Biodiversity and Conservation
https://doi.org/10.1007/s10531-023-02608-9

ORIGINAL RESEARCH

Assessment of multiple model algorithms to predict earthworm geographic distribution range and biodiversity in Germany: implications for soil-monitoring and species-conservation needs

Gabriel Salako[1,2] · **David J. Russell**[1] · **Andres Stucke**[3,4] · **Einar Eberhardt**[3]

Received: 29 November 2022 / Revised: 28 March 2023 / Accepted: 9 April 2023

Abstract
Identifying the potential distribution of soil-biodiversity with its density and richness relationships, including constituent species, is a pre-requisite for the assessment, conservation and protection of soil biodiversity and the soil functions it drives. Although the role of earthworms in improving soil quality has long been established, to quantitatively and spatially assess how this soil-animal group's distribution changes along environmental gradients and geographic space and the identification of the drivers of such change has not been fully investigated. This comprehensive study aimed at modelling and mapping earthworm spatial distribution and diversity patterns to determine their conservation needs and provide baseline reference data for Germany. The study compared multiple modelling algorithms to map earthworm community parameters and 12 species-specific distribution probabilities, calculate their geographic range sizes and determine responses to environmental predictor variables. Three general patterns of spatial distribution ranges were identified by the model predictions (large-range, mid-range, and restricted-range species) with the corresponding environmental contributions to the predictions. Modelled species responses to environmental predictors confirm observed environmental drivers of earthworm distribution in Germany. The range classes based both on distributional level and geographic space provide the necessary information for identifying conservation and decision-making priorities, especially for restricted-distribution species as well as those with clearly defined habitat preferences.

Keywords Density · Species richness · Spatial distribution · Modelling · Biodiversity

Communicated by David Hawksworth.

✉ Gabriel Salako
gabriel.salako@senckenberg.de

[1] Senckenberg Museum of Natural History, Görlitz, Germany

[2] Kwara State University, Malete, Nigeria

[3] Federal Institute of Geoscience and Natural Resources, Hannover, Germany

[4] Institute for Geography, University of Hildesheim, Hildesheim, Germany

Introduction

A major portion of the biodiversity of terrestrial ecosystems is represented by soil-dwelling fauna (FAO et al.. 2020). Furthermore, most ecosystem services provided by terrestrial habitats are based on soil functions (Adhikari and Hartemink 2015), whereby practically all soil functions derive from processes driven by soil organisms such as earthworms (Gardi and Jeffery 2009; Turbé et al. 2010). The enormity of soil biodiversity and its functional importance has led to the appreciation of the conservation needs of soil biocoenoses (Brussaard 1998; Lavelle et al. 2006; Bouma and Montanarella 2016; Briones 2018), which has been recognized by stakeholders and politics (e.g.,Turbé et al. 2010; FAO et al. 2020; To this aim, the recently adopted European soil-protection strategy framework specifically calls for protecting and conserving soil biodiversity through explicit monitoring programs. Many publications have outlined approaches for monitoring soil biodiversity (e.g., Gardi and Jeffery 2009; Cluzeau et al. 2012; Pulleman et al. 2012; Griffiths et al. 2016; Orgiazzi et al. 2016; van Leeuwen et al. 2017). Some concrete national soil-biodiversity monitoring programs specifically detail assessment approaches (e.g., Weeks 1998; Römbke et al. 2000; Rutgers et al. 2009; Cluzeau et al. 2012), whereby these base an assessment on a comparison with "reference values" derived from field surveys in selected reference sites. However, conservation and protection of soil biotic communities require evidence-based baseline information—derived at broader spatial scales—on their local and regional distributions, which is necessary for formulating reference values ("standard operational ranges") for soil-biodiversity monitoring and assessment (Huber et al. 2008; Cluzeau et al. 2012; Ockleford et al. 2017; Baritz et al. 2021). An approach for deriving generalizable soil-biodiversity baselines is to upscale local observational data to broader spatial scales using correlative modelling methodologies.

Species biodiversity is a function of community composition (*which* species occur), species richness (*how* many species co-occur), both total communities as well as individual species' densities, and the pattern of their occurrences in geographic space, all of which are vital for biodiversity assessments. Using only one or few of these metrics is not sufficient for assessing biodiversity as, e.g., two communities may be identical in richness but differ in densities and occurring species identities (Groves 2022). Identifying the potential distribution of soil-biodiversity coupled with its community composition, density and richness relationships is a pre-requisite for the assessment, conservation and protection of soil biodiversity and the soil functions it drives.

In recent years, a powerful tool for understanding biodiversity, its distribution and the potential drivers of this distribution has been the development of species distribution models (SDMs). SDMs statistically model species or community's correlations with environmental parameters and use these correlations to upscale their potential occurrences to larger spatial scales based on the spatial distribution of the environmental parameters (Guisan et al. 2017). Such mapping methodologies are a core solution for decision support of biodiversity and ecosystem-services conservation policies throughout the EU (Maes et al. 2012).

Various SDM modelling methods or algorithms exist today, such as Generalised Linear Models (GLM), Classification Tree Analysis (CTA), Multivariate Adaptive Regression Spline (MARS), Maximum Entropy (MaxEnt), Random Forest (RF) and Generalised Boosted Regression Models (GBM), among others. Each of these possess inherent strengths and weaknesses (Li and Wang 2013; Valavi et al. 2021), and a few of them are sensitive to sampling size, which greatly affect their capacity to predict accurately

(Kumar and Stohlgren 2009). Models' performance comparison is one of two methods for overcoming model-based uncertainty in SDM, the other is the use of ensemble methods (Marmion et al. 2008). Guisan et al. (2017) identified various evaluation metrics that can be used to select the best performing model for SDM projections and predictions, which range from conventional statistics such as R^2, root mean square error (RMSE), to threshold-independent "area under the receiver operating characteristic curve" (AUC_{ROC}) and Kappa statistics, among others. The choice of the metrics to use for evaluation should however be determined by the nature, type of data and the objectives of the studies (Guisan et al. 2017; Zurell et al. 2020).

Related to species' distribution ranges (and possibly used interchangeably, but different in methodology and purpose) is a "species' range size". While "species distribution range" describes the occurrence of a taxon or its arrangement within a geographic space without necessarily determining the range size or extent, "species geographic range size" is the geographic area in which a taxon is found, mostly measured in km^2. This latter metric has been used as an indicator for assessing the threat status of a species; a narrow-range species being much more vulnerable and having a higher probability of going extinct than a species with a wider range size (Gaston and Fuller 2009). The classification of species into threat categories has mostly been done using geographic range size (Sheth et al. 2020), which has been adopted as standard practice by the IUCN under criteria B1 and B2 both at global and regional scale (2012b, 2022). While this study did not intend to re-produce the red list for earthworms in Germany, the predicted species distribution maps and geographic range sizes can be used to classify species into distributions range classes for information on conservation guidance.

Although the role of earthworms in, e.g., improving soil structure and quality has long been established (Blanchart et al. 1999; Blouin et al. 2013), quantitatively assessing how this soil-animal group is distributed along environmental gradients and geographic space and the identification of the drivers of such change has only recently been investigated. Previous studies have focused either on specific taxa (i.e., Marchán et al. 2015: the species *Hormogaster elisae* Alvarez, 1977, Marchán et al. 2021: the endemic genera *Kritodrilus* Bouché, 1972 and related taxa, Marchán et al. 2016: the family Hormogastridae Michaelsen, 1900) or on earthworms in total (i.e., Palm et al. 2013; Rutgers et al. 2016; Phillips et al. 2019). They also investigated different spatial scales: local (e.g., Gabriac et al. 2022, Marchán et al. 2015), catchment (Palm et al. 2013), regional (Marchán et al. 2016, 2021; Marchán Marchán and Domínguez 2022) or continental/global (Rutgers et al. 2016; Phillips et al. 2019). Furthermore, such studies often employed a single modelling framework for predicting spatial distributions (i.e., GLM: Rutgers et al. 2016, GLMM: Phillips et al. 2019, MaxEnt: Marchán et al. 2015, 2016, BRT: Palm et al. 2013; but see Marchán et al. 2016; Marchán & Domínguez 2022 for ensemble methods). The current study's general goal was therefore to compare different modelling algorithms' abilities to spatially model and map the distribution of both earthworm communities and selected species at a national scale (Germany) for assessment of their distributional and conservation status. It therefore sought to achieve the following objectives: (1) apply and compare correlative modelling techniques for mapping the spatial distribution and the geographic range of earthworm community parameters as well as selected earthworm species in Germany, (2) identify and determine the importance of environmental predictors based on their contribution to correlative models, and (3) evaluate whether species related to different earthworm life-form types are predicted to generally react differently to environmental drivers.

Materials and methods

Study area and workflow

The spatial scope of the study encompasses the entire range of mainland Germany. In the models and for mapping algorithms, the latitudinal extent was set from 47.3209 to 54.9049° N and the longitudinal extent from 6.0470 to 14.8428° E (Fig. 1). This project's modelling workflow followed the Overview, Data, Model, Assessment and Prediction Protocols ("ODMAP", Zurell et al. 2020; Fig. 2), consisting of eight systematic steps: data collation of (1A) biological (earthworm) and (1B) environmental data, (2) variable selection, (3) multicollinearity tests, (4) data organization, (5) data partitioning into training and test data, (6) model calibration and fitting, (7) prediction (upscaling ["mapping"] model results) and (8) model evaluation and assessment.

Earthworm's data collation

Raw data were downloaded on 22 February 2021 from Senckenberg's soil-biodiversity data warehouse ("Edaphobase", Burkhardt et al. 2014) using the filters "Lumbricidae" & "Germany". Additional density data (i.e., from Bavaria, Brandenburg, Saxony-Anhalt) were obtained and included in datasets. Data were cleaned so that only community-level data was used, and each location of occurrence ("site") truly included a unique habitat and soil type. This resulted in a dataset consisting of 22,134 individual data records (rows) from 992 locations (sites of occurrence) within Germany (see Supplement 1 for the individual data sources). All available metadata concerning soil, habitat types, climate, etc. were downloaded with the earthworm data, being specifically linked to the individual data records (i.e., earthworm data per specific site of occurrence). To address the problem of sampling bias, we performed spatial thinning on the earthworm (occurrence records) using the thin function of the R Spthin package (Boria et al. 2014; Aiello-Lammens et al. 2015). For community-level modelling, the data from each location was aggregated for the parameters "total earthworm density" (harmonized to number of individuals per square meter [= individuals/m^2]) and "earthworm species richness" (average number of species found occurring in a site). For species occurrences, individual species were listed as occurring in a site (="present" [=1]), if data for that species existed for any sampling date of the site. It was assumed that any species not listed in the data for that site likely did not—or only rarely—occurred in the site (at least at the time of the sampling event(s)) and was listed as not occurring in the site (="pseudo-absences [=0]) While we modeled density and richness using 45 valid species, we selected only 12 species (Table 1) prepared into presence (p = 1) and "pseudo-absences (a = 0)) for distribution modelling.

Environmental data collation

We pre-selected 15 external predictor variables known to have physiological and ecological importance in the distribution of earthworm species (Rutgers et al. 2016; Phillips et al. 2019; Edwards and Arancon 2022). As many Edaphobase records did not contain all of these variables, external data was used to augment data gaps. Data on climate variables were downloaded from Climatologies at High resolution for the Earth Land Surface Areas (CHELSA, Karger et al. 2017, https://chelsa-climate.org/bioclim/): mean annual temperature (Bio1), total annual precipitation (Bio12), growing degree days with temperatures

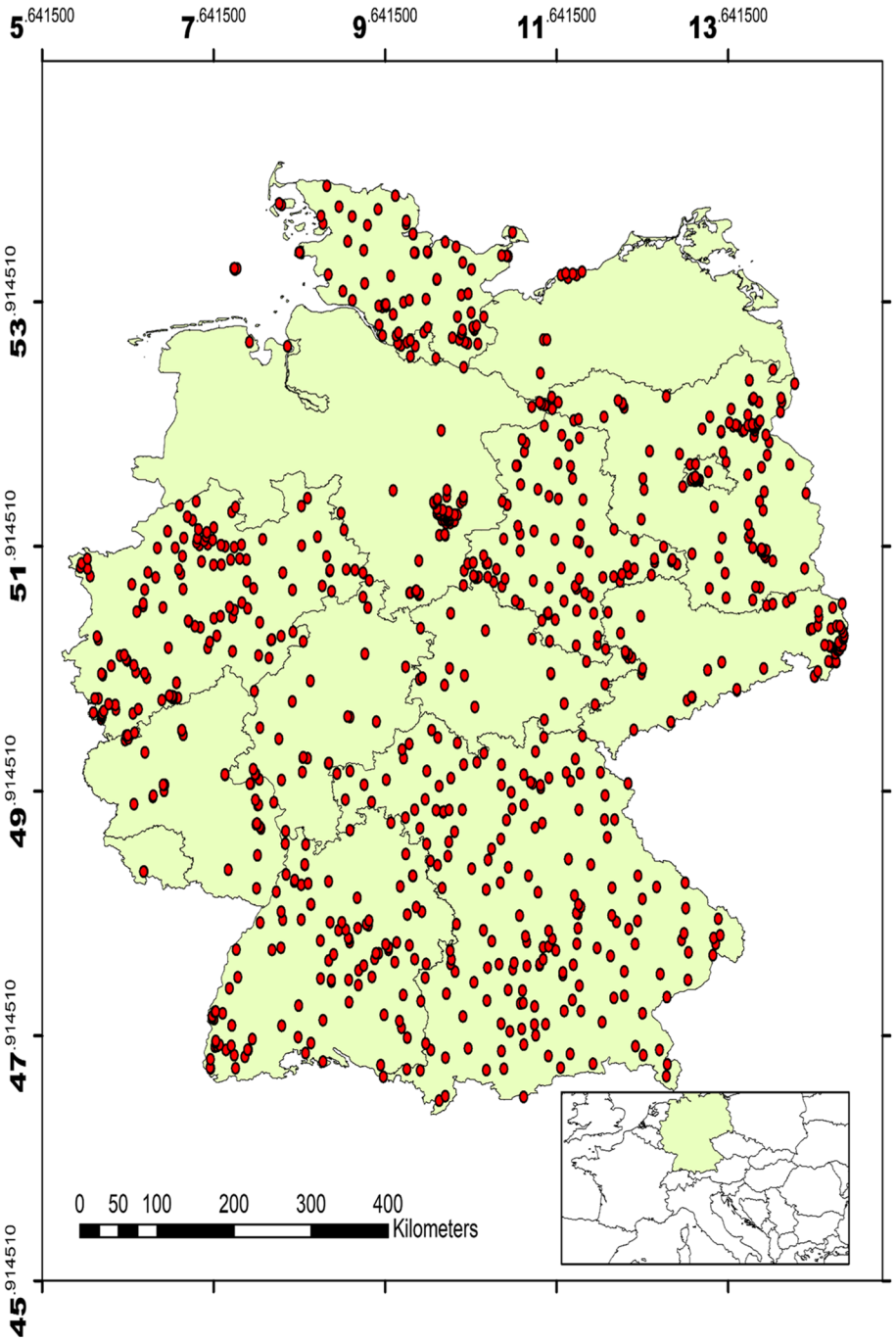

Fig. 1 Map of the study area (Germany) showing earthworm species occurrence data (dots) used for calibrating the distribution models. Occurrence-data source: Edaphobase (https://portal.edaphobase.org)

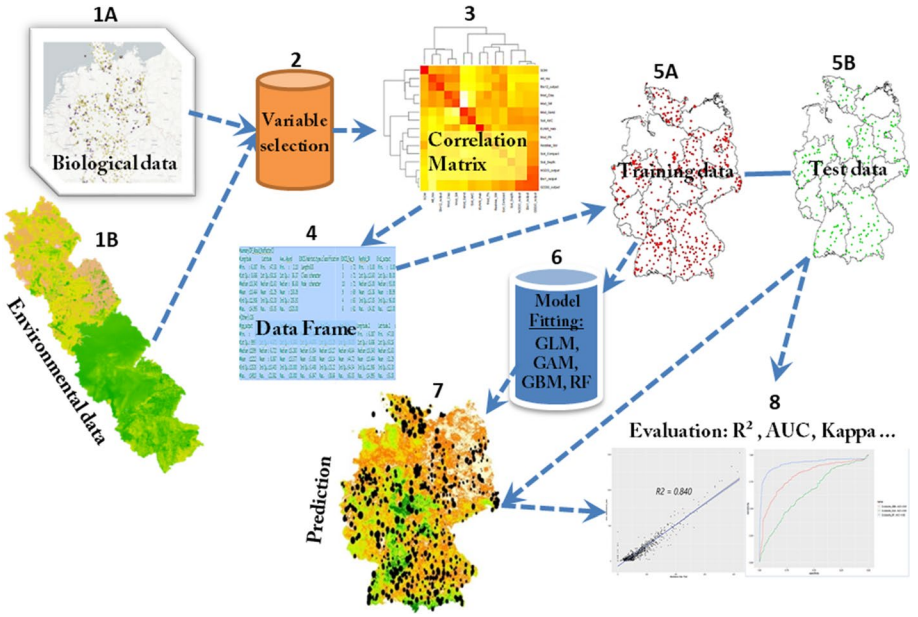

Fig. 2 Graphical model workflow according to ODMAP protocols

Table 1 Earthworm species selected for modelling species-specific occurrence probabilities

Species	Number of sites of occurrence	Ecological group	Selection criterion
Aporrectodea caliginosa (Savigny, 1826)	719	Endogeic	Common
Aporrectodea rosea (Savigny, 1826)	603	Endogeic	Common
Lumbricus terrestris Linnaeus, 1758	546	Anecic	Common
Lumbricus rubellus Hoffmeister, 1843	595	Epigeic	Common
Allolobophora chlorotica (Savigny, 1826)	304	Endogeic	Common
Lumbricus castaneus (Savigny, 1826)	319	Epigeic	Common
Aporrectodea longa (Ude, 1885)	143	Anecic	Limited range (?)
Dendrobaena octaedra (Savigny, 1826)	309	Epigeic	Unique habitat
Dendrobaena attemsi (Michaelsen, 1903)	29	Epigeic	Limited range (?)
Bimastos eiseni (Levinsen, 1884)	43	Epigeic	Limited range (?)
Aporrectodea limicola (Michaelsen, 1890)	67	Endogeic	Unique habitat
Lumbricus badensis Michaelsen, 1907	2	Anecic	Endemic

above 10 °C (NGD10), and growing degree days with temperatures above 0 °C (GDDO). This climate data represented average values between 1981 and 2010 at 30 arc-second resolution, which corresponds to the main time frame of the biological data. External data on soil variables (texture [sand, silt, and clay content], soil depth, bulk density, air capacity [=porosity] and organic–matter content) were obtained from German Federal Institute of Geoscience and Natural Resources (BGR) at a 250 m resolution. Soil moisture (average % soil water content at 2.5 cm soil depth) was obtained from the European Space Agency's

(ESA) Climate Change Initiative (CCI) Soil Moisture product (https://www.esa-soilmoisture-cci.org/data; Dorigo et al. 2017; Gruber et al. 2019). Topographic data (elevation at m.a.s.l.) was derived from USGS-NASA (https://earthexplorer.usgs.gov/). Land-use/habitat-type data was obtained from the Ecosystem Types of Europe (https://www.eea.europa.eu/data-and-maps/data/ecosystem-types-of-europe-1) based on the EUNIS (European Nature Information System) habitat classification level-1 details, version 3.1, 2019 at 100 m resolution. The habitat raster data was grouped into 11 habitat-type classes (marine influenced, coastal, inland waters, mires and bogs, grasslands, heathland, woodland and forest, sparse vegetation, arable land, urban and artificial habitats, and habitat complexes). We initially referenced all pre-selected variables to the World Geographic coordinate System (WGS84), cropped and masked them to the German shapefile and resampled or disaggregated to a common resolution of 250 m to match the resolution of the earthworm data.

SDM modelling workflow

Model predictor variable selection

All modelling procedures were performed in the R programming language version 0.1.19, 2021 (R Core Team 2021). As issues of predictor autocorrelation and multicollinearity affect model performance and accuracy (Guisan and Zimmermann 2000; Mod et al. 2016; Salako et al. 2021), during model calibration significant and independent environmental predictor variables were selected. For this, we first used principal component analysis (PCA) and Pearson correlation to identify autocorrelation among predictors. The correlation threshold was set at $R^2 \geq 0.7$ (Johnson et al. 2002; Bobrowski et al. 2021). Subsequently, a variance inflation factor (VIF; Vifstep) was used to remove any further multicollinearity from the predictor variables. These procedures, including the use of Akaike information criterion (AIC) during model building, resulted in a final set of 11 environmental predictor variables: soil depth, soil pH, clay & silt content, soil moisture, soil bulk density, soil porosity (= air capacity), average annual temperature and precipitation, and habitat type.

Model calibration and fitting

We tested four different model algorithms to model earthworm distribution predictions based on model classification into traditional regression and machine learning algorithms as well as their broad usage in modern studies (Li and Wang 2013; Valavi et al. 2021): Generalized Linear regression Models (GLMs), Generalized Additive Models (GAMs), Generalized Boosting Models (GBM) and Random Forest models (RF). These model algorithms have been described in detail in several publications (e.g., Guisian and Zimmermnan 2000; Li and Wang 2013; Guisan et al. 2017; supplementary Table 1). Model fitting for the selected algorithms (GLM, GAM, GBM, RF) were performed with the eleven environmental predictor variables and response variables being species observation presence/absence data (P/A) to predict species-specific distribution probabilities, and community total density (ind./m^2) and richness (number of co-occurring species) data to predict these community-level metrics. We ultimately used macro-ecological models (MEM) to directly predict and map earthworm species richness because of its relative advantages over stack species distribution models (SSDM) (Biber et al. 2020); since data for rare species were very patchy resulting in underprediction of this metric in SSDMs.

GLM and GAM models were formulated by calling the following functions, with added syntax ("poly") for polynomial regression in GLMs and ("s") for smoothing in GAMs.

$$Y < -glm\left(a \sim x_0 + x_1 + x_2 + x_3 + x_4 \ldots\ldots\ldots xn,\ data = DF\right)\text{(linear regression)},$$

Family $=''$ *gaussian*$''$ (*for p/a*) and $''$*poisson*$''$ (*for density and richness*)

$$Y1 < -glm\left(a \sim poly\left(x_{0,2}\right) + poly\left(x_{1,2}\right) + poly\left(x_{2,2}\right) + poly\left(x_{3,2}\right)\right.$$
$$\left.poly\left(x_{4,2}\right) \ldots\ldots\ldots poly\left(x_{n,2}\right), data = DF\right)\text{(polynomial regression)}$$

$$Y2 < -gam(a \sim s(x) + s\left(x_1\right) + s(x_2) + s(x_3) + \ldots\ldots\ldots s(xn),\ data = DF)$$

whereby a = the response variable, x = environmental/predictor variables, DF = training/test data (data frame combining both the response and predictor variables). We selected the best models in a backward stepwise regression based on their Akaike Information Criterion (AIC). The models with the lowest AIC were selected (Whittingham et al. 2006). The final models included the predictor variables listed above.

Model fitting for GBM was performed in both the DISMO (Thuiller et al. 2021) and BIOMOD 2 packages (Hijmans et al. 2020) to maximally utilize all features associated with the two packages, using the same data as for GLM. We set the *learning rate* (*lr*) at 0.001 and a *bag.fraction* of 0.5 with a maximum cross validation of 10 (*cv.folds* = 10). Final selection was based on the model with lowest predictive deviance. The following function for GBM was used:

$$Y < -gbm.step\ (data = DF,\ gbm.x =,\ gbm.y =,\ family =''\ gaussian''\ or\ ''$$
$$Poisson'',\ tree.complexity = 2,\ learning.rate = 0.001,$$
$$bag.fraction = 0.5,\ cv.folds = 10)$$

whereby DF = the training/test data (data frame combining both the response and predictor variables), *gbm.x* = the predictor columns and *gbm.y* = the response column.

The random forest (RF) model was formulated with the following function:

$$Y < -random\ Forest(x = n[, 5 : 15], y = n[, 3], n\ tree = 1000,\ node\ size = 10,\ importance = T)$$

whereby x = the n data frame columns for predictors, y = the n data frame column for the response variable and *ntree* = the number of trees.

To determine selected earthworm distributional range size, species range size estimates were implemented in R (Rangemap; Cobos et al. 2021), using the earthworm occurrence data to calculate species extent of occurrence (EOO) and area of occupancy (AOO) using minimum convex hull polygon (Gaston and Fuller 2009; IUCN 2012a, b) at a national scale (Table 4). We therefore used the combination of predicted distribution maps and geographic range sizes to classify species into spatial distributions ranges: (1) large-range distribution are species with widespread distribution and AOO > 2000 km^2 (2) mid-range are species with AOO > 1000 km^2 but less than 2000 km^2, (3) restricted or small range are species with AOO < 500 km^2 and endemic and unique habitat species are those with AOO < 200 km^2.

All model predictions (as spatial raster files) were imported to a GIS environment for visualization. To later more precisely project modelling results onto the spatial maps,

we transformed the result raster files to the Europe-focused projection system "ETRS89 LAEA" (EPSG code: 3035) for enhanced map visualization. Furthermore, as a first approximation of earthworm potential diversity, we produced a GIS "overlay" of the earthworm community total-density and species-richness modelling results.

Model assessment/evaluation

To assess model performance, we used a split-sample cross validation (CV) method by splitting the data into training and test datasets at a 70:30 ratio (Phillips et al. 2008; Hijmans and Elith 2019; Guisan et al. 2017). The train datasets were used to fit the models, while the test datasets were used to evaluate model predictive performance. For the quantitative response variables of community species richness and total density, the coefficient of determination of a regression (R^2) between the observations (field data) and predictions, as well as the concordance index (C-index) were used. We evaluated the predictive ability of the species-specific distribution-probability models using the threshold independent statistic of "Area Under the receiver operating characteristic Curve" (AUC_{ROC}) (Jiménez-Valverde 2011) and Cohen's Kappa coefficient statistic. AUC_{ROC} ranges from 0 to 1 with an AUC of 0.5 or lower described as not better than a random prediction, 0.7 to 0.8 considered acceptable, 0.8 to 0.9 considered excellent, and more than 0.9 considered outstanding (Manel et al. 2001; Salako et al. 2015; Guisan et al. 2017). Kappa scores range from − 1 to 1, with 0 (no agreement; random) and 1 (perfect agreement), and the rare occasions of negative values signifying less agreement than expected by chance. We obtained the relative contribution of environmental variables to model predictions and checked model ecological plausibility by extracting their response curve "partial plots".

To test models' reliability performances, we applied Friedman's one way analysis of variance by rank on the results of all the model performances by each evaluation metrics, implemented in the base R function *rstatix* (Alboukadel 2021*)*.

Results

Model performance and final model selection

When calibrating the GLM models, a polynomial effect was very marginal when assessed by AIC. Therefore, to maintain model comparability, we only retained the linear GLM model. We also subsequently dropped the GAMs, as there was no significant performance difference compared to GLMs (data not shown). The Friedman test resulted in $P < 0.00112$, indicating a significant difference in model performance. The evaluation of the tested models' prediction performance on earthworm community species richness and total density showed that GLM performance exhibited very low R^2 and C-Index values (Table 2), indicating poor prediction of both metrics. RF had the highest R^2 for total density and species richness; while the R^2 values for GBM were much lower (Table 2). The C-Index scores were also higher for RF than GBM for both total density and species richness.

For the selected species' occurrence probabilities, the AUC_{ROC} and Kappa scores of the three remaining algorithms showed that mean scores ranged from 0.601–0.982 to 0.096–0.982, respectively (Table 3). RF resulted in the highest mean scores with less

Table 2 Prediction performance (goodness-of-fit) metrics for the different tested model algorithms for the two community-level response variables

Model	R^2	C-index
Density		
GLM	0.089	0.603
GBM	0.155	0.657
RF	0.840	0.861
Richness		
GLM	0.029	0.583
GBM	0.115	0.658
RF	0.574	0.819

variation relative to the other models. GLM resulted in the lowest AUC values, with an average of 0.683 (Table 3). We therefore chose RF for the subsequent modelling and mapping.

Predicted spatial distribution of earthworm community total density and species richness

The predicted earthworm community total density ranged from 10 to maximally 600 ind./m^2, with an average of 350 ind./m^2 per site (Fig. 3, left). Species richness predictions ranged from 1 to 12 species, with an average of 4–5 species per site (Fig. 3, right). Higher total community densities (> 400 ind./m^2) were predicted especially in grasslands and arable land in north-eastern Germany (Fig. 3, left). However, in these regions, species richness was predicted to be relatively poor with an average of 2 species per site (Fig. 3, right). A comparison with predicted species distributions (Fig. 5) revealed that primarily epigeic species such as *D. octaedra* (but also *L. rubellus* and the endogeic *A. caliginosa*) were responsible for these predicted high total densities. The GIS overlay of densities and richness produced approximate earthworm diversity (Fig. 4).

Predicted geographic distribution and range size of selected earthworm species

The modelled distribution predictions for individual species were based on probability of occurrences, with the scale ranging from (0) no occurrence probability to (1) maximum occurrence probability. We measured the geographic range size in km^2 using extent of occurrence (EOO) and area of occupancy (AOO) of the probability predictions. Based on AOO assessment and the distribution-model results, three general patterns of spatial distribution ranges were identified (see Table 4 for geographical range sizes used as grouping criteria): (1) species with large distribution ranges (*A. caliginosa*, *A. rosea*, *L. rubellus* and *L. terrestris*), (2) species with mid-range distributions (*A.chlorotica*, *L castaneus*, *D. octaedra* and *A. longa*), (3) species with restricted or small distributional ranges (*D. attemsi*, *B. eiseni* and *A. limicola*), including endemic species (*L. badensis*) or those limited to unique habitats (Fig. 5).

Table 3 Mean AUC_{ROC} and Cohen's Kappa scores of the evaluated models for each of the selected species, as well as ranges (min., max.) and average of all species

Model	*A. eiseni*	*D. attemsi*	*D. octaedra*	*L. castaneus*	*L. rubellus*	*A. caliginosa*	*A. chlorotica*	*A. limicola*	*A. rosea*	*A. longa*	*L. badensis*	*L. terrestris*	Min.	Max.	Mean
AUC															
GLM	0.825	0.706	0.654	0.624	0.680	0.621	0.628	0.775	0.645	0.601	0.796	0.646	0.601	0.825	0.683
GBM	0.908	0.890	0.807	0.762	0.763	0.772	0.734	0.884	0.752	0.742	0.896	0.771	0.734	0.908	0.807
RF	0.938	0.941	0.919	0.925	0.920	0.922	0.924	0.995	0.994	0.880	0.982	0.936	0.880	0.995	0.940
Kappa															
GLM	0.761	0.367	0.108	0.163	0.145	0.143	0.128	0.148	0.138	0.096	0.132	0.187	0.096	0.761	0.210
GBM	0.864	0.768	0.437	0.364	0.369	0.567	0.432	0.432	0.362	0.146	0.457	0.357	0.146	0.864	0.463
RF	0.903	0.923	0.974	0.778	0.876	0.938	0.718	0.669	0.941	0.357	0.982	0.974	0.357	0.982	0.836

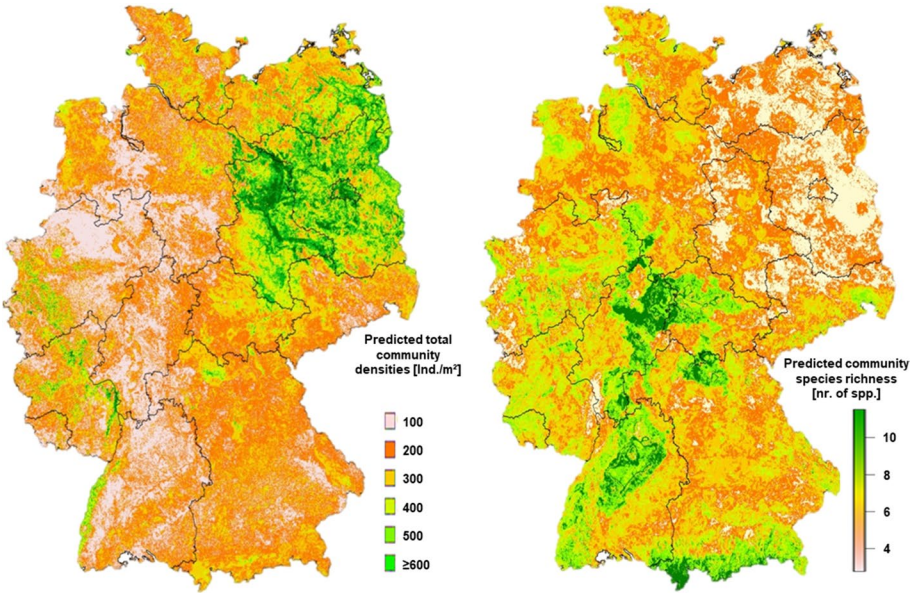

Fig. 3 Earthworm community total density (left) and species richness (right) predicted by the Random Forest model

Environmental variables' relative contribution to model predictions and community/species responses to predictor variables

Community-level analyses

Contributions of environmental predictor variables The environmental variables contributing most to predicted earthworm total density were soil moisture, habitat type, and average annual precipitation, which together contributed to almost 50% of the model results (Fig. 6). The remaining predictors for total density contributed fairly equally (5 to 10% each) to the model results, with the exceptions of average annual temperature and soil pH, which interestingly only accounted for less than 5%. The environmental contributions to predicted species richness were dominated by clay content (as a proxy for soil texture) and habitat type, which contributed 17 and 19% respectively (Fig. 6). Climate variables (average annual precipitation and temperature) further contributed more than 10% to the species-richness predictions. As opposed to the total density predictions, pH also influenced the species-richness predictions by almost 10%, while the remaining variables almost equally contributed between 4 and 8%. Soil depth played the most minor role in the species-richness predictions.

Community responses to environmental predictors Regarding climate, density predictions increased below 500 mm/a *total annual precipitation*, but decreased above this threshold (Supplementary Fig. 1). *Average annual temperature* did not affect density predictions below 10 °C, but increases were predicted above 10 °C. Interestingly, these climate parameters had somewhat opposing effect on species-richness predictions, which increased above 500 mm/a precipitation and decreased between 6 and 10 °C average annual temperature.

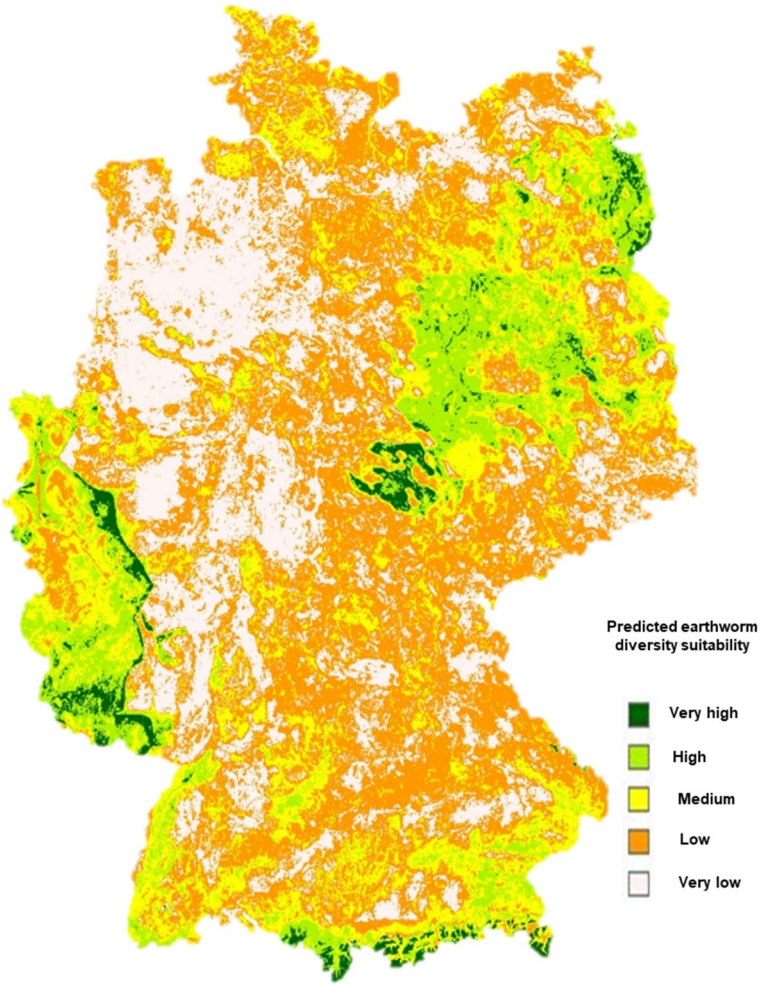

Fig. 4 Predicted habitat suitability for earthworm diversity (overlay of Random-Forest predictions of earthworm total densities and species richness)

Concerning soil parameters, increasing *soil moisture* (as of ca. 7%) led to lower total density predictions (Supplementary Fig. 1), but did not strongly influence species-richness projections until 20% water content, at which level species richness increased. *Soil acidity* exhibited a similar threshold for both total density and species-richness projections, dramatically reducing at about pH 4, only to increase again above pH 5. The predictions of species richness, on the other hand, remained very low below pH 4, and strongly increased above this value. Although *soil organic matter* contributed only slightly to the model results, density predictions increased above 6% SOM content, while showing little influence on species-richness predictions. An influence of soil texture was best represented by *clay content*, where predictions of both total density and species richness increased above 20–30% clay. *Silt content* had apparently little influence on species richness, but density predictions strongly decreased above 30%. Besides soil texture, soil structure also influenced the earthworm community predictions. Density projections strongly increased at a

Table 4 Classification of species' Geographic range size based on IUCN category B criteria ("1" and "2"; see "Methods" for explanations)

Species	EOO (km^2)*	AOO (km^2)*	++Geographic range size class
A. caliginosa	364,624.4	2420	Large range
A. rosea	365,299.4	2092	Large range
L. terrestris	365,187.7	2084	Large range
L. rubellus	375,796.7	2008	Large range
A. chlorotica	350,969.0	2004	Mid-range
L. castaneus	355,539.2	1040	Mid-range
D. octaedra	365,051.9	1020	Mid-range
A. longa	354,399.3	572	Mid-range
D. attemsi	201,500.8	108	Restricted range
B. eiseni	192,725.0	100	Restricted range
A. limicola	262,682.8	196	Restricted range
L. badensis	NA	NA	Restricted range/endemic (Lehmitz et al. 2016)

EOO extent of occurrence, *AOO* area of occupancy, *NA* occurrence data insufficient for calculating range sizes

++The large distribution range are species with widespread distribution and AOO > 2000 km^2 (2) the mid-range are species with AOO > 1000 km^2 but less than 2000 km^2, (3) the restricted or small range are species with AOO < 500 km^2 and the endemic and unique habitat species are those with AOO < 200 km^2

*Thresholds for EOO are 100 km^2, 5000 km^2 and 20,000 km^2 for Critically Endangered (CR), Endangered (EN) and Vulnerable (VU) species, while the equivalent values for AOO are 10 km^2, 500 km^2 and 2000 km^2, respectively

total porosity above 15%, but species-richness projections decreased continuously with increasing porosity. *Soil bulk density* apparently negatively affects earthworm communities, as both total-density and species-richness predictions abruptly and dramatically decreased above bulk densities of 1.5 g/cm^3.

Species-specific analyses

Environmental variables contribution Regarding species' occurrence-probability models, four variables were predicted to be the principal drivers of earthworm species' distributions in Germany: precipitation and associated soil moisture, habitat type, and soil pH. For some species, average annual temperature and soil organic matter were also apparently important (Table 5).

Total annual precipitation and the related *soil moisture* often accounted for up to 20% or more of the predictions in many species and were found to be essential predictor variables (together contributing more than 33%) for the modelled distribution of *A. limicola*, *A. chlorotica*, and *L. badensis*, but also for *L. terrestris*, *L. rubellus* and *L. castaneus* (ca. 30%) as well as somewhat (> 25%) for *A. rosea*, *A. caliginosa* and *D. attemsi* (Table 5). Therefore, these hydrological parameters were highly important predictors of the occurrence of ¾ of the tested species. The occurrences of *D. attemsi*, *B. eiseni*, *L. rubellus* and *A. chlorotica* were predicted to also be highly dependent on climate *(average annual precipitation and temperature*), exhibiting large contributions of these variables (often > 21%) to their

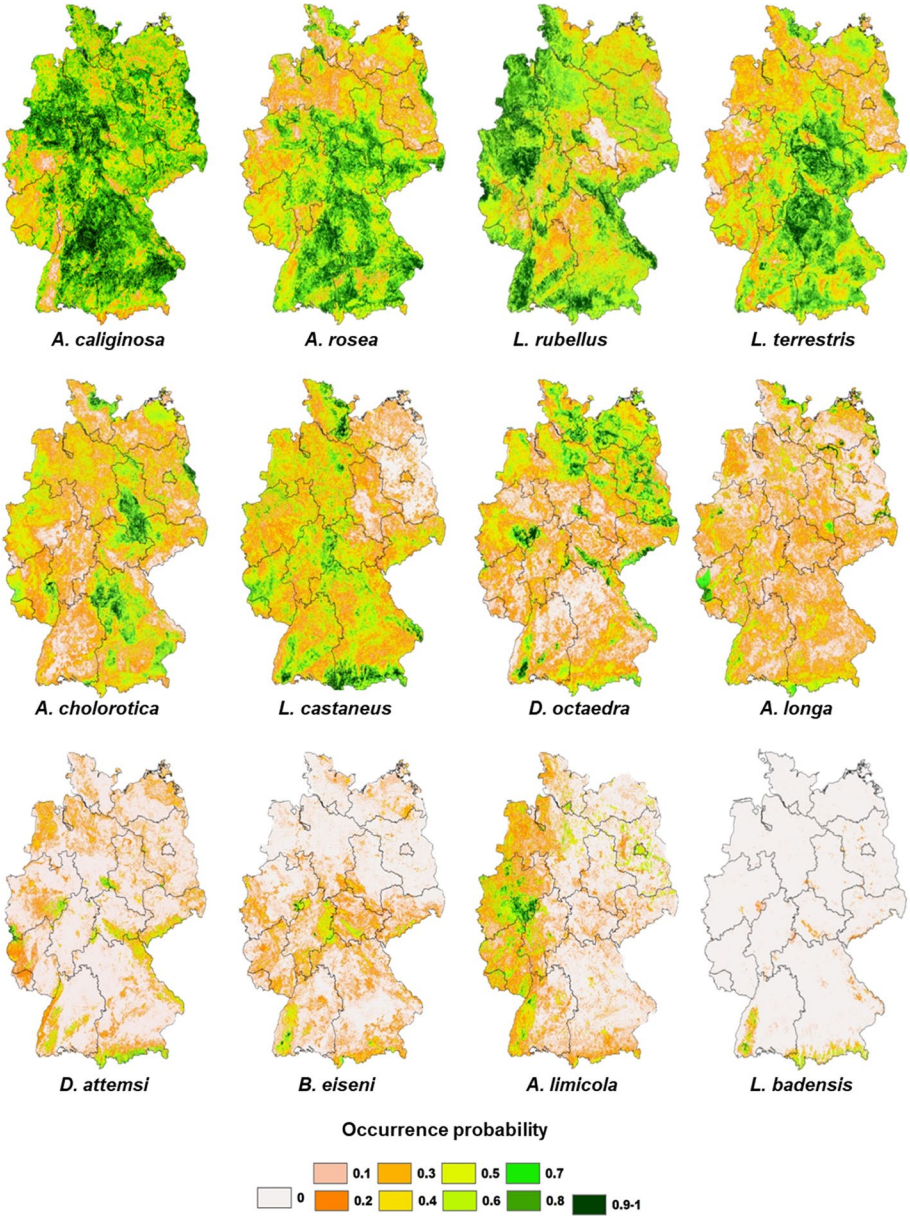

Fig. 5 Distribution probabilities of selected species predicted by the Random-Forest models, grouped into different predicted range sizes (top row: large-range species; middle row: mid-range species, bottom row: restricted-range and endemic species)

occurrence predictions. Also, the occurrence probability of *D. octaedra* was predicted to be highly related (16%) to annual average temperature.

Habitat type was the next environmental predictor identified as contributing importantly to species' occurrence probabilities. With an average contribution of 13% and

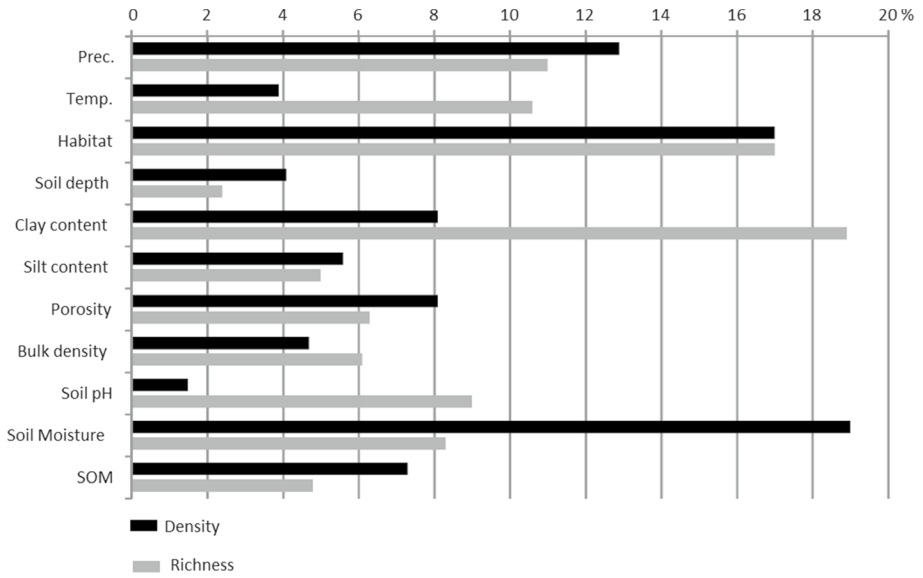

Fig. 6 Contribution (in %) of the environmental predictor variables to the predictions of earthworm community total density and species richness. *Prec., Temp.* average annual precipitation and temperature, respectively, *habitat* habitat type (see "Methods"), *SOM* Soil Organic Matter; all (soil) contents in %

contributions to individual species ranging from 9.2 to 29%, this predictor influenced all species occurrence probabilities. Only for *B. eiseni* was habitat type less important (7%).

Concerning soil parameters, the contribution of *soil pH* to the occurrence predictions was moderate across all species, on average ranging from 6 to 22%, with an average of ca. 10%. *A. longa* and *A. caliginosa* were predicted to be highly dependent on soil pH, presenting a high contribution of > 20%. The occurrence of *A. limicola* and *L. badensis* appeared to be less dependent on soil pH, with contributions of only 2.6 and 0%, respectively. The contribution of *soil organic matter* was predicted to be higher mostly for epigeic species (ca. 7–13%; except for *D. octaedra* [4.4%]) and endogeic species such as *A. limicola* and *A. caliginosa*, compared to much lower prediction contributions of less than 6% for the majority of endogeic and anecic species. Soil texture *(clay and silt content*) had comparatively less influence on species' occurrences (< 10%), except for *A. longa* and *D. octaedra*, where silt contributed 15% to the occurrence predictions; the probabilities of *L. terrestris* (silt = 11%) and *B. eiseni* (clay = 10%) were moderately influenced by soil texture. Soil structure (*porosity and bulk density*) in general only moderately influenced model results (ca. 7% on average), although occurrence probabilities of species such as *L. rubellus* and *L. badensis* were highly influenced by bulk density and *L. castaneus* and *A. rosea* by porosity. *Soil depth* was the variable contributing least to the model results, influencing the occurrence probabilities of most species by only 2–5%.

Species' responses to environmental predictor variables While the relative contributions of predictors to model results allow identification of potentially important drivers of earthworm distribution, response curves to quantitative predictors can give insight to how these

Table 5 Relative contributions (in %) of environmental predictor variables to the selected species' modelled occurrence probabilities

Species	Prec	Temp	Habitat	Soil depth	Clay content	Silt content	Porosity	Bulk density	Soil pH	Moisture	SOM
B. eiseni	20.0	15.0	7.0	2.4	10.0	4.0	11.2	3.1	7.0	3.0	9.0
D. attemsi	12.0	21.0	29.0	3.4	4.0	1.3	0.9	4.5	9.0	12.0	13.2
D. octaedra	8.0	16.0	9.4	4.5	6.5	15.0	7.9	8.0	8.0	11.0	4.4
L. castaneus	17.0	5.2	15.0	4.5	9.0	6.6	11.0	6.5	6.4	10.0	7.2
L. rubellus	20.0	10.0	12.0	2.2	5.0	4.0	7.0	17.0	6.0	9.0	8.0
A. caliginosa	6.0	9.7	15.1	2.4	5.2	3.9	9.4	4.3	22.0	9.2	12.5
A. chlorotica	12.0	13.0	12.0	1.8	8.2	7.0	4.5	3.0	8.0	26.0	3.2
A. limicola	12.0	5.2	17.0	4.2	3.0	2.3	7.7	2.2	2.6	29.0	15.0
A. rosea	10.0	7.5	9.2	3.0	8.5	9.6	10.0	9.4	11.0	16.0	5.4
A. longa	8.6	6.3	12.0	2.3	4.8	15.0	5.1	6.3	23.0	12.0	4.1
L.badensis	20.0	0.0	9.0	0.0	0.0	2.0	3.0	21.0	0.0	15.0	1.2
L. terrestris	10.0	9.0	9.2	2.5	5.1	11.0	5.0	6.5	13.0	21.0	5.7
Mean	13.0	9.8	13.0	2.8	5.8	6.8	6.9	7.7	9.7	14.4	7.4

Horizontal dotted lines group species to life-form types (top: epigeic species, middle: endogeic species, bottom: anecic species). *Temp., Prec.* average annual temperature [in °C] and precipitation (in mm/a), respectively. Soil depth in cm. *Clay, Silt* soil clay and silt content, respectively, in mass%. *Porosity* total pore space in vol%. *Bulk dens.* soil bulk density in g/cm^3. *Soil moisture* soil water content in vol%. *SOM* soil organic matter in mass%

drivers affect species' potential distribution. Climate parameters *(average annual temperature* and *precipitation)* and the related *soil moisture* often presented consistent patterns of influence on occurrence predictions, which differed between epigeic and endogeic species (Fig. 6). The occurrence probabilities of epigeic species generally decreased with increasing annual mean temperature (usually above 6–8 °C) and inversely with decreasing mean annual precipitation and soil moistures. The precipitation thresholds at which occurrence probabilities increased were species specific, where the probabilities of, e.g., *D. attemsi*, *L. castaneus*, and *L. rubellus* increased above precipitations of 500 mm/and those of *B. eiseni* and *D. octaedra* above ca. 1000 mm/a (Supplementary Fig. 2a). Increased occurrence probabilities of these species were predicted at different levels of soil moisture except for *B. eiseni*, whose occurrence predictions decreased above soil-H_2O contents of ca. 6%. In contrast, the occurrence predictions of endogeic species increased with higher annual mean temperature and inversely often decreased with lower annual mean precipitation and soil moisture (Supplementary Fig. 2b). Again, the thresholds at which occurrence predictions changed were species specific. An exception represented *A. limicola*, whose occurrence probabilities increased with lower annual mean temperature (below ca. 7 °C) and higher annual precipitation (ca. 800 mm/a) and average soil moisture (> ca. 18% H_2O). Anecic species did not show a common pattern and were species specific (Supplementary Fig. 2b).

Species' prediction responses to *soil pH* were also different between epigeic and endogeic/anecic species. (Supplementary Fig. 2a, b). The occurrence probabilities of epigeic species were generally larger at lower pH values, while that of endogeic and anecic species were larger at higher pH values. Exceptional was the epigeic *L. castaneus*, whose prediction response increased at higher pH values. Interestingly, the threshold at which occurrence probabilities either increased or decreased were generally around pH 4.0, with an optimum (for species with higher probabilities at higher pH values) of around pH 6–7. Exceptions were found for, i.e., *A. chlorotica*, *L. rubellus* and *L. terrestris*, whose thresholds were around pH 5, or *A. longa* with increasing occurrence probabilities above soil-pH values of 6.

No consistent response to *soil organic matter* (SOM) was observed among life-form groups, and the prediction responses were instead species specific (Supplementary Fig. 2a, b). The epigeic *D. octaedra* and *D. attemsi*, the endogeic *A. chlorotica* and *A. rosea*, as well as the anecic *A. longa* and *L. terrestris* all showed reduced occurrence probabilities with increasing SOM content. These species optima where generally predicted to be < 6% SOM. Contrarily, the epigeic *L. castaneus* and *L. rubellus*, the endogeic A. *caliginosa* and *A. limicola* showed positive responses to increasing SOM, with maximum probabilities shown to be around 8–12% SOM content.

Soil texture parameters (*clay and silt content*) only contributed substantially to the occurrence predictions of a few species (Supplementary Fig. 2a, b), and were not related to life-form type. For instance, the predicted occurrence probabilities were larger at both higher clay and silt contents for *B. eiseni* (epigeic), *A. rosea*, *A. caliginosa* (both endogeic) and *L. terrestris* (anecic)—as did those of *A. limicola* at higher clay content—suggesting a preference for finer textured soils by these species. Contrarily, the responses of *L. rubellus* (epigeic) and *A. longa* (anecic) generally decreased with higher clay and silt content—and that of *D. attemsi* (epigeic) and *A. chlorotica* decreased with higher clay content—suggesting preferences for coarser soils. The other species only showed irregular responses or very low prediction contributions to soil texture.

Concerning soil structure, differences between life-form types were observed (Supplementary Fig. 2a, b). Most epigeic species exhibited higher occurrence probabilities at higher levels of porosity and lower bulk densities, suggesting a preference for looser soil.

Contrarily, the predicted responses of some endogeic species (i.e., *A. caliginosa* and *A. rosea*) and the anecic *L. terrestris* where larger at lower levels of porosity and higher bulk densities (as was the response of *A. limicola* at larger bulk densities), suggesting a preference for denser soils. The predicted response of the other endogeic and anecic species was irregular and/or with very low contributions to predicted occurrence probability.

As a categorical explanatory variable, species response curves could not be calculated for *habitat*. Nonetheless, the predicted species' occurrence probabilities to the levels (individual habitat types) of this variable can indicate species responses to habitat type. These results showed that the largest occurrence probabilities were found primarily for forests, grassland, and arable land, while marine-influenced (i.e., islands) and coastal habitats as well as bogs & fens and scrubland were the least preferred (Table 6). Most species were predicted to occur in all habitats with at least low occurrence probabilities (Table 6), except for *L. badensis*, which was predicted to occur primarily in forests and grassland. A few species were predicted to occur in many habitat types with high probabilities, indicating a generalist nature, i.e., *A. caliginosa*, *A. longa*, *A. rosea*, *L. rubellus* as well as *L. castaneus* (in more developed habitat types).

The occurrence predictions also identified some species as potentially having unique or main habitat preferences, such as *D. octaedra* and *B. eiseni* in forests, *L. terrestris* in grassland and forests, or *A. chlorotica* and *D. attemsi* in agricultural habitats (arable land and grassland). *A. limicola* was predicted to have high occurrence probabilities in wetland habitats such as floodplains and bogs & fens. Interesting were the high occurrence probabilities predicted for a few species in urban gardens and parks as well as post-mining areas. Habitat types were generally not predicted to harbor specific life-form types, except for bogs & fens (where mostly endogeic and epigeic species had the largest occurrence probabilities) or grasslands (with endogeic and anecic species predicted to have the highest occurrence probabilities; Table 6).

Discussion

This study followed standard ODMAP protocols for implementing species distribution models (SDMs) (Zurell et al. 2020), using multiple model algorithms to compare model performance and select the best performing model for projecting earthworm spatial distribution throughout Germany. SDM objectives, focus taxa and spatial scales of a study are major components in ODMAP protocols, as these determine the methods used in SDMs. Furthermore, we used several methods (expert judgement, statistical techniques, ecological-relevance analysis) to select environmental predictor variables, thus enabling us to cover a large spectrum of relevant environmental variables for robust modelling of earthworm species distribution.

Machine learning algorithms such as Random Forests (RF), Generalised Boosted regression Models (GBM) or MaxEnt have been suggested to perform better than traditional regression models such as Generalized Linear regression Models (GLM) or Generalized Additive regression Models (GAM) (Elith et al. 2006; Li and Wang 2013; Valavi et al. 2022). Although RF has hitherto only rarely been used and its potential underutilized in SDMs, its high prediction performance has recently attracted attention in applied ecological studies (e.g., Mi et al. 2017). RF and GBM have been described as ensemble classifiers, which consist of and use several alternative trees in decision making while building model predictions (Li and Wang 2013; Guisan et al. 2017). Previous studies on

Table 6 Predicted occurrence probabilities of the individual species in the different habitat types (= levels of the categorical variable "habitat")

	Marine-influenced	Coastal	Floodplains	Bogs and fens	Grasslands	Scrubland	Forests	Sparse vegetation	Arable land	Urban and post-mining
B. eiseni	**0.01**	0.20	0.20	**0.11**	0.48	**0.12**	0.65	**0.04**	0.45	0.40
D. attemsi	**0.10**	**0.06**	**0.13**	0.35	0.30	0.25	0.45	0.24	0.55	**0.10**
D. octaedra	**0.10**	0.20	0.25	0.20	0.26	0.20	0.64	0.43	0.42	0.34
L. castaneus	0.20	**0.00**	0.50	**0.01**	0.35	**0.01**	0.59	0.11	0.48	0.44
L. rubellus	0.30	**0.02**	0.35	**0.01**	0.56	**0.10**	0.60	**0.04**	0.48	0.34
A. caliginosa	**0.05**	0.20	0.51	**0.15**	0.55	0.50	0.55	0.34	0.52	0.56
A. cholorotica	**0.03**	**0.05**	**0.04**	**0.10**	0.60	**0.10**	0.30	**0.10**	0.60	0.30
A. limicola	0.34	0.20	0.55	0.46	0.50	0.20	0.58	0.30	0.52	**0.10**
A. rosea	**0.03**	**0.02**	0.40	0.30	0.58	**0.01**	0.60	**0.12**	0.59	0.45
A. longa	**0.05**	**0.12**	**0.07**	**0.15**	0.51	0.38	0.52	0.30	0.56	0.41
L. badensis	**0.02**	**0.02**	**0.05**	**0.02**	0.30	**0.00**	0.50	**0.10**	0.20	**0.02**
L. terrestris	**0.01**	**0.04**	0.35	**0.12**	0.59	**0.03**	0.62	**0.11**	0.20	0.25
Cumulative	1.24	1.13	3.40	1.98	5.58	1.90	6.60	2.23	5.57	3.71

Probabilities are scaled between 0 and 1. "cumulative" = the cumulative *absolute* probabilities of all species in the habitat type (and not of *relative* probabilities within the habitat type), and therefore can reach values > 1; these are shown to indicate which habitat types are predicted to be most populated by the modelled species. Based on an overall average probability (of all species throughout all habitat types) of 0.28, occurrence probabilities higher than 0.2 are conservatively considered to represent a probable species' habitat preference; lower values (in bold) likely represent random predictions

earthworm distribution modelling had often used single algorithms such as GLM (Rutgers et al. 2016), BRT = GBM (Palm et al. 2013), or MaxEnt (Marchán et al. 2015, 2016), based on the specific goals of these studies. Given the potential of machine learning models with high predictive performance, comparative assessment of different modelling algorithms could aid in identifying best fitting models for upscaling observed distributional data to the national scale. Although exceeding the performance of GLM and GAM in this study, GBM only predicted within the 3rd quantile range of the total-density and species-richness data. The comparison of the goodness-of-fit statistics (R^2, CI, AUC, and Kappa), the observed:predicted data fits (data not shown) as well as the resultant maps of predicted density and richness by all models illustrated the good performance of RF. For instance, this algorithm was able to predict beyond the 3rd quartile range of density field data, including maximum densities over 600 ind/m^2. Other studies corroborate our finding of RF algorithms having the best predictive performance (e.g., Marmion et al. 2008; Mi et al. 2017; Valavi et al. 2022). We did, however, observe goodness-of-fit improvements for density predictions in RF after including additional data from Bavaria, confirming reports that RF can be data sensitive (Valavi et al. 2021; Yiu 2021). The resulting partial response curves explaining the relationships between communities (or species) and environment further exemplify how the RF models produce ecologically informative results (Cutler et al. 2007; Mi et al. 2017). A promising direction for future studies would be the use of ensemble models incorporating multiple algorithms (as in Marchán et al. 2021; Marchán and Domínguez 2022).

The high goodness-of-fit results for the RF models notwithstanding, any predictive model is only as good as the underlying data used for calibration. With over 20,000 data records from close to 1000 sites of occurrence, the biological background data can be considered large and highly sufficient. As Marchán et al. (2016) noted that prediction performance can be influenced by the amount of data used for training models, this large data set likely contributed to the high observed goodness-of-fit. Although the earthworm data records also included data on the environmental predictors in over 40% of the cases, providing high thematic association, this is patchy and needed to be augmented by external data. This is critical for soil parameters, for which it is often difficult to obtain comprehensive, nationwide data that is not based on broad interpolations (inappropriate due to the high small-scale heterogeneity of soil). However, not all relevant parameters could be included. For instance, Creamer et al. (2019, in Baritz et al. 2021) regarded indicators of soil organic-matter *quality* (i.e., C:N, N:P relationships) to also be highly important for soil organisms, which was not available for Germany. We are also aware of the potential difficulties of using external habitat data since a temporal disconnect between earthworm observation and habitat-type overviews may contain land-use changes. Fortunately, habitat type was the most common environmental metadata included with the earthworm data, ensuring a broader 1:1 association for model calibration. Finally, only abiotic variables were considered as predictor variables; any interactions with other organisms (i.e., between earthworm species, other soil fauna or microorganisms) were not considered (cf. Palm et al. 2013), since large-scale data for other organisms is also not available. Despite not being able to consider every *potential* driver of earthworm distribution, the models did include a high number of the most important environmental parameters known to effect earthworm fitness (e.g., Lee 1985; Edwards and Arancom 2022).

Although the model predictions have not yet been validated in the field, expert collations of earthworm species' autecology confirm a majority of the predictions. Notable are the predicted species' responses to soil pH, which identified many acidophobous and some acidophilus or -tolerant species, with a threshold between pH 4 and 5. Graefe and Beylich

(2003) also reported a strong species-specific differentiation, with a common threshold of pH 4.2, except for i.e., the acidophobous *A. longa* with a threshold of pH 5, as also predicted by our models. Our predictions of species' responses to soil acidity are also widely confirmed by, e.g., Sims and Gerard (1999), Jänsch et al. (2013), Krück (2018), and Sherlock (2018). These authors as well as Römbke et al. (2018) and Lehmitz et al. (2016) also described species-specific preferences for soil organic-matter (SOM) content, which were almost completely confirmed by the model predictions. Some of these authors also considered preferences for clay content, which were generally but not always confirmed by the current model predictions. For instance, Jänsch et al. (2013) reported on *D. octaedra*'s preference for soils with low clay content and *A. chlorotica*'s slight preference for clay soils, both of which were contradicted by our results. Also, our predicted positive response of *L. terrestris* to soils with lower clay and silt content is contrary to the assessment of Sims and Gerard (1999) and Sherlock (2018) that this species prefers clay-rich soils (these authors, however, regard UK populations).

Our study supported earlier work on the effects of precipitation and soil moisture, where these variables accounted for population increases and the distribution of adult earthworms (Lavelle 1978; Lavelle and Spain 2005; Kalu et al. 2015; Rajwar et al. 2022). Philips et al. (2019) via simpler statistical methods identified climate (average annual precipitation and temperature) as the almost exclusive driver of earthworm communities (total density, species richness) at a global scale. While our study confirmed the combined role of temperature, precipitation and soil moisture, it also identified, i.e., habitat type, soil pH and soil organic matter as important drivers. Since at global scales climate also drives, e.g., natural vegetation and partly also soil genesis, it is plausible that statistical methods will identify such broader-scale predictors as drivers of distribution predictions at very large spatial scales over other environmental parameters which exhibit higher small-scale variability. Marchán et al. (2015, 2016), Rutgers et al. (2016) and Gabriac et al. (2022) noted the overlap and correlation between large-scale variables and soil (micro-)variables. Depending on the scale of previous spatial studies, different predictors have been found to be significant: i.e., climate at global scales (Philips et al. 2019); climate, vegetation/land-use and topology at continental to sub-continental scales (Rutgers et al. 2016; Marchán et al. 2016; Marchán and Domínguez 2022); while soil parameters were important predictors at smaller spatial scales (Marchán et al. 2015; Marchán and Domínguez 2022; Gabriac et al. 2022). Our study at a regional scale also demonstrated the importance of habitat type (land use) and soil parameters in addition to climate variables as drivers of earthworm distribution. Therefore, the relevant drivers of earthworm biodiversity are apparently scale dependent; climate parameters being important at global and continental scales, while vegetation/habitat type and soil factors become more important at smaller spatial scales. At local scales, soil factors may increase in importance, with anthropogenic land-use measures importantly influencing earthworm biodiversity at the smallest scales (Palm et al. 2013).

This study predicted occurrence probabilities beyond the traditional forest, grassland and arable land by encompassing all terrestrial EUNIS level-1 habitat types. The predictions highly corresponded to the suggested range-size classifications. For instance, most of the "large-range" species were predicted both to be broadly distributed throughout many regions in Germany and to occur equally in many diverse habitat types, often with probabilities > 50–60%, thus indicating their ecologically generalist nature. The Red List of Germany lists these species as being "very common" (Lehmitz et al. 2016), and they have been reported to occur in many different habitat types (e.g., Sims and Gerard 1999; Jänsch et al. 2013; Römbke et al. 2018; Sherlock 2018), confirming our results. Although *L. terrestris* is generally viewed as being eurytopic, it has sometimes been noted as having a slight

preference for grassland sites (Sims and Gerard 1999; Jänsch et al. 2013; Sherlock 2018), which is confirmed by our predictions that, however, also equally predicted forest habitats. This species has been said to be disturbance intolerant (Lehmitz et al. 2016; Römbke et al. 2018), which may explain its low predicted probabilities in naturally disturbed sites (i.e., floodplains, bogs) as well as anthropogenically influenced habitat types.

The species we classified as "mid-range" were also predicted to occur widely in Germany, albeit in much lower probabilities. The German Red List lists them all as being "common". Although predicted to occur in many different habitat types, they appear to be more habitat discriminant, with preference optima in specific habitat types. For instance, *A. chlorotica* was predicted to occur more strongly in agrarian sites (arable fields or grassland), which has been reported from observational data (i.e.,Jänsch et al. 2013; Römbke et al. 2018). On the other hand, *D. octaedra* was predicted to occur mostly in forest habitats, as also noted by, e.g., Jänsch et al. (2013), Römbke et al. (2018), Sherlock (2018). Considering its acidophilus nature, its preference is likely for coniferous forests (cf. Sherlock 2018). While *A. castaneus* seems to be more generalist in nature, we predicted its highest occurrence probabilities to be in forests and floodplains, which is contradicted by, i.e., Jänsch et al. (2013), Römbke et al. (2018), and Krück (2018), who consider its preference to also be for grasslands. Interestingly, *A. longa* was predicted by our models to also be somewhat generalist, occurring in different habitat types, but to be missing in wetter habitats (e.g., islands, coastal, floodplains, bogs). This is confirmed by Krück (2018), who attested a preference for dryer habitats, but contradicted by Sims and Gerard (1999), who noted its occurrence in floodplains of the UK.

The species identified as "restricted-range" all showed higher occurrence probabilities limited to specific regions and habitat types. The Red List of Germany lists them all as being "rare" or "very rare". The highest distribution probabilities of *D. attemsi* were more in hilly or mountainous regions of Germany; its highest probabilities were predicted for arable land (and secondarily in forests), which has not been noted by previous authors (except Sherlock 2018). *B. eiseni* was predicted by the models to most likely occur in forests (of central and southern Germany), as also documented by Römbke et al. (2018) and Lehmitz et al. (2016). *A. limicola* is known to be hydrophilous (Sims and Gerard 1999; Lehmitz et al. 2016; Krück 2018; Römbke et al. 2018; Sherlock 2018). Accordingly, the models predicted it to occur in floodplains with high probabilities, as well as in grasslands and forests (which may also be located in floodplains or similar, but possibly misclassified to more general habitat types in the data). The models predicted it to occur mostly in western Germany (and most strongly along the Rhine River valley), confirming Krück (2018) who noted its very rare occurrence in northeastern Germany. *L. badensis* is an endemic species of Germany, occurring in forests of the High Black Forest (southwestern Germany) (Lehmitz et al. 2016), as confirmed by the model predictions.

Interesting are the few species predicted to occur in "fringe" habitats. For instance, *A. limicola*, *L. rubellus* and *L. castaneus* were predicted to occur in marine-influenced habitats (i.e., islands), and *A. caliginosa*, *A. chlorotica* and *A. limicola* in coastal sites; all however with low (< 35%) occurrence probabilities, indicating patchy occurrence in these habitats. Notable were the large number of species predicted in moderate probabilities to occur in urban, industrial, and other anthropogenic sites. Although the occurrence probabilities in these habitat types were often low, these results represent the first of their kind and can help assessment of soil-biodiversity surveys in such areas.

At the community level, the inverse relationship between species richness and total density (as found primarily for north-eastern Germany) is a common occurrence in ecology, where an area may exhibit high individual densities, but low species richness (Verberk

2011). Less favorable environmental conditions may only allow the occurrence of few species, but these in large populations, perhaps due to reduced competition from other species (Groves 2022). This could possibly explain the high individual density, but low species richness predicted for north-eastern Germany, which is known for dryer, sandy soils and where forests are usually coniferous plantations. In this regard, the predicted high occurrence probabilities of *D. octaedra* and *L. rubellus* in these areas are conspicuous, both of which are epigeic acidophilous (or –tolerant) species with a predicted affinity for forests. Personal observations have often shown high population densities of very few epigeic species in forests on sandy soils, rendering this explanation plausible. On the other hand, the Bavarian Alps and Rhine valley were predicted to be among the few regions with high earthworm biodiversity (both total density and species richness) in Germany. The Rhine Valley is known for rich soils and high general biodiversity and the predictions in alpine regions support the Alpine convention declaration of the Alps being one of the richest regions in Europe in terms of plant and animal diversity (Alpine convention 2014).

Range size has long been recognized as being a good indicator for assessing a species' threat status, which has rarely been used for soil organisms (but see Marchán and Domínguez 2022 for a good example). This study did not intend to create a red list, as this already exists for earthworms in Germany (Lehmitz et al. 2016). However, mapping earthworm species' spatial distribution and determining their geographic range size helped categorize species into range-size groups, which enables a rapid assessment of species' threat status and their conservation needs and provides valuable information for setting conservation priorities (IUCN 2012a, b, 2022). Within Germany, no species was considered threatened under extent-of-occurrence (EOO) criteria (all studied species' EEOs exceeded the minimum threshold of 20,000 km^2; IUCN 2012a, b). Nonetheless, a comparison of the predicted distribution maps and the calculated area of occupancy (AOO) shows that certain species should be of concern due to their restricted AOO range in Germany, i.e., *B. eiseni*, *D. attemsi* and *A. limicola* or due to being endemic in Germany, such as *L. badensis*.

Divergent opinions exist on the status of *B. eiseni* in Germany; while Bouché (1972) and Graff (1953) classified it as rare in France and Germany, respectively, this was contradicted by Römbke et al. (2018) who considered the species to be common. Our findings tend to partly support the earlier opinions of Graff (1953) and Bouché (1972) and the intermediary position of Lehmitz et al. (2016), who judged it to be moderately common (we prefer the term "restricted range "). While our predictions confirmed the restricted occurrence of *B. eiseni* in Hessian forests (Römbke et al. 2018) and a few other clusters, caution must be exercised. The species is assumed to be arboreal and corticolous, and the limited observational data may be methodologically biased, as common earthworm extraction methods may not sufficiently sample this species' preferred microhabitat (Lehmitz et al. 2016; Römbke et al. 2018). *A. limicola* is the only species studied here that is listed as endangered in the German Red List of earthworms. Its predicted occurrence in wetland sites, corroborating its status as a hydrophilous species, as well as its predicted restricted distribution in Germany, supports its endangered status. The remaining species should be viewed as species of focus in future soil- biodiversity surveys.

Since earthworm taxonomy and systematics are constantly in flux, especially with recent molecular studies (e.g.,Pérez-Losada et al. 2012; Domínguez et al. 2015), caution is perhaps advised regarding taxa that may include different morphotypes or cryptic species, which can exhibit different ecological preferences and, therefore, drivers of their distribution. Two examples are *A. caliginosa* (i.e.,Pérez-Losada et al. 2009; Briones 2011; Fernández et al. 2012) and *A. chlorotica* (see Lowe and Butt 2008; Dupont et al. 2011, 2022). It is impossible to reevaluate retrospectively the species identifications included in the analyzed

dataset, so that this remains an issue of taxonomy and species identification and not of the modelling procedures. Nonetheless, it must be mentioned that data used in the current study can partly represent species complexes, and the predicted distributions may subsume these.

Especially noted is furthermore *L. badensis,* an endemic species found in the high Black Forest region, and which is likely endangered (Lehmitz et al. 2016). Although insufficient observational data was available for calculating EOO or AOO, it was predicted to have a very restricted and narrow distributional range, corroborating its assessment as endangered. This species also highlights an important aspect of species distribution modelling: although the models predict potential occurrence e.g., in the Bavarian Alps, the species has never been found to occur there. The model results primarily show the high potential habitat *suitability* for the species in the Alps, and are not proof of occurrence, thus underscoring its status as endemic to southwest Germany.

Conclusion

This study is, to the best of our knowledge, the first comprehensive analysis modelling earthworm distribution at a national scale, including the most important species and differentiating among multiple environmental drivers. Earlier earthworm SDM studies generally used single modelling frameworks (GLM, GLMM, GAM, MaxEnt or BRT; i.e., Rutgers et al. 2016; Philips et al. 2019; Marchán et al. 2016; Palm et al. 2013; but see Marchán et al. 2016; Marchán and Domínguez 2022 for ensemble species distribution models (ESDMs)). Given the potential of machine learning models with high predictive performance, this study compared the predictive performance of traditional regression models (GLM, GAM) with machine learning algorithms (GBM and RF) to identify the best statistical model for predicting earthworm biodiversity across Germany. The predictive performance of RF was outstanding.

These predictions, including classifying species into different range-size groups as well as community and species- specific responses to a broad spectrum of environmental variables, provide an effective national-scale approximation of earthworm distribution and its drivers in Germany. Such information is invaluable for future scientific field studies and a prerequisite for soil-biodiversity monitoring programs, which require standardized baseline values for result assessment. A tool is currently being developed to extract reference values from the model results based on specific site conditions, explicitly for use in soil-biodiversity monitoring programs. While such programs will help validate the model results, we call for wide-spread recording of environmental (especially soil) parameters concomitantly with biodiversity surveys, to improve the thematic association between species and environmental drivers and, thereby, model precision.

Modelling and mapping earthworm distributions further allowed grouping species into geographic range-size classes, providing vital information for decision making on conservation priorities. Previous studies were improved by projecting species distribution into 10 habitat classes at the top hierarchy level (EUNIS level-1 habitat types). Despite this improvement, earthworm distribution can still be highly variable within level-1 habitat types. For instance, “forest” can be subdivided into deciduous, coniferous, mixed deciduous/coniferous forests, among others; management measures in agricultural habitats have a strong influence on earthworm communities. Availability of high-resolution raster data at EUNIS level-2 or finer hierarchies will increase model precision on species’ habitat

preferences. Importantly, recording of habitat types at a more differentiated level during soil-biodiversity surveys will highly improve future data syntheses, allowing better conservation decisions. Although the RF model was able to predict the potential distribution of some species in largely un-sampled areas of northern Germany, thereby demonstrating its ability as a "non-overfitting" model to predict beyond the training datasets, this cannot replace true field observations. We therefore also call for more data to be collected in these un-sampled areas and be made readily available for future synthesis analyses.

Attention should be given to species with restricted ranges, such as *D. attemsi*, *B. eiseni*, and *L. badensis*. For species with clearly defined habitats, such as *A. limicola* in wetlands and *D. octaedra* in forests, the habitats in which they can be found should be monitored for possible habitat degradation (Global 2022). We further suggest detailed studies on the endemic *L. badensis*, which would allow more precise SDMs and calculation of geographic range sizes, providing a better assessment of its realized distribution and protection needs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10531-023-02608-9.

Acknowledgements We thank the German Federal Institute of Geoscience and Natural Resources (BGR) for provision of soil data.

Author contributions DJR and EE conceived the study; GS, DJR and AS collated and cleaned the data; GS ran the models; GS, DJR wrote the draft manuscript. All authors participated in writing, editing and revising the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was funded by the German Environmental Agency (UBA) under reference number FKZ 3719-71-206-0.

Data availability Data used in this study are contained in open access in the Edaphobase data warehouse and the specific data sets are available upon request via the corresponding author. R code script files used are deposited at https://team.eudaphobase.eu/index.php/s/rJBcPJKdDCT4W5s.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Adhikari K, Hartemink AE (2015) Linking soils to ecosystem services: a global review. Geoderma 262:101–111

Aiello-Lammens MA, Boria RA, Radosavljevic A, Vilela B, Anderson RP (2015) spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography 38:1–10

Alboukadel K (2021) rstatix: pipe-friendly framework for basic statistical tests. R package version 0.7.0. https://CRAN.R-project.org/package=rstatix

Alpine Convention (2014) Guidelines for climate change adaptation at the local level in the Alps. Alpine signals 7. Permanent Secretariat of the Alpine Convention, Innsbruck. ISBN: 9788897500247.

Baritz R, Amelung W, Antoni V, Boardman J, Horn R. Prokop, G, Römbke J, Romkens P, Steinhoff-Knopp B, Swartjes F, Trombetti M, de Vries W (2021) Soil monitoring in Europe. Indicators and thresholds for soil quality assessments. EEA ETC/ULS Report. European Environmental Agency. https://ww.eea.europa.eu/publications#c7=en&c11=25&c14=&c12=&b_start=0&c13=soil

Biber MF, Voskamp A, Niamir A, Hickler T, Hof C (2020) A comparison of macroecological and stacked species distribution models to predict future global terrestrial vertebrate richness. J Biogeogr 47:114–129

Blanchart E, Albrech A, Alegre J, Duboisset A, Giloe C, Pashanas B, Lavelle P, Brussaard L (1999) Effects of earthworms on soil structure and physical properties. In: Lavelle P, Brussaard L, Hendrix P (eds) Earthworm management in tropical agroecosystems. CAB International, pp 149–172

Blouin M, Hodson ME, Delgado EA, Baker G, Brussaard L, Butt KR, Dai J, Dendooven L, Peres G, Tondoh JE, Cluzeau D, Brun JJ (2013) A review of earthworm impact on soil function and ecosystem services. Eur J Soil Sci 64:161–182

Bobrowski M, Weidinger J, Schwab N, Schickhoff U (2021) Searching for ecology in species distribution models in the Himalayas. Ecol Model 458:109693. https://doi.org/10.1016/j.ecolmodel.2021.109693

Boria AR, Olson LE, Goodman SM, Anderson RP (2014) Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol Model 275(2014): 73–77

Bouché MB (1972) Lombriciens de France. Ann Zool Ecol Anim 72(2):1–671

Boumal J, Montanarella L (2016) Facing policy challenges with inter- and transdisciplinary soil research focused on the UN sustainable development goals. Soil 2:135–145

Briones MJI (2011) A taxonomic revision of the *Allolobophora caliginosa* complex (Oligochaeta, Lumbricidae): a preliminary study. Can J Zool 74:240–244

Briones MJI (2018) The serendipitous value of soil fauna in ecosystem functioning: the unexplained explained. Front Environ Sci 6:149. https://doi.org/10.3389/fenvs.2018.00149

Brussaard L (1998) Soil fauna, guilds, functional groups and ecosystem processes. Appl Soil Ecol 9(123):135

Burkhardt U, Russell DJ, Decker P, Döhler M, Höfer H, Lesch S, Rick S, Römbke J, Trog C, Vorwald J, Wurst E, Xylander WE (2014) The Edaphobase project of GBIF-Germany—A new online soil-zoological data warehouse. Appl Soil Ecol, 83;3–12

Cluzeau D, Guernion M, Chaussod R, Martin-Laurent F, Villenave C, Cortet J, Ruiz-Camacho N, Pernin C, Mateille T, Philippot L, Bellido A, Rougé L, Arrouays D, Bispo A, Pérès G (2012) Integration of biodiversity in soil quality monitoring: baselines for microbial and soil fauna parameters for different land-use types. Eur J Soil Biol 49:63–72

Cobos ME, Barve V, Barve N, Jimenez-Valverde A, Nuñez-Penichet C (2021) rangemap: simple tools for defining species ranges. https://cran.r-project.org/web/packages/rangemap/index.html

Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler J (2007) Random forest classification in ecology. Ecology 88:2783–2792

Domínguez J, Aira M, Breinholt JW, Stojanovic M, James SW, Pérez-Losada M (2015) Underground evolution: new roots for the old tree of lumbricid earthworms. Mol Phylogenet Evol 83:7–19

Dorigo WA, Wagner W, Albergel C, Albrecht F, Balsamo G, Brocca L, Chung D, Ertl M, Forkel M, Gruber A, Haas E, Hamer PD, Hirschi M, Ikonen J, de Jeu R, Kidd R, Lahoz W, Liu YY, Miralles D, Mistelbauer T, Nicolai-Shaw N, Parinussa R, Pratola C, Reimer C, van der Schalie R, Seneviratne SI, Smolander T, Lecomte P (2017) ESA CCI soil moisture for improved earth system understanding: state-of-the art and future directions. Remote Sens Environ. https://doi.org/10.1016/j.rse.2017.07.001

Dupont L, Lazreka F, Porco D, King RA, Rougerie R, Symondson WOC, Livet A, Richard B, Decaëns T, Butt KR, Mathieu J (2011) New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. Pedobiologia 54:217–224

Dupont L, Audusseau H, Porco D, Butt KR (2022) Reproductive strategies in a complex of simultaneously hermaphroditic species, the *Allolobophora chlorotica* case study. BioRxiv. https://doi.org/10.1101/2022.01.31.475338

Edwards CA, Arancon NQ (2022) Biology and ecology of earthworms, 4th edn. Springer, New York

Elith J, Graham CH, Anderson RP et al (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151

FAO, ITPS, GSBI, CBD, and EC (2020) State of knowledge of soil biodiversity: status, challenges and potentialities. Report 2020. FAO, Rome

Fernández R, Almodóvara A, Novo M, Simancas B, Díaz Cosín DJ (2012) Adding complexity to the complex: new insights into the phylogeny, diversification and origin of parthenogenesis in the *Aporrectodea caliginosa* species complex (Oligochaeta, Lumbricidae). Mol Phylogenet Evol 64:368–379

Gabriac Q, Ganault P, Barois I, Aranda-Delgado E, Cimetière L, Cortet J, Gautier M, Hedde M, Marchán DF, Pimentel Reyes JC, Stokes A, Decaëns T (2022) Environmental drivers of earthworm communities

along an altitudinal gradient in the French Alps. BioRxiv 5:1–10. https://doi.org/10.1101/2022.10.13.512055
Gardi C, Jeffery S (2009) Soil biodiversity. JRC Sci Techn Rep. https://doi.org/10.2788/7831
Gaston KJ, Fuller RA (2009) The sizes of species' geographic range. J Appl Ecol 46:1–9
Global Invasive Species Database (2022) Species profile: *Dendrobaena octaedra*. http://www.iucngisd.org/gisd/species.php?sc=1710. Accessed 14 April 2022.
Graefe U, Beylich A (2003) Critical values of soil acidification for annelid species and the decomposer community. Newslett Enchytr 8:51–55
Graff O (1953) Die regenwürmer deutschlands. Schrift Forschung Landwirt 7:1–70
Griffiths BS, Römbke J, Schmelz RM, Scheffczyk A, Faber JH, Bloem J, Pérès G, Cluzeau D, Chabbi A, Suhadolc M, Sousa JP, Martins da Silva P et al (2016) Selecting cost effective and policy-relevant biological indicators for European monitoring of soil biodiversity and ecosystem function. Ecol Ind 69:213–223
Groves CP (2022) "Biogeographical region" in encyclopedia britanica. www.britanica.com/science/biogeog. Accessed 26 April 2022
Gruber A, Scanlon T, van der Scalie R, Wagner W, Dorigo W (2019) Evolution of the ESA CCI soil moisture climate data records and their underlying merging methodology. Earth Syst Sci Data 11:717–739. https://doi.org/10.5194/essd-11-717-2019
Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecol Model 135:147–186
Guisan A, Thuiller W, Zimmermann NE (2017) Habitat suitability and distribution models: with applications in R. Cambridge University Press, Cambridge, p 478
Hijmans RJ, Elith J (2019) Spatial distribution models, spatial data science with R. https://rspatial.org/sdm/SDM.pdf
Hijmans RJ, Phillips S, Leathwick J, Elith J (2020) dismo: species distribution modeling. R package version 1.3-3. https://CRAN.R-project.org/package=dismo
Huber S, Prokop G, Arrouays D, Banko, G, Bispo A, Jones RJA, Kibblewhite MG, Lexer W, Möller A, Rickson RJ, Shishkov T, Stephens M, Toth G, Van den Akker JJH, Varallyay G, Verheijen FGA, Jones AR (eds) (2008) Environmental Assessment of Soil for Monitoring. Volume I: Indicators & Criteria. EUR 23490 EN/1. Office for the Official Publications of the European Communities, Luxembourg
IUCN (2012a) Guidelines for application of IUCN red list criteria at regional and national levels: version 4.0. IUCN, Gland. www.iucnredlist.org/technical-documents/categories-and-criteria
IUCN (2012b) IUCN red list categories and criteria: version 3.1. 2nd edn. IUCN, Gland. www.iucnredlist.org/technicaldocuments/categories-and-criteria
IUCN Standards and Petitions Committee (2022) Guidelines for Using the IUCN red list categories and criteria. Version 15. https://www.iucnredlist.org/documents/RedListGuidelines.pdf
Jänsch S, Steffens L, Höfer H, Horak F, Roß-Nickoll M, Russell D, Toschki A, Römbke J (2013) State of knowledge of earthworm communities in German soils as a basis for biological soil quality assessment. Soil Organ 85(3):215–233
Jiménez-Valverde A (2011) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob Ecol Biogeogr 21(4):498–507. https://doi.org/10.1111/j.1466-8238.2011.00683
Johnson CM, Johnson LB, Richard C, Beasley V (2002) Predicting the occurrence of amphibians: An assessment of multiple-scale models. In: Scott JM, Heglund PJ, Samson F, Haufler J, Morrison M, Raphael M, Wall B (eds) Predicting species occurrences: issues of accuracy and scale. Island Press, Covelo, pp 157–170
Kalu S, Koirala M, Khadaka RJ (2015) Earthworm population in relation to different land use and soil characteristics. J Ecol Nat Environ 7(5):124–131
Karger DN, Conrad O, Böhner J, Kawohl T, Kreft H, Soria-Auza RW, Zimmermann NE, Linder P, Kessler M (2017) Climatologies at high resolution for the Earth land surface areas. Sci Data 4:170122. https://doi.org/10.1038/sdata.2017.122
Krück S (2018) Bildatlas zur Regenwurmbestimmung. Natur+Text, Rangsdorf
Kumar S, Stohlgren TJ (2009) Maxent modelling for predicting suitable habitat for threatened and endangered tree *Canacomyrica monticola*. New Caledonia J Ecol Nat Environ 1:94–98
Lavelle P (1978) Les vers de terre de la savane de lamto (Cote d'voire) peuplements, population et functions dans l'ecosytems. Disseration, Université Paris VI/ENS
Lavelle P, Spain VA (2005) Soil ecology. Springer, Dordrecht
Lavelle P, Decaëns T, Aubert M, Barot S, Blouin M, Bureau F, Margerie P, Mora P, Rossi J-P (2006) Soil invertebrates and ecosystem services. Eur J Soil Biol 42:S3–S15

Lee KE (1985) Earthworms: their ecology and relationships with soils and land use. Academic Press, Ann Arbor, p 411
Lehmitz R, Römbke J, Graefe U, Beylich A, Krück S (2016) Rote liste und gesamtartenliste der regenwürmer (lumbricidae et criodrillidae) Deutschlands. Nat Biol Vielfalt 70(4):565–590
Li X, Wang YL (2013) Applying various algorithms for species distribution modelling. Integr Zool 8:124–135
Lowe CN, Butt KR (2008) *Allolobophora chlorotica* (Savigny, 1826): evidence for classification as two separate species. Pedobiologia 52:81–84
Maes J, Egoh B, Willemen L, Liquete C et al (2012) Mapping ecosystem services for policy support and decision making in the European Union. Ecosyst Serv 1:31–39
Manel S, Ceri Williams H, Ormerod SJ (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. J Appl Ecol 38:921–931
Marchán DF, Domínguez J (2022) Evaluating the conservation status of a North-Western Iberian Earthworm (*Compostelandrilus cyaneus*) with insight into its genetic diversity and ecological preferences. Genes 13:337. https://doi.org/10.3390/genes13020337
Marchán DF, Refoyo P, Novo M, Fernandez R, Trigo D, Díaz Cosín DJ (2015) Predicting soil micro-variables and the distribution of an endogeic earthworm species through a model based on large-scale variables. Soil Biol Biochem 81:124–127
Marchán DF, Refoyo P, Fernandez R, Novo M, de Sosa I, Cosín Díaz DJ (2016) Macroecological inferences on soil fauna through comparative niche modeling: the case of Hormogastridae (Annelida, Oligochaeta). Eur J Soil Biol 75:115–122
Marchán DF, Csuzdi C, Decaëns T, Szederjesi T, Pizl V, Domínguez J (2021) The disjunct distribution of relict earthworm genera clarifies the early historical biogeography of the Lumbricidae (Crassiclitellata, Annelida). J Zool Syst Evol Res 59:1703–1717. https://doi.org/10.1111/jzs.12514
Marmion M, Parviainen M, Luoto M, Heikkinen RK, Thuiller W (2008) Evaluation of consensus methods in predictive species distribution modelling. Divers Distrib 15:56–69
Mi C, Huettman F, Guo Y, Wen L (2017) Why choose random forest to predict rare species distribution with few samples in large undersampled area? Three Asian crane species models provide supporting evidence. PeerJ 5:e2849. https://doi.org/10.7717/peerj.2849
Mod HK, Scherrer D, Luoto M, Guisan A (2016) What we use is not what we know: environmental predictors in plant distribution models. J Veg Sci 27:1308–1322
Ockleford C, Adriaanse P, Berny P, Brock T et al (2017) Scientific opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms. EFSA PPR Panel EFSA J 15(2):4690. https://doi.org/10.2903/j.efsa.2017.4690
Orgiazzi A, Panagos P, Yigini Y, Dunbar MB, Gardi C, Montanarella L, Ballabio C (2016) A knowledge-based approach to estimating the magnitude and spatial patterns of potential threats to soil biodiversity. Sci Total Environ 545–546:11–20
Palm J, van Schaika NLMB, Schröder B (2013) Modelling distribution patterns of anecic, epigeic and endogeic earthworms at catchment-scale in agro-ecosystems. Pedobiologia 56:23–31
Pérez-Losada M, Ricoy M, Marshall JC, Domínguez J (2009) Phylogenetic assessment of the earthworm *Aporrectodea caliginosa* species complex (Oligochaeta: Lumbricidae) based on mitochondrial and nuclear DNA sequences. Mol Phylogenet Evol 52:293–302
Pérez-Losada M, Bloch R, Breinholt JW, Pfenninger M, Domínguez J (2012) Taxonomic assessment of Lumbricidae (Oligochaeta) earthworm genera using DNA barcodes. Eur J Soil Biol 48:41–47
Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2008) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 19:181–197
Phillips HRP, Guerra CA, Bartz MLC, Briones MJI, Brown G et al (2019) Global distribution of earthworm diversity. Science 366(6464):480–485
Pulleman M, Creamer R, Hamer U, Helder J, Pelosi C, Pérès G, Rutgers M (2012) Soil biodiversity, biological indicators and soil ecosystem services—an overview of European approaches. Current Opinion in Environmental Sustainability 4:529–538
R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/
Rajwar N, Singh V, Bhatt S, Singh Bisht S (2022) Earthworm population dynamics in three different land use systems along an altitudinal gradient (208–2609 m asl) in Kumaun Himalayas, India. Trop Ecol 63:134–140
Römbke J, Dreher P, Beck L, Hammel W, Hund K, Knoche H, Kördel W, Kratz W, Moser T, Pieper S, Ruf A, Spelda J, Woas S (2000) Bodenbiologische Bodengüte-Klassen. Umweltbundesamt, Berlin

Römbke J, Dorow WHO, Jänsch S (2018) Distribution and diversity of earthworms (Lumbricidae) in Hesse (Central Germany): current knowledge. Soil Organ 90(3):171–185
Rutgers M, Schouten AJ, Bloem J, van Eekeren N, de Goede RGM, JagersopAkkerhuis GAJM, van der Wal A, Mulder C, Brussaard L, Breure AM (2009) Biological measurements in a nationwide soil monitoring network. Eur J Soil Sci 60:820–832
Rutgers M, Orgiazzi A, Gardi C, Römbke J, Jänsch S, Keith AM, Neilson R, Boag B, Schmidt O et al (2016) Mapping earthworm communities in Europe. Appl Soil Ecol 97:98–111
Salako G, Chandalin B, Aliyu MB, Sawyerr H (2015) Modeling the suitability index of selected conifers on Mambilla Plateau Taraba State, Nigeria: implication on planted forest. Int J Agrofor Remote Sens GIS 1(1):1–9
Salako G, Oyebanji OO, Olagunju TE, Howe GT (2021) Potential impact of climate change on the distribution of some selected legumes in Cameroon and adjoining Nigeria border. Afr J Ecol 1:1–17. https://doi.org/10.1111/aje.1291
Sherlock E (2018) Key to the earthworms of the UK and Ireland, 2nd edn. FSC Publication, Shrewsbury
Sheth SN, Morueta-Holme N, Angert AL (2020) Determinants of geographic range size in plants. New Phytol 226:650–665
Sims RW, Gerard BM (1999) Earthworms. FSC Publication, Shrewsbury
Thuiller W, Georges D, Gueguen M, Engler R, Breiner F (2021) Package 'biomod2': ensemble platform for species distribution modeling. https://cran.r-project.org/web/packages/biomod2/index.html
Turbé A, de Toni A, Benito P, Lavelle P, Ruiz Camacho N, van der Putten WH, Labouze E, Mudgal S (2010) Soil biodiversity: functions, threats and tools for policy makers. Report to the European DG Environment
Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G (2021) Modelling species presence-only data with random forests. Ecography 44(12):1731–1742
Valavi R, Guillera-Arroita G, Lahoz-Monfort JJ, Elith J (2022) Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecol Monogr 92(1):e0148. https://doi.org/10.1002/ecm.1486
van Leeuwen JP, Saby NPA, Jones A, Louwagie G, Micheli E, Rutgers M, Schulte RPO, Spiegel H, Toth G, Creamer RE (2017) Gap assessment in current soil monitoring networks across Europe for measuring soil functions. Environ Res Lett 12:124007. https://doi.org/10.1088/1748-9326/aa9c5c
Verberk W (2011) Explaining general patterns in species abundance and distributions. Nat Educ Knowl 3(10):38
Weeks JM (1998) A demonstration of the feasibility of SOILPACS. Environmental Agency, London
Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol. 75(5):1182–1189
Yiu T (2021) Understanding Random forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
Zurell D, Franklin J, König C, Bouchét PJ, Dormann CF, Elith J, Fandos G, Feng X, Guillera-Arroita G, Guisan A et al (2020) A standard protocol for reporting species distribution. Ecography 43:1261–1277