spatial variable selection strategy could increase the model's performance and enable predictions beyond training locations. In the case of the nematode map, we also achieved a nearly global applicability; hence, we expect that the performance holds on average for the entire global prediction domain (Figure 8). It could now be suspected that, in turn, the simplification of the models comes at the expense of prediction performance for areas densely covered by reference data. However, this could not be confirmed since the random cross-validation performance, which reflects the prediction quality within clustered reference data, also increased. In the maps of the expected RMSE based on the dissimilarity (Figures 8 and 9), we showed that regions close to reference samples can still be predicted with higher accuracy. For the two case studies that were based on heavily clustered reference data, the original overly complex models were hence not required and fewer predictors were sufficient to predict densely covered areas.

In the case of specific leaf area with significantly clustered reference samples, the estimation of the area of applicability revealed that the simplified model was still not applicable to the whole global domain. Regardless of the cross-validation strategy, it was not possible to test the ability of the model to make predictions for the degree of dissimilarity that is required for global predictions. This leads us to the conclusion that if reference samples are too sparse and only cover few environments, achieving reliable (and assessed) global predictions is hard to accomplish with a single model. If no global predictions can be reliably made without significant loss of prediction performance for the densely sampled areas, it might be advantageous to move away from the ambition of a global map: regionally trained models allow one to include a larger set of (higher resolution) predictors, which may lead to a model that provides more accurate predictions for the region it was trained for. A simplified global model, in contrast, often reflects general global patterns but might fail in reflecting local and regional variations (and the accuracy and uncertainty therein). Accurate regional information, however, is required for risk assessment (Maxwell et al., 2021) and nature conservation (Schmidt-Traub, 2021; Wyborn & Evans, 2021) where most decisions are made on a regional or national scale. Nonetheless, globally consistent maps of the environment are of high relevance, for example, in global conservation strategies (Jung et al., 2021; Sayre et al., 2020). To overcome the current trade-off between detailed regional and more general global maps, it might be a consideration for future modelling studies to develop approaches where different sets of predictors can be used (e.g., random forest with surrogates as in the updated version of the specific leaf area map), depending on their availability, applicability and performance in different regions.

5 | CONCLUSION

Producing global prediction models is nowadays comparably easy but we have shown that the applicability of the models to the global domain is often limited. In order to be useful for any further application, we think that it should be in the obligation of the producer of global maps to limit predictions to the area of applicability, alongside the careful assessment and interpretation of prediction

performances. We showed that model simplification by spatial variable selection can increase the prediction performance for studies that are initially based on a large set of predictor variables. But we have also shown that even when models are simplified by this approach, using sparse and clustered reference data with the ambition of producing globally continuous maps remains challenging. We hope that this study contributes to the ongoing discussion on the training and assessment of machine learning models that are intended for global mapping of the environment.

ACKNOWLEDGMENT

We want to acknowledge all data providers from the original studies and the TRY database. Please refer to the original publications for more information. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

The work was funded by the Federal Ministry for Economic Affairs and Climate Action of Germany (project number 50EE2009). This research was further supported by the European Research Council under the ERC-SyG-2019 USMILE project (grant agreement 855187).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Marvin Ludwig https://orcid.org/0000-0002-3010-018X

REFERENCES

Amani, M., Ghorbanian, A., Ahmadi, S. A., Kakooei, M., Moghimi, A., Mirmazloumi, S. M., Moghaddam, S. H. A., Mahdavi, S., Ghahremanloo, M., Parsian, S., Wu, Q., & Brisco, B. (2020). Google earth engine cloud computing platform for remote sensing big data applications: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 5326–5350.

Barbiero, P., Squillero, G., & Tonda, A. (2020). Modeling generalization in machine learning: A methodological and computational study. *arXiv:2006.15680 [cs, stat]*.

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.

Batjes, N. H., Ribeiro, E., & van Oostrum, A. (2020). Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, *12*(1), 299–320.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15.

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.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., & Smets, B. (2020). Copernicus global land cover layers—Collection 2. *Remote Sensing*, *12*(6), 1044.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315.

Hassine, K., Erbad, A., & Hamila, R. (2019). Important complexity reduction of random Forest in multi-classification problem. In *2019 15th international wireless communications & mobile computing conference (IWCMC)* (pp. 226–231). IEEE.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, 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*(2), e0169748.

Holloway, J., & Mengersen, K. (2018). Statistical machine learning methods and remote sensing for sustainable development goals: A review. *Remote Sensing*, *10*(9), 1365.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Jung, M., Arnell, A., de Lamo, X., García-Rangel, S., Lewis, M., Mark, J., Merow, C., Miles, L., Ondo, I., Pironon, S., Ravilious, C., Rivers, M., Schepaschenko, D., Tallowin, O., van Soesbergen, A., Govaerts, R., Boyle, B. L., Enquist, B. J., Feng, X., … Visconti, P. (2021). Areas of global importance for conserving terrestrial biodiversity, carbon and water. *Nature Ecology & Evolution*, *5*(11), 1499–1509.

Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Tautenhahn, S., Werner, G. D. A., Aakala, T., Abedi, M., Acosta, A. T. R., Adamidis, G. C., Adamson, K., Aiba, M., Albert, C. H., Alcántara, J. M., Alcázar C, C., Aleixo, I., Ali, H., … Wirth, C. (2020). TRY plant trait database – enhanced coverage and open access. *Global Change Biology*, *26*, 119–188. https://doi.org/10.1111/gcb.14904

Kuhn, M. (2008). Building predictive models in *R* using the caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., & Wegner, J. D. (2022). Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. *Remote Sensing of Environment*, *268*, 112760.

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, *23*(7), 811–820.

Ma, H., Mo, L., Crowther, T. W., Maynard, D. S., van den Hoogen, J., Stocker, B. D., Terrer, C., & Zohner, C. M. (2021). The global distribution and environmental drivers of aboveground versus belowground plant biomass. *Nature Ecology & Evolution*, *5*, 1110–1122.

Martin, L. J., Blossey, B., & Ellis, E. (2012). Mapping where ecologists work: Biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment*, *10*(4), 195–201.

Maxwell, A. E., Sharma, M., Kite, J. S., Donaldson, K. A., Maynard, S. M., & Malay, C. M. (2021). Assessing the generalization of machine learning-based slope failure prediction to new geographic extents. *ISPRS International Journal of Geo-Information*, *10*(5), 293.

Merow, C., Smith, M. J., Edwards, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., & Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, *37*(12), 1267–1281.

Meyer, H., & Ludwig, M. (2022). *CAST: 'caret' applications for spatial-temporal models*.

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, *12*(9), 1620–1633.

Meyer, H., & Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, *13*(1), 2208.

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.

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*, 108815.

Milà, C., Mateu, J., Pebesma, E., & Meyer, H. (2022). Nearest neighbour distance matching leave-one-out cross-validation for map validation. *Methods in Ecology and Evolution*, *13*(6), 1304–1316.

Moreno-Martinez, A., Camps-Valls, G., Kattge, J., Robinson, N., Reichstein, M., van Bodegom, P., Kramer, K., Cornelissen, J. H. C., Reich, P., Bahn, M., Niinemets, U., Peñuelas, J., Craine, J., Cerabolini, B. E. L., Minden, V., Laughlin, D. C., Sack, L., Allred, B., Baraloto, C., … Running, S. W. (2018). A methodology to derive global maps of leaf traits using remote sensing and climate data. *Remote Sensing of Environment*, *218*, 69–88.

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.

Pya, N. (2021). *Scam: Shape constrained additive models*.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

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.

Rodríguez-Veiga, P., Wheeler, J., Louis, V., Tansey, K., & Balzter, H. (2017). Quantifying Forest biomass carbon stocks from space. *Current Forestry Reports*, *3*(1), 1–18.

Roe, K. D., Jawa, V., Zhang, X., Chute, C. G., Epstein, J. A., Matelsky, J., Shpitser, I., & Taylor, C. O. (2020). Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance. *PLoS One*, *15*(4), e0231300.

Sabatini, F. M., Jiménez-Alfaro, B., Jandt, U., Chytrý, M., Field, R., Kessler, M., Lenoir, J., Schrodt, F., Wiser, S. K., Arfin Khan, M. A. S., Attorre, F., Cayuela, L., De Sanctis, M., Dengler, J., Haider, S., Hatim, M. Z., Indreica, A., Jansen, F., Pauchard, A., … Bruelheide, H. (2022). Global patterns of vascular plant alpha diversity. *Nature Communications*, *13*(1), 4683.

Sayre, R., Karagulle, D., Frye, C., Boucher, T., Wolff, N. H., Breyer, S., Wright, D., Martin, M., Butler, K., Van Graafeiland, K., Touval, J., Sotomayor, L., McGowan, J., Game, E. T., & Possingham, H. (2020). An assessment of the representation of ecosystems in global protected areas using new maps of world climate regions and world ecosystems. *Global Ecology and Conservation*, *21*, e00860.

Schmidt-Traub, G. (2021). National climate and biodiversity strategies are hamstrung by a lack of maps. *Nature Ecology & Evolution*, *5*(10), 1325–1327.

Schramm, M., Pebesma, E., Milenković, M., Foresta, L., Dries, J., Jacob, A., Wagner, W., Mohr, M., Neteler, M., Kadunc, M., Miksa, T., Kempeneers, P., Verbesselt, J., Gößwein, B., Navacchi, C., Lippens, S., & Reiche, J. (2021). The openEO API–Harmonising the use of earth observation cloud services using virtual data cube functionalities. *Remote Sensing*, *13*(6), 1125.

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.

Scowen, M., Athanasiadis, I. N., Bullock, J. M., Eigenbrod, F., & Willcock, S. (2021). The current and future uses of machine learning in ecosystem service research. *Science of the Total Environment*, *799*, 149263.

Skidmore, A. K., Wang, T., de Bie, K., & Pilesjö, P. (2019). 'Comment on "the global tree restoration potential". *Science*, *366*(6469), eaaz0111.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). BLOCKCV: An R package for generating spatially or environmentally separated folds for *k* -fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, *10*(2), 225–232.

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.

van den Hoogen, J., Robmann, N., Routh, D., Lauber, T., van Tiel, N., Danylo, O., & Crowther, T. W. (2021). *A geospatial mapping pipeline for ecologists*, *Preprint*, *Ecology*.

Veldman, J. W., Aleman, J. C., Alvarado, S. T., Anderson, T. M., Archibald, S., Bond, W. J., Boutton, T. W., Buchmann, N., Buisson, E., Canadell, J. G., Dechoum, M. D. S., Diaz-Toribio, M. H., Durigan, G., Ewel, J. J., Fernandes, G. W., Fidelis, A., Fleischman, F., Good, S. P., Griffith, D. M., … Zaloumis, N. P. (2019). Comment on "The global tree restoration potential". *Science*, *366*(6463), eaay7976.

Venter, Z. S., & Sydenham, M. A. K. (2021). Continental-scale land cover mapping at 10 m resolution over Europe (ELC10). *Remote Sensing*, *13*(12), 2301.

Wadoux, A. M.-C., Heuvelink, G. B., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692.

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation: *Model transferability*. *Methods in Ecology and Evolution*, *3*(2), 260–267.

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in *C++* and *R*. *Journal of Statistical Software*, *77*(1), 1–17.

Wyborn, C., & Evans, M. C. (2021). Conservation needs to break free from global priority mapping. *Nature Ecology & Evolution*, *5*(10), 1322–1324.

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. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ludwig, M., Moreno-Martinez, A., Hölzel, N., Pebesma, E., & Meyer, H. (2023). Assessing and improving the transferability of current global spatial prediction models. *Global Ecology and Biogeography*, *32*, 356–368. https://doi.org/10.1111/geb.13635