

Supplementary Materials for

Global distribution of earthworm diversity

Helen R. P. Phillips*, Carlos A. Guerra, Marie L. C. Bartz, Maria J. I. Briones, George Brown, Thomas W. Crowther, Olga Ferlian, Konstantin B. Gongalsky, Johan van den Hoogen, Julia Krebs, Alberto Orgiazzi, Devin Routh, Benjamin Schwarz, Elizabeth M. Bach, Joanne Bennett, Ulrich Brose, Thibaud Decaëns, Birgitta König-Ries, Michel Loreau, Jérôme Mathieu, Christian Mulder, Wim H. van der Putten, Kelly S. Ramirez, Matthias C. Rillig, David Russell, Michiel Rutgers, Madhav P. Thakur, Franciska T. de Vries, Diana H. Wall, David A. Wardle, Miwa Arai, Fredrick O. Ayuke, Geoff H. Baker, Robin Beauséjour, José C. Bedano, Klaus Birkhofer, Eric Blanchart, Bernd Blossey, Thomas Bolger, Robert L. Bradley, Mac A. Callahan, Yvan Capowiez, Mark E. Caulfield, Amy Choi, Felicity V. Crotty, Andrea Dávalos, Darío J. Diaz Cosin, Anahí Dominguez, Andrés Esteban Duhour, Nick van Eekeren, Christoph Emmerling, Liliana B. Falco, Rosa Fernández, Steven J. Fonte, Carlos Fragoso, André L. C. Franco, Martine Fugère, Abegail T. Fusilero, Shaieste Gholami, Michael J. Gundale, Mónica Gutiérrez López, Davorka K. Hackenberger, Luis M. Hernández, Takuo Hishi, Andrew R. Holdsworth, Martin Holmstrup, Kristine N. Hopfensperger, Esperanza Huerta Lwanga, Veikko Huhta, Tunsisa T. Hurisso, Basil V. Iannone III, Madalina Iordache, Monika Joschko, Nobuhiro Kaneko, Radoslava Kanianska, Aidan M. Keith, Courtland A. Kelly, Maria L. Kernecker, Jonatan Klaminder, Armand W. Koné, Yahya Kooch, Sanna T. Kukkonen, H. Lalthanzara, Daniel R. Lammel, Iurii M. Lebedev, Yiqing Li, Juan B. Jesus Lidon, Noa K. Lincoln, Scott R. Loss, Raphael Marichal, Radim Matula, Jan Hendrik Moos, Gerardo Moreno, Alejandro Morón-Ríos, Bart Muys, Johan Neiryneck, Lindsey Norgrove, Marta Novo, Visa Nuutinen, Victoria Nuzzo, Mujeeb Rahman P, Johan Pansu, Shishir Paudel, Guénola Pérès, Lorenzo Pérez-Camacho, Raúl Piñeiro, Jean-François Ponge, Muhammad Imtiaz Rashid, Salvador Rebollo, Javier Rodeiro-Iglesias, Miguel Á. Rodríguez, Alexander M. Roth, Guillaume X. Rousseau, Anna Rozen, Ehsan Sayad, Loes van Schaik, Bryant C. Scharenbroch, Michael Schirrmann, Olaf Schmidt, Boris Schröder, Julia Seeber, Maxim P. Shashkov, Jaswinder Singh, Sandy M. Smith, Michael Steinwandter, José A. Talavera, Dolores Trigo, Jiro Tsukamoto, Anne W. de Valença, Steven J. Vanek, Iñigo Virto, Adrian A. Wackett, Matthew W. Warren, Nathaniel H. Wehr, Joann K. Whalen, Michael B. Wironen, Volkmar Wolters, Irina V. Zenkova, Weixin Zhang, Erin K. Cameron, Nico Eisenhauer

*Corresponding author. Email: helen.phillips@idiv.de

Published 25 October 2019, *Science* **366**, 480 (2019)
DOI: 10.1126/science.aax4851

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S6
Tables S1 to S4
References

Materials and Methods

Literature Search

Web of Science was searched on 18th December 2016, using the following search term: ((*Earthworm** OR *Oligochaeta* OR *Megadril** OR *Haplotaxida* OR *Annelid** OR *Lumbric** OR *Clitellat** OR *Acanthodrili** OR *Ailoscolec** OR *Almid** OR *Benhamiin** OR *riodrili** OR *Diplocard** OR *Enchytraeid** OR *Eudrilid** OR *Exxid** OR *Glossoscolec** OR *Haplotaxid** OR *Hormogastrid** OR *Kynotid** OR *Lutodrili** OR *Megascolec** OR *Microchaetid** OR *Moniligastrid** OR *Ocnerodrili** OR *Octochaet** OR *Sparganophilid** OR *Tumakid**) AND (*Diversity* OR “*Species richness*” OR “*OTU*” OR *Abundance* OR *individual** OR *Density* OR “*tax* richness*” OR “*Number*” OR *Richness* OR *Biomass*)

This search returned 7783 papers. All titles and abstracts of papers post-2000 were screened (6140 papers), and were excluded if they did not reference data suitable for the analysis (suitability discussed below). Since it was anticipated that raw data would need to be requested, papers published before 2000 were not screened, as it was unlikely that available author contact details were up-to-date. We note however that earlier publications may be useful for future research, e.g., focusing on long-term monitoring and temporal analyses. After this initial screening, PDFs of all remaining papers (n = 986) were manually screened to determine whether data were suitable.

In order to be suitable for the analysis, papers had to present (or make reference to) the following information and data:

1. Sampled earthworm communities using standard earthworm extraction methodologies, which would adequately capture quantitative information of the earthworm community, such as hand-sorting of a sufficient soil volume [e.g., (39)] or chemical expulsion from a quadrat [e.g., (40)] at two or more sites. At a minimum, total fresh biomass and/or total abundance of the earthworms at each site had to be measured. Ideally, there was data on identification of all individuals to species level, with the abundance/biomass data of each species;
2. Available geographic coordinates for all sampled sites, or maps that could be georeferenced;
3. Measurements of at least one soil property at each site (see below);
4. Information on the habitat cover and/or land use;
5. Differences in land use/habitat cover or soil properties (see below for information on the land use/habitat cover and soil properties) across the sites.

Where possible, all suitable data were taken from the 477 papers that were identified as containing suitable data. Data were extracted from figures where necessary (using IMAGEJ (41)). If data were not provided in the text or the supplementary materials, authors were contacted to obtain the raw data from each site. As some datasets remain unpublished, or are yet to be published, individual earthworm researchers were also contacted to enquire as to whether they had suitable data. Including unpublished data helps to reduce publication bias (42).

Data collation

The data taken or requested from one publication or an unpublished field campaign was considered a ‘dataset’. If a dataset contained data sampled using different methodologies, we split it into different ‘studies’ based on the methodology, as measured diversity of earthworms is highly dependent on the methods used (43). For datasets where sites were repeatedly sampled

over time, both within years and across years, we used only the first and the last sampling campaign and these were split into two studies. The modelling approach used (linear mixed-effects models, with random effects accounting for different studies) dealt with non-independence of such datasets (44).

Site level information

Sites were described as a location of one or more samples, which, when taken together, adequately captured the earthworm community. Sampling methodology, and therefore the number of samples per site, were determined by the original data collectors. But sampling effort was constant within a study. For each dataset, we collated the following information into a standardised data template: geographic coordinates for each of the sampled sites, start and end dates of sampling (month and year), and the sampling method used. For each dataset, we requested at least one soil property (pH, cation exchange capacity (CEC) or base saturation, organic carbon, soil organic matter, C/N ratio, soil texture, soil type, soil moisture) for each site, but only pH, CEC, organic carbon and soil texture (silt and clay) variables were used for this analysis. Most sites contained pH values (63.7%), 14% of sites contained organic carbon, 40% of sites contained silt and clay, but only 7.3% contained CEC. Any missing soil properties were filled with SoilGrids data, described below. If soil properties were given for different soil depths, then we calculated a weighted average (maximum soil depth = 1 m, but typically collected down to 30 cm). Using information within the published articles, and additional information provided by the data collectors, the habitat cover at each site was classified into categories based on the ESA CCI-LC 300m map (<http://maps.elie.ucl.ac.be/CCI/viewer/index.php>; Table S1).

Recorded community metrics

For each dataset, the following site-level community metrics were calculated where possible: total (adults and juveniles) abundance of earthworms at the site, total (adults and juveniles) fresh biomass of earthworms at the site, and number of species at the site. Using the area sampled at the site, both abundance and biomass were transformed to individuals per m² and grams per m², respectively, if they were not already given in that unit, to standardize the data into commonly used units. Species richness of each site was calculated from available species lists, if not already provided. Two issues arose when calculating species richness of earthworms. Firstly, many specimens were not identified to species level. Where data collectors identified a specimen as a unique morphospecies (species delineation based solely on morphological characteristics, typically identified to genus level with a unique ID differentiating from other species of the same genus, as determined by the original data collector), they were included in the species richness estimate as an additional species. Records that were not identified to species level, or identified as a morphospecies, were excluded. Secondly, typically only adult specimens of many earthworm species can be identified to species level (43), so juveniles were excluded from the calculation. Therefore, a more appropriate term would be ‘number of identified adult (morpho-) species’, but for brevity this will be referred to as ‘species richness’. Species richness was not calculated per unit area (i.e., density), as within each study the sampled area was consistent. Thus, due to the modelling framework, issues of diversity increasing with sampled area were accounted for.

Species identity

For datasets where the earthworms had been identified to species level, all species names were checked for spelling errors and synonyms. Scientific names were standardised using expert opinion (MJIB, GB, MLCB) and DriloBASE (<http://drilobase.org/drilobase>). Following standardisation, earthworm species were categorised into the three main ecological groups:

epigeics, endogeics, and anecics (45), plus a fourth minor group, epi-endogeic (species which exhibit traits of both epigeics and endogeics). Earthworms provide a variety of ecosystem functions, for example, increasing crop yield by enhancing decomposition and nutrient mineralization rates (12), but each ecological group contributes in different ways, often on the basis of their feeding or habitat preferences (45). Epigeic species are typically found in the upper layers of the soil and litter, and, amongst other roles, are important in the first stages of decomposition through burial of the litter layer (11, 46, 47). Endogeic species live in the mineral soil layers, creating horizontal burrows (45). One function they have been shown to provide is a decrease in the density of root-pathogenic nematodes (48, 49), reducing nematode populations and disease incidence, which can contribute to increased crop yields (50, 51). Anecic species mix the litter and mineral soil via surface cast production (45, 46). In addition, the vertical burrows created by anecic species increase water infiltration into deeper soil layers, increasing water holding capacity (52, 53), and regulating water availability.

Data extraction and harmonisation across global layers

In order to predict earthworm communities across the globe, we required harmonised sets of spatially distributed variables. We collected 15 globally distributed layers that are described as predictors of earthworm distribution (Table S2). For the SoilGrids data (54; <https://soilgrids.org>; modelled global layers of soil properties based on soil profiles and remotely-sensed products), which provides soil properties for different layers within the soil profile, we calculated the weighted average for the values of the top four layers (corresponding to the top 30 cm of the soil profile, which matches the soil depth of the earthworm sampling techniques). For sites missing one or more sampled soil properties, the soil properties associated with the 1km pixel corresponding to the site's geographical coordinates were used in the analyses. For CEC, for all sites, values were taken from SoilGrids.

Where possible, the land cover global layer (ESA CCI-LC 300 m; <https://www.esa-landcover-cci.org/>) was re-categorised to amalgamate similar habitat cover categories matching the ones collected within the dataset (see Table S1). Where not possible, the categories were ignored (i.e., classified as NA) during later steps, as estimates could not be produced for unknown habitat cover categories.

No climate variables were taken from the papers or raw data provided, as there was little consistency in climate variables across the papers. Instead, five global climate layers (climatologies) obtained from the CHELSA climate dataset (55) were used (annual mean temperature, temperature seasonality, temperature annual range, annual precipitation, and precipitation seasonality) and, from other sources, the number of months of snow cover (56), and the aridity index and potential evapotranspiration (PET; 57, 58). The within-year standard deviation of PET (PETSD) was calculated as well. Finally, a globally distributed layer of elevation (59) was also included in the analysis. For all of these layers, the value within the 1 km pixel that matched the site's coordinates was used in the analyses.

For an initial harmonisation across all global layers, it was necessary to aggregate or disaggregate - when appropriate - the spatial resolution of the different layers to match a one-kilometre square grid. A nearest neighbour disaggregation algorithm was applied without changing the pixel values, but changing the pixel resolution using the one-kilometre square resolution from SoilGrids as a reference.

Following the spatial harmonisation, the global layers were matched with the collated dataset based on the geographic coordinates of the sampled sites. In the case of the climate layers, all variables were appended to the dataset. Soil variables were only appended if the sites were missing sampled measures, with all studies lacking at least one soil property.

To help prevent extrapolation, all global layers were truncated to values represented by each subset of data, i.e., the minimum and maximum values used in each of the three community metric models. The exception was the number of months of snow cover, which was truncated at four months, thus any sites or areas of the globe with a greater number of months than four were modelled and predicted (respectively) as four months. This ensured an even spread across the range of values (many sites were within 0-4 months, only 9% of sites were greater than four).

Mixed effects modelling

Earthworm species richness, abundance, and biomass models

Three (generalised) linear mixed effects models were constructed, using lme4 (60), one for each of the site-level community metrics: species richness, total abundance (individuals per m²), and total biomass (grams per m²). Prior to modelling, the full dataset was split into three subsets, based on the response variables (i.e., a dataset containing all sites with a species richness value). Within each dataset, we tested for multicollinearity between the elevation, climate, and soil variables using Variance Inflation Factors (VIFs) and removing the variable with highest VIF in turn until all remaining variables were below the predetermined threshold of 3 (61).

Abundance and biomass were log transformed ($\log(x + 1)$) prior to modelling and were then modelled using a Gaussian error structure. Species richness was not log transformed, but instead modelled with a Poisson error structure. All models had random effects that accounted for variation between each of the different studies, using an intercept only structure. Fixed effects included habitat cover, elevation, soil properties, and climate variables. All continuous variables (i.e., elevation, all soil variables, and most climate variables) were centred and scaled (variables were centred on the mean value and divided by the standard deviation) to aid model fitting and interpretability. Number of months of snow cover was modelled as a categorical variable (and therefore not centred and scaled) to allow for a non-linear relationship. As it is expected that earthworm diversity will peak with some snow cover, due to increased precipitation, and soil protection during freezing months (62), but be restricted by prolonged snow cover (63). This also improved the modelling process, as sites were skewed towards the lower number of months, with not enough data at the higher latitudes to fit a non-linear regression.

For each of the three models, the structure of the fixed effects in the maximal model was the same. Habitat cover and elevation were included as additive effects with no interactions. The other variables were grouped into four themes: 'soil', 'precipitation', 'temperature' and 'water retention' (Table S3). For example, all precipitation variables that remained (i.e., were not removed due to their VIF score) were grouped together. Within the soil and two climate groups, all two-way interactions were considered. The water retention group contained specific two-way interactions between soil structure variables (clay and silt percentage) and climate variables relating to water availability that were present in the two climate themes (annual precipitation, precipitation seasonality, PET, PETS, and aridity). These specific interactions were to account for soil moisture and how quickly moisture might leave the soil.

Each maximal model was then simplified using Akaike information criterion (AIC) values. All interactions were tested first, and removed if AIC values were reduced compared to the more

complicated model. Any main effects that were not involved with interactions were tested, and removed if AIC values were reduced (44, 64) (Table S3).

Ecological group responses

The same modelling approach was used to investigate changes in earthworm ecological groups across the different habitat types. For each site, the diversity, abundance and biomass of the three main ecological groups (epigeic, endogeic, and anecic) and one minor ecological group (epi-endogeic) were calculated, based on the category assigned following species name standardisation. Three (generalised) linear mixed effects models with diversity, abundance and biomass as response variables were constructed as detailed above, with the exception that habitat cover interacted with the ecological group (i.e., the biomass of epigeics, endogeics, and anecics at each site). The model was simplified following details given above.

The community metrics of each ecological group in each habitat cover was then predicted, using the ‘predict’ function in ‘lme4’ (when all other variables were at zero, i.e., the mean). The predicted values for the three main ecological groups (epigeic, endogeic and anecic, which had sufficient underlying data. Epi-endogeics were modelled but did not have enough underlying data for robust predictions) were plotted using the ‘triangle.plot’ function in ‘ade4’ (65). The predicted total biomass, i.e., the total of the predicted biomass of the three main ecological groups, was used to determine size of the points within the triangle plot.

Creating maps of earthworm communities

The global patterns of earthworm communities (species richness, abundance, and biomass) were predicted using each of the three models. The values from the relevant global layers (i.e., those corresponding to the variables that remained in each model following simplification) were used in the ‘predict’ function in the ‘lme4’ package, being predicted based on the coefficients of the final models.

A global layer of predicted values was then presented as maps of local communities of earthworms. Although all global layers had been capped at values represented in the underlying dataset, extrapolation still occurred during the prediction (there were instances where grid cells in multiple layers were at the extreme values, and such combinations were not represented in the underlying data, most evident in the predictions of earthworm biomass, see ‘Interpreting the model validation’). To prevent outliers skewing the visualization of results, the colour of maps were curtailed at the extreme low and high values. Curtailing was based on where the majority of values laid. Thus, values lower or higher than the number marked on the scale are coloured the same but may represent a large range of values.

Variable Importance

In order to determine which themes (soil, elevation, habitat cover, precipitation, temperature, water retention) were the most influential in driving earthworm communities, Variable Importance was performed using random forest models (66, 67).

For each of the three community metrics, random forest models were constructed (67) with all the variables that were present in the final (i.e. simplified) model. Random forest models use multiple regression trees to classify data (67). This method was chosen as these models can handle non-linear data, whilst interactions are not specified but can be learnt from the data (68). Random forest models are an ensemble of individual regression (or classification) trees (66, 67). Each tree is created using around two-thirds of the available data, i.e., “out-of-bag” regression, and the process is repeated until the ‘forest’ is complete (ntree default = 500 trees). At each node

in the tree, the subset of response variables is split using the best predictor variable. Unlike regression trees, where at each node the best predictor is used from all available predictor variables, random forest models use only a random sample of the predictor variables (“Mtry”) to determine the best predictor to split the response variable at each node (66, 67). The default Mtry value was used (number of predictors divided by 3), so in our case of 10 to 12 predictor variables $Mtry = 3$ (biomass model) and 4 (species richness and abundance models). The “out-of-bag” data is then predicted using the average prediction of all trees (67).

In addition, random forest models can be used to assess the importance of individual variables (66). One such measure is the mean decrease in node impurity calculated from the decrease in the residual sum of squares for the variable that was used at the node. The average decrease for each variable is averaged across all the trees to create the node impurity (67). An alternative importance variable is the mean decrease in accuracy. For each tree, when the “out-of-bag” data (~one-third of the data) is being predicted, a single predictor variable is permuted, and the increase in prediction error calculated (67). This mean decrease in accuracy is often considered the best of those available (69), but results between the decrease in node impurity and mean decrease in accuracy often correlate well (70).

For each of the three random forest models, the mean decrease in node impurity and mean decrease in accuracy was calculated (using the ‘importance’ function in ‘randomForest’) for each predictor variable in the random forest. In order to determine which theme of variables (habitat cover, elevation, soil, temperature, precipitation, and water retention) was most important in driving patterns in earthworm communities, the mean decrease in node impurity was averaged for all variables within each theme and weighted by the number of times each variable was used in the random forest compared to the other variables in the same theme.

Model validation and sensitivity analysis

A number of additional analyses were performed to determine the robustness of the models and the ability to predict new values. Firstly, the influence of combining both measured soil properties and soil properties from SoilGrids was tested. For the three response variables (species richness, abundance, and biomass) models were created that only included data from SoilGrids. The same modelling process was used as described above (using VIFs to determine starting variables, then simplification). Once the final model had been identified for each of the three community metrics, 10-fold cross-validation was performed (71).

Cross-validation was performed in two ways. Firstly by randomly splitting the dataset underlying each of the three models into 10 nearly-equal size groups. Using the model structures produced following simplification, the model was built using 9 of the groups of data. The 10th group of data was predicted from the re-built model. The predicted data was plotted against the observed data. This process was repeated until all 10 groups of data had been predicted. This process was done for the models that contained only SoilGrids data, and the main models (that used a mixture of soil property data, Fig. S3). Secondly, by splitting the dataset into 10 near-equal sized groups based on study. Thus, the site-level community metrics for a 10th of the studies were predicted by the remaining 9/10 of the data. This process was only performed on the main models (Fig. S4).

For the site-level cross-validated models, the mean squared error (MSE) was calculated from the results of the cross-validation. MSE measures the ability of the model to predict new data, and the result are easily interpretable as they are on the same scale as the original data. MSE was calculated for the total of all models, but as the models may be better at predicting

certain ranges of values, MSE was also calculated for the tertiles of the observed data (i.e., the ability of the model to predict the low, medium, and high values of earthworm communities).

Secondly, the R^2 values of all models (the main models, and the models with only SoilGrids data) were calculated using the MuMIn package (72). The R^2 values describe the fit of the model to the data. The R^2_{marginal} is the variance explained by the fixed effects, whilst the $R^2_{\text{conditional}}$ is the variance explained by the fixed and random effects.

To determine the confidence of the globally predicted values, we followed methods of (9). For each of the biodiversity models, we investigated how well the underlying data represented the full multivariate environmental covariate space of the global layers. We performed a Principal Components based approach on each of the datasets. The centering values, scaling values, and eigenvectors were then used to transform all global layers into the same PCA spaces. Then, we created convex hulls for each of the bivariate combinations from the first 6 (total biomass data) and 7 (richness and total abundance data) i.e., half of the number of variables within the model, principal components to cover more than 90% of the sample space variation. Using the coordinates of these convex hulls, we classified whether each pixel of each global layer falls within or outside each of these convex hulls. Therefore, if all global layers within a pixel were within the convex hull, the interpolation percentage would be 100%, while if only half of the layers were within the convex hull, the interpolation percentage would be 50%. This analysis was performed in Google Earth Engine (73).

Regional latitudinal diversity gradients

In order to ascertain whether there is a species diversity gradient with latitude, the site-level diversity data underlying the species richness model (i.e., contained sites with species level or morphospecies identification) was used. The sites were split into latitude zones that contained roughly equal numbers of sites. Sites were assigned to a zone based on their latitude, with the intention that each zone would contain close to 250 sites. However, all sites with the same coordinates were kept within the same band, so the number of sites within a zone did vary (min = 209, max = 341, mean = 267.6). The number of unique species, based on species binomials, across all sites within each zone was calculated. Within each zone, it was also assumed that each uniquely named morphospecies was different from any of the named species (number of morphospecies across zones, min = 0, max = 21, mean = 3.05). Some of the sites also contained genus-level only identification. When this was the case, a genus was included as one additional species if the genus was unique within the zone (i.e., no named species belonged to that genus). As the number of morphospecies was biased with latitude (i.e., greater taxonomic expertise in the temperate regions, Table S4), the analysis was repeated excluding morphospecies (Fig. S2). The two methods resulted in similar patterns, but reduced richness in some of the zones in the tropics.

All statistics, data manipulation and processing of global data layers was implemented in R [version 3.3.1; (74)].

Supplementary Text

Interpreting the model validation

The results of the biomass model highlighted an issue with the modelling technique used. All of the global data layers were cut at values represented by the underlying datasets. However, during the prediction, it was often the case that multiple data layers were at the extreme ends of the possible range of values. This led to, especially in the case of the biomass model, and to

some extent the abundance model, unrealistically high values being predicted. This issue could only be fixed with additional data, but does not affect the visual maps produced in this study. For the global predictions of biomass, values greater than 2 kg per m² were deemed to be unrealistic. This threshold is over 4 times the maximum recorded biomass of earthworm communities (75), and thus is highly unlikely to be realistic. 98.9% of pixels were less than 400 g per m² [the maximum recorded earthworm community biomass recorded in the temperate region (75)].

Overall, the models had reasonably good fit to the data, assessed using the R² values (Table 1C). However, the predictive power of the models was variable. With all models, the total MSE (Table 1A) increased mainly due to the ill-fitting of the sites with higher values. It is unclear why high values cannot be fitted well with the models; however, it is highly likely that increasing the number of sites would help either identify the issue or improve model fit.

For the majority of the datasets (182 out of 228 studies), the models contained the measured soil properties for some of the variables. Where this was missing, we used the SoilGrids data. Models which contained only SoilGrids data had a better fit to the data (Table 1C) and were typically better at predicting during cross-validation (lower MSE values; Table 1B). However, in most instances, the change in MSE was negligible between the different types of models (Table 1A versus B). Despite the models that contained only SoilGrids data performing slightly better in terms of R² and MSE, there are other reasons why using a mixture of the measured variables and the SoilGrids variables is the best option in the modelling process. Firstly, modelled global estimates of the soil properties may not accurately depict site-level conditions (76), which could result in the variables appearing less important than they would be if they matched the measured communities. Secondly, some of the coordinates within a study were identical which would result in identical SoilGrids data (for these datasets, often small-scale field experiments, the measured soil properties variables were not identical). Using only SoilGrids data would reduce the gradient of soil properties within each study, reducing the number of available gradient comparisons across all datasets. And given that a number of studies (106 out of 228 studies) had identical climate variables across all sites, having variety in all other variables prevented this being an issue within the modelling framework. We call on soil ecologists to collect data on soil properties when they measure diversity of soil taxa, as this permits more robust modelling at both the small scale, and across larger scales.

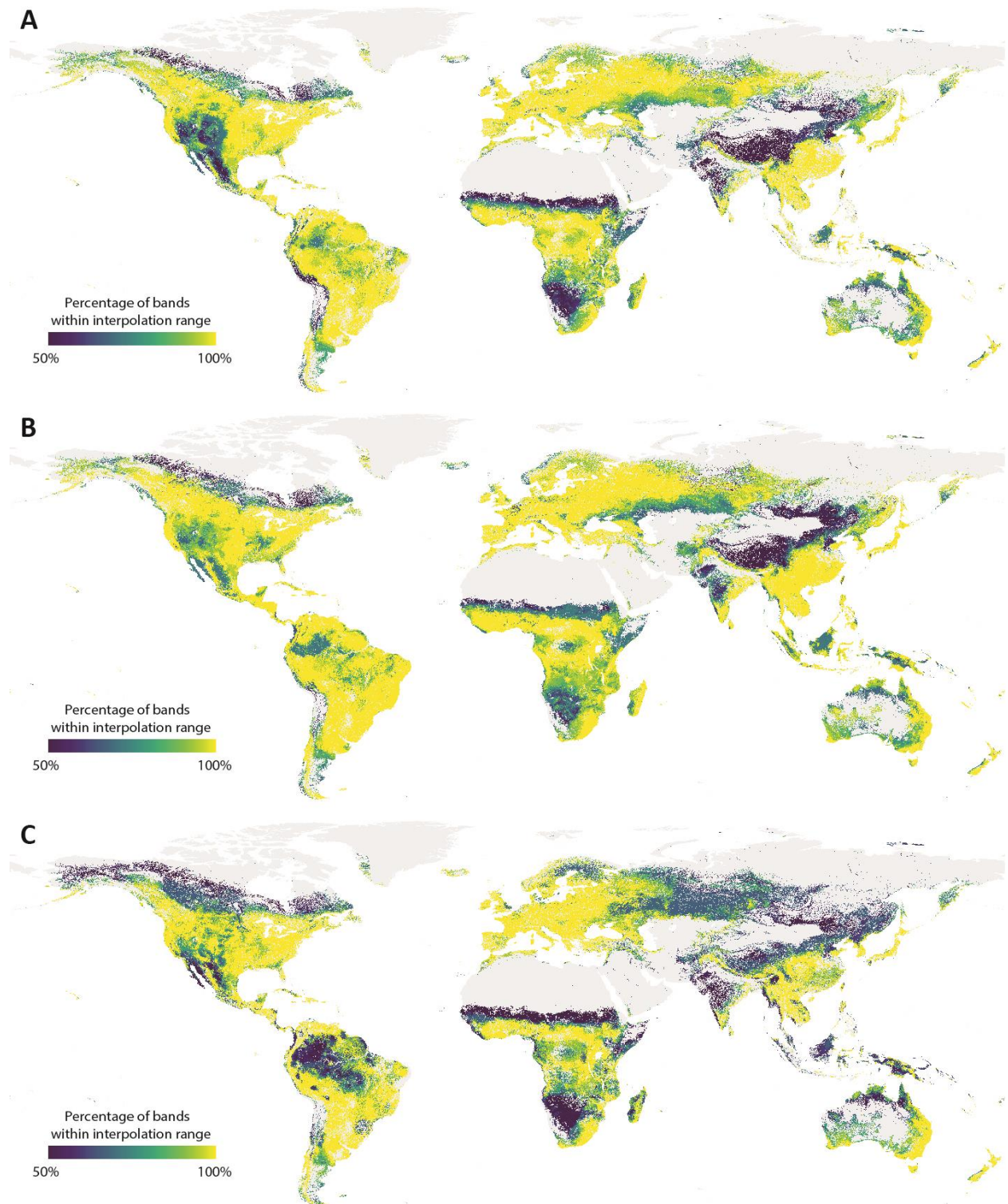
Regardless of whether the model contained measured soil properties or only SoilGrids data, the models were consistently worse at predicting when observed values were high (Table 1). This is likely due to the small number of studies where sampled values were high. Only 5 studies had more than 10 species of earthworms in at least one site, and only 6 studies had more than 300 grams per m² of earthworm biomass in at least one site. There were a greater number of studies that contained high abundance of earthworms, with 34 studies having at least one site that contained more than 600 individuals per m². Increasing the number of studies and sites would help identify whether this, or another cause, is responsible. Ideally, this would improve the predictive power of the models. It is hoped that efforts will continue to collate earthworm diversity data from across the globe.

When cross-validation was performed at the study level (Fig. S4) the predictions were not scattered around the 1:1 line. However, this is to be expected, as when sites are randomly selected and predicted, the study level random-effect is most likely still present in the model. This ensures that the community metrics of each sites can be predicted using the variance from the study it is within. When an entire study is removed, and so no random-effect level exists for

it in the model, all study-level random effects are averaged in order to produce the prediction. Thus, the prediction error is increased, and more concentrated around the overall mean.

For the species richness (Fig. S1A) and total abundance data (Fig. S1B), the interpolation percentage across the globe was relatively high (i.e., the underlying datasets adequately captured the majority of the multivariate environmental conditions). Regions surrounding the Eurasian Steppe, and the Himalayas were some of the most extrapolated regions, with arid regions in Africa and boreal regions also having lower interpolation percentages. For the total biomass data, more regions of the globe had low interpolation percentages (Fig. S1C). These low-value regions were spread across the tropics, particularly Brazil and Indonesia, and large parts of Africa, the sub-tropics, such as India, and temperate regions, including northern China and Russia. Overall, we would expect the globally predicted values of the biomass model to be more extrapolated, than the diversity and total abundance models.

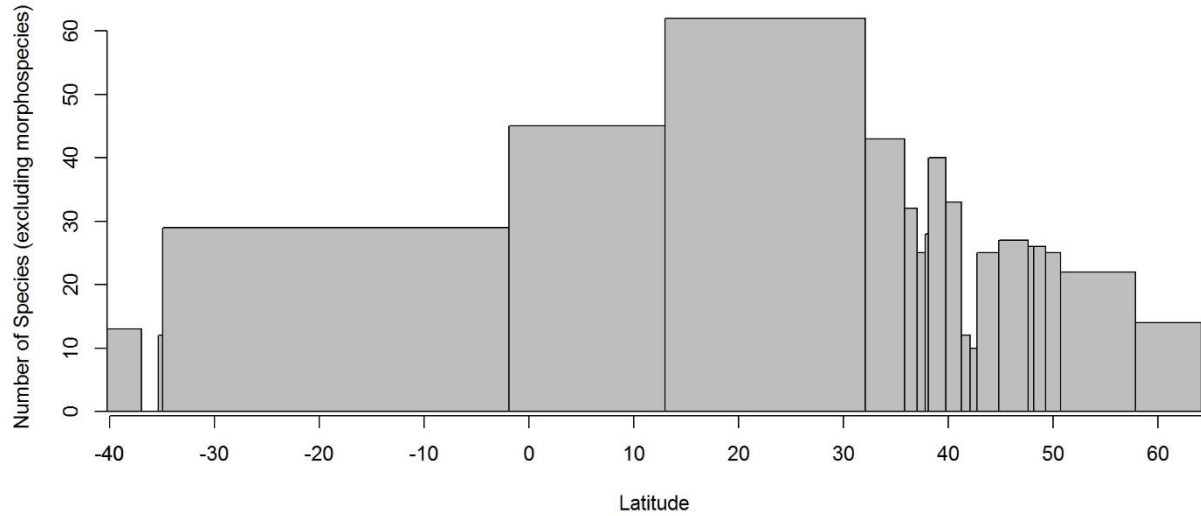
Fig. S1.



Assessment of global extrapolation and interpolation for the (A) species richness data, (B) total abundance data, and (C) total biomass data. Scale shows the percentage of pixels (from each of

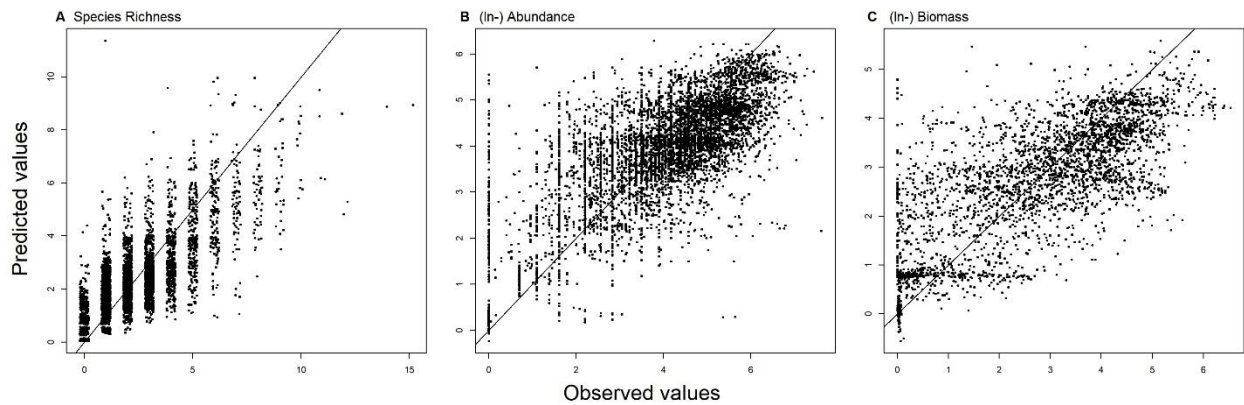
the global layers) falling within the convex hull spaces of the first 6 (biomass) and 7 (richness, abundance) Principal Components collectively explaining >90% of the variation. Low interpolation percentage values (in blue) indicate that few global layers were represented by data, thus extrapolation would have occurred during prediction, whilst high interpolation percentages (in yellow) indicated that many or all global layers were represented by data, thus interpolation would have occurred during prediction.

Fig. S2.



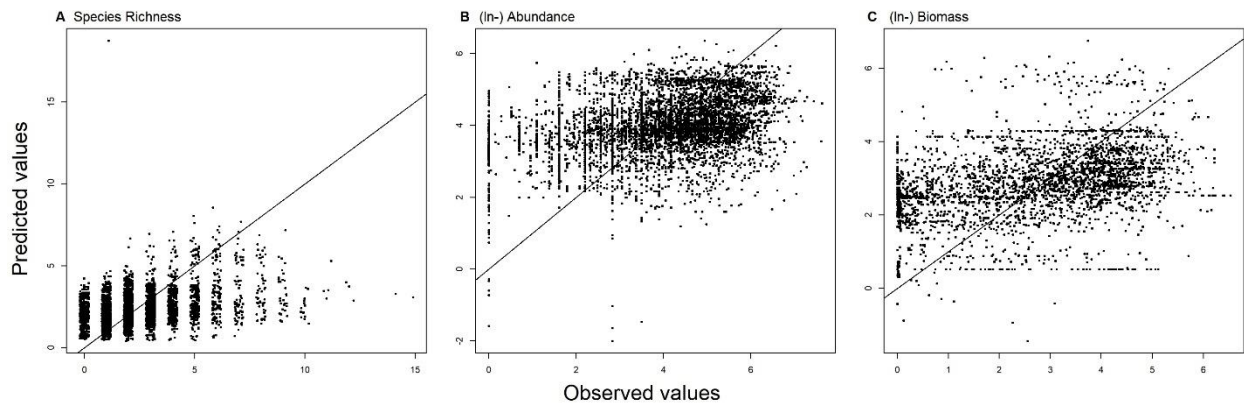
The number of unique species within each latitudinal zone, when the number of sites within each zone was kept relatively equal. The height of the bar indicates the number of unique species across all sites. The width of the bar shows the latitude range the sites cover. Within each zone only the species with binomials, or genera with no other identified species, were included in the calculations (morphospecies were excluded).

Fig. S3.



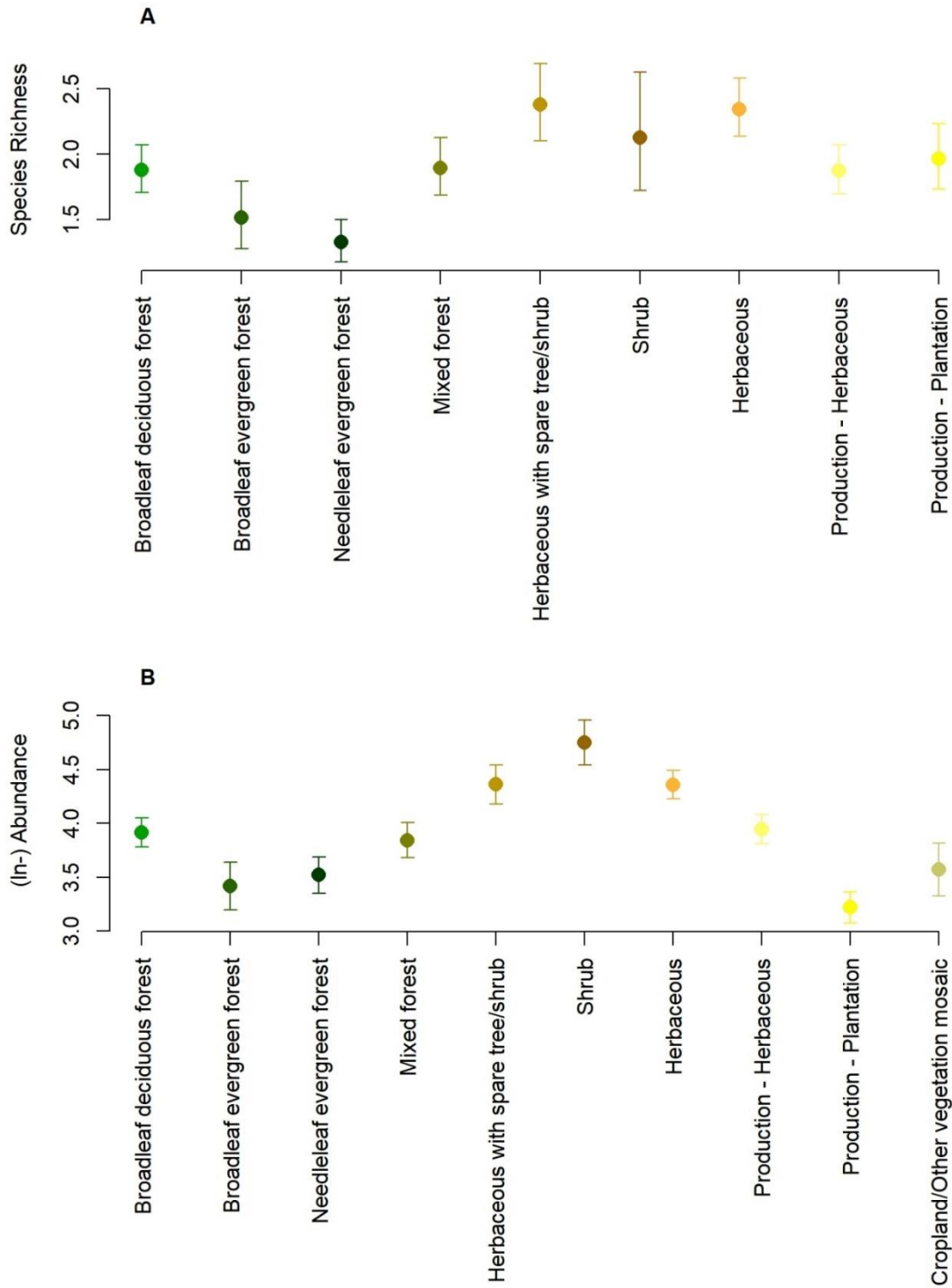
10-fold cross validation of the three main community metric models, (A) species richness, (B) ln-abundance, and (C) ln-biomass. X-axis shows the observed value, and Y-axis the predicted value, black line is the 1:1 line. The underlying dataset of each model was randomly split into 10 nearly-equal size groups. Using the model structures produced following simplification, the model was built using 9 of the groups of data. The 10th group of data was predicted from the rebuilt model. This process was repeated until all 10 groups of data had been predicted. The predicted data was plotted against the observed data.

Fig. S4.



10-fold cross validation of the three main community metric models, (A) species richness, (B) ln-abundance, and (C) ln-biomass. X-axis shows the observed value, and Y-axis the predicted value, black line is the 1:1 line. The underlying dataset of each model was randomly split into 10 nearly-equal size groups, so that each group contained all the data of a tenth of the studies. Using the model structures produced following simplification, the model was built using 9 of the groups of studies. The 10th group of studies was predicted from the re-built model. This process was repeated until all 10 groups of studies had been predicted. The predicted data was plotted against the observed data.

Fig. S5.

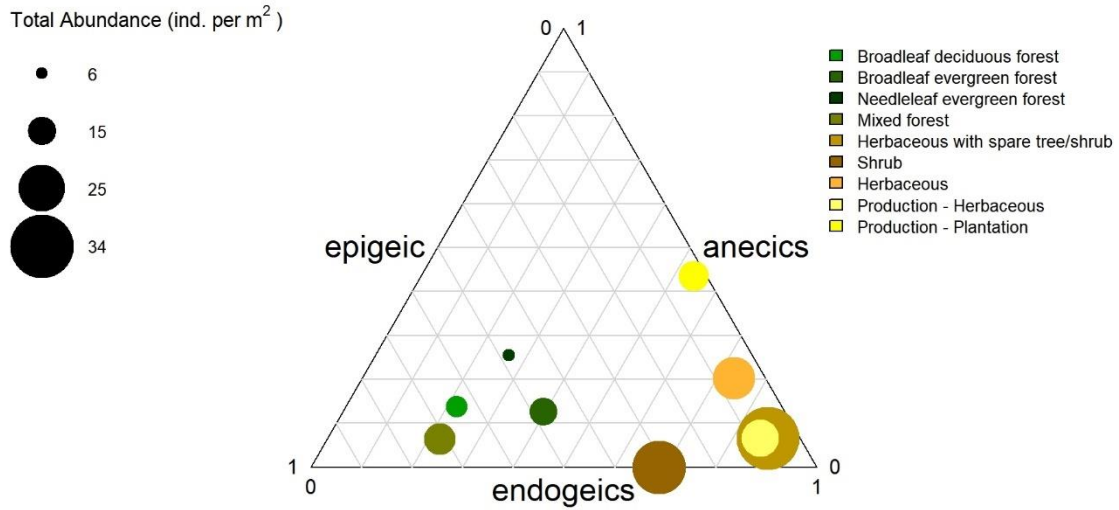


Changes in (A) species richness and (B) ln-abundance across the different habitat cover categories (+/- SD). Values of species richness and abundance are predicted from the main

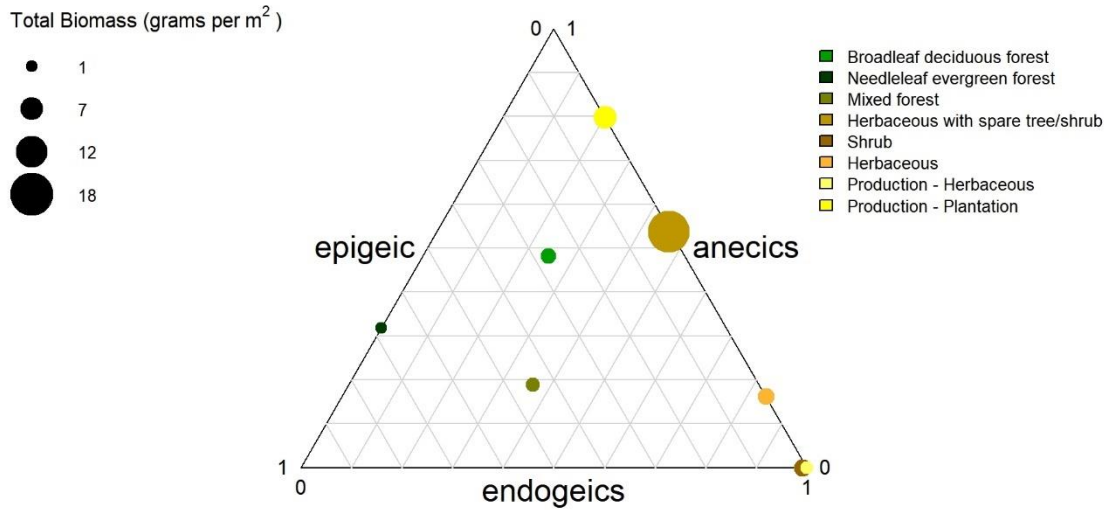
models when all other variables are at zero, i.e., the mean. Not all habitat cover categories had sampled estimates (i.e., species richness could not be estimated for ‘Cropland/Other vegetation mosaic’).

Fig. S6.

A



B



The (A) total abundance and (B) total biomass of the three ecological groups (epigeic, endogeic and anecic earthworms) within each habitat cover category based on modelled estimates. Circle size is relative to the total biomass predicted for the habitat cover, and circle colour indicates the habitat cover. Position within the three axes indicates the proportion of each of the three ecological groups within the community, based on the interaction term between habitat cover and ecological group. During simplification, the interaction term between habitat cover and ecological group was removed in the species richness model, thus those results are not shown.

Not all habitat cover categories had sampled estimates (i.e., biomass could not be estimated for 'Broadleaf evergreen forest' or 'Cropland/Other vegetation mosaic'). This figure shows, for example, that "Broadleaf deciduous forests" have a rather even predicted biomass distribution across the three ecological groups (but low total biomass), while "Production sites" ("Plantation" and "Herbaceous") have high total earthworm abundance, but are dominated by endogeic species.

Table S1.

Original habitat cover variable	Reclassified habitat cover
No Data	NA
Cropland, rainfed	Production - Herbaceous
Cropland - herbaceous cover	Production - Herbaceous
Cropland - Tree or shrub cover	Production - Plantation
Cropland, irrigated or post-flooding	NA
Mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%)	Cropland/Other vegetation mosaic
Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%)	Cropland/Other vegetation mosaic
Tree cover, broadleaved, evergreen, closed to open (>15%)	Broadleaf evergreen forest
Tree cover, broadleaved, deciduous, closed (>40%)	Broadleaf deciduous forest
Tree cover, broadleaved, deciduous, open (15-40%)	Broadleaf deciduous forest
Tree cover, needleleaved, evergreen, closed (>40%)	Needleleaf evergreen forest
Tree cover, needleleaved, evergreen, open (15-40%)	Needleleaf evergreen forest
Tree cover, needleleaved, deciduous, closed (>40%)	Needleleaf deciduous forest
Tree cover, needleleaved, deciduous, open (15-40%)	Needleleaf deciduous forest
Tree cover, mixed leaf type (broadleaved and needleleaved)	Mixed forest
Mosaic tree and shrub (>50%) / herbaceous cover (<50%)	Tree open
Mosaic herbaceous cover (>50%) / tree and shrub (<50%)	Herbaceous with spare tree/shrub
Shrubland	Shrub
Grassland	Herbaceous
Lichens and mosses	NA
Sparse vegetation (tree, shrub, herbaceous cover) (<15%)	Sparse vegetation
Tree cover, flooded, fresh or brackish water	NA
Tree cover, flooded, saline water	NA
Shrub or herbaceous cover, flooded, fresh/saline/brackish water	NA
Urban areas	Urban
Bare areas - consolidated	Bare area (consolidated)
Bare areas - unconsolidated	Bare area (unconsolidated)
Water bodies	Water bodies
Permanent snow and ice	NA

The re-categorisation of the ESA habitat cover variable. Habitat cover at a sampled site was classified based on the 'Reclassified habitat cover' column. As not all categories of habitat were available in the data (i.e., due to too detailed categories, or in habitats typically devoid of sampling), some of the categories of the original habitat cover variable (left-hand column) were reclassified (right-hand column). Usually, this meant that categories were grouped together (i.e., to reduce the categories based on 'openness').

Table S2.

	<i>Variable</i>	<i>Source</i>	<i>Original Spatial Resolution</i>	
	Habitat Cover	ESA CCI-LC	300 m	
	Elevation	(59)	1 km	
Soil Parameters	pH (H2O)		1 km	
	Organic carbon		1 km	
	Soil clay content		1 km	
	Soil silt content		1 km	
	CEC		SoilGrids (54)	1 km
Temperature	Annual Mean Temperature		1 km	
	Temp. seasonality		1 km	
	Temp. annual range		CHELSA (55)	1 km
	PET		(57, 58)	1 km
	PETSD			1 km
Precipitation	Annual precipitation		1 km	
	Precip. seasonality		CHELSA (55)	1 km
	Number of Months with Snow		(56)	1 km
	Aridity Index		(57, 58)	1 km

Information for each of the 15 global layers detailed in the methods. Abbreviations: CEC = Cation exchange capacity, Temp. = Temperature, Precip. = Precipitation, PET = Potential evapotranspiration, PETSD = within year standard deviation of PET

Table S3.

Global Data Layers	Main Effect Only	pH (H2O)	Organic carbon	Soil clay content	Soil silt content	CEC	Annual Mean Temperature	Temp. seasonality	Temp. annual range	PET	PETSD	Annual precipitation	Precip. seasonality	Number of Months with Snow	Acidity Index
Habitat Cover	★ ★ ★ ★ ★														
Elevation	—														
pH (H2O)	★ ★ ★ ★ ★	↑ ↓	↑ ↓	↑ ↓	↑ ↓	↑ ↓									
Organic carbon	★ ★ ★ ★ ★	↑ ↓	↑ ↓	↑ ↓	↑ ↓	↑ ↓				↓ ↓			↓ ↓		
Soil clay content	★ ★ ★ ★ ★	↑ ↓	↑ ↓	↑ ↓	↑ ↓	↑ ↓				↓ ↓			↓ ↓		↑
Soil silt content	★ ★ ★ ★ ★	↑ ↓	↑ ↓	↑ ↓	↑ ↓	↑ ↓				↓ ↓			↓ ↓		↑
CEC	★ ★ ★ ★ ★	↑ ↓	↑ ↓	↑ ↓	↑ ↓	↑ ↓				↓ ↓			↓ ↓		↑
Annual Mean Temperature															
Temp. seasonality															
Temp. annual range	—														
PET	—														
PETSD															
Annual precipitation	—														
Precip. seasonality	★ ★ ★ ★ ★	↑ ↓	↑ ↓	↑ ↓	↑ ↓	↑ ↓									
Number of Months with Snow	★ ★ ★ ★ ★	↑ ↓	↑ ↓	↑ ↓	↑ ↓	↑ ↓									
Acidity Index	★ ★ ★ ★ ★	↑ ↓	↑ ↓	↑ ↓	↑ ↓	↑ ↓									

Results following model simplification of the three community metric models. ‘Main Effect Only’ column shows the slope for the main effect of each variable in the final species richness (turquoise), total abundance (green) and total biomass (yellow) models. ‘+’ indicates the slope was positive, ‘-’ indicates a negative slope, and ‘*’ indicates that the variable was categorical (with intercepts and slopes depending on the category). The remaining columns show the interactions between the variables. An upwards arrow indicates that the slope of one variable would become more positive as the other variable is increased. A downwards arrows indicates that the slope of the one variable would become more negative as the other variable is increased. However, it may not necessarily indicate that the slope changes direction. Black symbols indicate that the coefficient was significant ($p < 0.05$) within the model, and grey/hatched symbols indicate they were not significant [NB. P-values are for illustrative purposes only, as models were simplified based on AIC values]. Habitat cover and elevation were only in the models as main effects. Also noted is the variable theme in which the variable was grouped. Variables that interacted within the ‘water retention’ theme are not shown explicitly, but can be deduced based on interactions between a climate variable and soil property variable. Abbreviations: CEC = Cation exchange capacity, Temp. = Temperature, Precip. = Precipitation, PET = Potential evapotranspiration, PETSD = within year standard deviation of PET.

Table S4.

<i>Latitude</i>	<i>Number of sites</i>	<i>Number of named species</i>	<i>Mean Latitude Range</i>	<i>Number of Morphospecies</i>	<i>% Native</i>	<i>% Non-native</i>	<i>% Unknown</i>
(65,70]	55	11	63.66	0	21.02	57.32	21.66
(60,65]	255	14	56.91	1	0	0	100
(55,60]	157	18	55.11	0	1.57	0	98.43
(50,55]	960	35	37.47	3	11.74	2.85	85.41
(45,50]	1136	38	32.53	1	1	54.42	44.58
(40,45]	1080	54	29.63	2	7.5	13.7	78.8
(35,40]	308	47	34.49	1	6.05	7.74	86.22
(30,35]	113	18	28.15	3	0	0	100
(25,30]	47	18	56.81	0	0	0	100
(20,25]	9	12	10.78	3	22.86	0	77.14
(15,20]	30	11	6.17	2	16.87	45.78	37.35
(10,15]	26	4	14.18	3	0	0	100
(5,10]	40	27	5.53	3	27.11	19.88	53.01
(0,5]	127	10	14.08	12	0	2.17	97.83
(-5,0]	146	10	14.21	14	0	0	100
(-10,-5]	22	1	46.08	6	0	0	100
(-15,-10]	5	0	NA	1	0	0	100
(-20,-15]	0	NA	NA	NA	NA	NA	NA
(-25,-20]	5	8	14.9	4	40	43.33	16.67
(-30,-25]	0	NA	NA	NA	NA	NA	NA
(-35,-30]	150	12	86.6	1	39.66	37.93	22.41
(-40,-35]	679	16	62.45	0	24.93	74.45	0.62
(-45,-40]	3	4	105.78	0	0	0	100

Details of the number of sites, and composition of the earthworm species for each 5 degree latitudinal band. Based on the geographical coordinates, each site was classified into 5 degree latitudinal bands. The number of species is based on the binomial names given by the original data collector, then revised for consistency. For each named species in each band the latitudinal range (the difference between the minimum latitude and the maximum latitude, based on all sites within the dataset that the species occurred) was calculated, and the average taken from all species within the band. Morphospecies are individuals that were identified to genus level, and identified by the data collectors as morphologically distinct from other (morpho-) species, but were not identified to species-level. The percentage of native and non-native species is based on information provided by the original data collectors, and is therefore often incomplete (depicted in the ‘% unknown’ column).

References and Notes

1. R. D. Bardgett, W. H. van der Putten, Belowground biodiversity and ecosystem functioning. *Nature* **515**, 505–511 (2014). [doi:10.1038/nature13855](https://doi.org/10.1038/nature13855) [Medline](#)
2. N. Eisenhauer, P. M. Antunes, A. E. Bennett, K. Birkhofer, A. Bissett, M. A. Bowker, T. Caruso, B. Chen, D. C. Coleman, W. de Boer, P. de Ruiter, T. H. DeLuca, F. Frati, B. S. Griffiths, M. M. Hart, S. Hättenschwiler, J. Haimi, M. Heethoff, N. Kaneko, L. C. Kelly, H. P. Leinaas, Z. Lindo, C. Macdonald, M. C. Rillig, L. Ruess, S. Scheu, O. Schmidt, T. R. Seastedt, N. M. van Straalen, A. V. Tiunov, M. Zimmer, J. R. Powell, Priorities for research in soil ecology. *Pedobiologia* **63**, 1–7 (2017). [doi:10.1016/j.pedobi.2017.05.003](https://doi.org/10.1016/j.pedobi.2017.05.003) [Medline](#)
3. L. Tedersoo, M. Bahram, S. Pölme, U. Kõljalg, N. S. Yorou, R. Wijesundera, L. Villarreal Ruiz, A. M. Vasco-Palacios, P. Q. Thu, A. Suija, M. E. Smith, C. Sharp, E. Saluveer, A. Saitta, M. Rosas, T. Riit, D. Ratkowsky, K. Pritsch, K. Põldmaa, M. Piepenbring, C. Phosri, M. Peterson, K. Parts, K. Pärtel, E. Otsing, E. Nouhra, A. L. Njouonkou, R. H. Nilsson, L. N. Morgado, J. Mayor, T. W. May, L. Majuakim, D. J. Lodge, S. S. Lee, K.-H. Larsson, P. Kohout, K. Hosaka, I. Hiiesalu, T. W. Henkel, H. Harend, L. D. Guo, A. Greslebin, G. Grelet, J. Geml, G. Gates, W. Dunstan, C. Dunk, R. Drenkhan, J. Dearnaley, A. De Kesel, T. Dang, X. Chen, F. Buegger, F. Q. Brearley, G. Bonito, S. Anslan, S. Abell, K. Abarenkov, Global diversity and geography of soil fungi. *Science* **346**, 1256688 (2014). [doi:10.1126/science.1256688](https://doi.org/10.1126/science.1256688) [Medline](#)
4. M. Delgado-Baquerizo, A. M. Oliverio, T. E. Brewer, A. Benavent-González, D. J. Eldridge, R. D. Bardgett, F. T. Maestre, B. K. Singh, N. Fierer, A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018). [doi:10.1126/science.aap9516](https://doi.org/10.1126/science.aap9516) [Medline](#)
5. M. Bahram, F. Hildebrand, S. K. Forslund, J. L. Anderson, N. A. Soudzilovskaia, P. M. Bodegom, J. Bengtsson-Palme, S. Anslan, L. P. Coelho, H. Harend, J. Huerta-Cepas, M. H. Medema, M. R. Maltz, S. Mundra, P. A. Olsson, M. Pent, S. Pölme, S. Sunagawa, M. Ryberg, L. Tedersoo, P. Bork, Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018). [doi:10.1038/s41586-018-0386-6](https://doi.org/10.1038/s41586-018-0386-6) [Medline](#)

6. H. Hillebrand, On the generality of the latitudinal diversity gradient. *Am. Nat.* **163**, 192–211 (2004). [doi:10.1086/381004](https://doi.org/10.1086/381004) [Medline](#)
7. E. K. Cameron, I. S. Martins, P. Lavelle, J. Mathieu, L. Tedersoo, M. Bahram, F. Gottschall, C. A. Guerra, J. Hines, G. Patoine, J. Siebert, M. Winter, S. Cesarz, O. Ferlian, H. Kreft, T. E. Lovejoy, L. Montanarella, A. Orgiazzi, H. M. Pereira, H. R. P. Phillips, J. Settele, D. H. Wall, N. Eisenhauer, Global mismatches in aboveground and belowground biodiversity. *Conserv. Biol.* **33**, 1187–1192 (2019). [doi:10.1111/cobi.13311](https://doi.org/10.1111/cobi.13311) [Medline](#)
8. N. Fierer, M. S. Strickland, D. Liptzin, M. A. Bradford, C. C. Cleveland, Global patterns in belowground communities. *Ecol. Lett.* **12**, 1238–1249 (2009). [doi:10.1111/j.1461-0248.2009.01360.x](https://doi.org/10.1111/j.1461-0248.2009.01360.x) [Medline](#)
9. J. van den Hoogen, S. Geisen, D. Routh, H. Ferris, W. Traunspurger, D. A. Wardle, R. G. M. de Goede, B. J. Adams, W. Ahmad, W. S. Andriuzzi, R. D. Bardgett, M. Bonkowski, R. Campos-Herrera, J. E. Cares, T. Caruso, L. de Brito Caixeta, X. Chen, S. R. Costa, R. Creamer, J. Mauro da Cunha Castro, M. Dam, D. Djigal, M. Escuer, B. S. Griffiths, C. Gutiérrez, K. Hohberg, D. Kalinkina, P. Kardol, A. Kergunteuil, G. Korthals, V. Krashevskaya, A. A. Kudrin, Q. Li, W. Liang, M. Magilton, M. Marais, J. A. R. Martín, E. Matveeva, E. H. Mayad, C. Mulder, P. Mullin, R. Neilson, T. A. D. Nguyen, U. N. Nielsen, H. Okada, J. E. P. Rius, K. Pan, V. Peneva, L. Pellissier, J. Carlos Pereira da Silva, C. Pitteloud, T. O. Powers, K. Powers, C. W. Quist, S. Rasmann, S. S. Moreno, S. Scheu, H. Setälä, A. Sushchuk, A. V. Tiunov, J. Trap, W. van der Putten, M. Vestergård, C. Villenave, L. Waeyenbergh, D. H. Wall, R. Wilschut, D. G. Wright, J. I. Yang, T. W. Crowther, Soil nematode abundance and functional group composition at a global scale. *Nature* **572**, 194–198 (2019). [doi:10.1038/s41586-019-1418-6](https://doi.org/10.1038/s41586-019-1418-6) [Medline](#)
10. T. Decaëns, Macroecological patterns in soil communities. *Glob. Ecol. Biogeogr.* **19**, 287–302 (2010). [doi:10.1111/j.1466-8238.2009.00517.x](https://doi.org/10.1111/j.1466-8238.2009.00517.x)
11. C. A. Edwards, Ed., *Earthworm Ecology* (CRC Press, ed. 2, 2004).
12. M. Blouin, M. E. Hodson, E. A. Delgado, G. Baker, L. Brussaard, K. R. Butt, J. Dai, L. Dendooven, G. Peres, J. E. Tondoh, D. Cluzeau, J. J. Brun, A review of earthworm

- impact on soil function and ecosystem services. *Eur. J. Soil Sci.* **64**, 161–182 (2013). [doi:10.1111/ejss.12025](https://doi.org/10.1111/ejss.12025)
13. D. Craven, M. P. Thakur, E. K. Cameron, L. E. Frelich, R. Beauséjour, R. B. Blair, B. Blossey, J. Burtis, A. Choi, A. Dávalos, T. J. Fahey, N. A. Fisichelli, K. Gibson, I. T. Handa, K. Hopfensperger, S. R. Loss, V. Nuzzo, J. C. Maerz, T. Sackett, B. C. Scharenbroch, S. M. Smith, M. Vellend, L. G. Umek, N. Eisenhauer, The unseen invaders: Introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis). *Glob. Change Biol.* **23**, 1065–1074 (2017). [doi:10.1111/gcb.13446](https://doi.org/10.1111/gcb.13446)
14. See supplementary materials.
15. M. Rutgers, A. Orgiazzi, C. Gardi, J. Römcke, S. Jänsch, A. M. Keith, R. Neilson, B. Boag, O. Schmidt, A. K. Murchie, R. P. Blackshaw, G. Pérès, D. Cluzeau, M. Guernion, M. J. I. Briones, J. Rodeiro, R. Piñeiro, D. J. Diaz Cosín, J. P. Sousa, M. Suhadolc, I. Kos, P. H. Krogh, J. H. Faber, C. Mulder, J. J. Bogte, H. J. van Wijnen, A. J. Schouten, D. de Zwart, Mapping earthworm communities in Europe. *Appl. Soil Ecol.* **97**, 98–111 (2016). [doi:10.1016/j.apsoil.2015.08.015](https://doi.org/10.1016/j.apsoil.2015.08.015)
16. P. F. Hendrix, P. J. Bohlen, Exotic earthworm invasions in North America: Ecological and policy implications. *Bioscience* **52**, 801–811 (2002). [doi:10.1641/0006-3568\(2002\)052\[0801:EEIINA\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2002)052[0801:EEIINA]2.0.CO;2)
17. T. G. Pearce, The calcium relations of selected lumbricidae. *J. Anim. Ecol.* **41**, 167 (1972). [doi:10.2307/3511](https://doi.org/10.2307/3511)
18. D. J. Spurgeon, A. M. Keith, O. Schmidt, D. R. Lammertsma, J. H. Faber, Land-use and land-management change: Relationships with earthworm and fungi communities and soil structural properties. *BMC Ecol.* **13**, 46 (2013). [doi:10.1186/1472-6785-13-46](https://doi.org/10.1186/1472-6785-13-46) [Medline](#)
19. J. Mathieu, T. J. Davies, Glaciation as an historical filter of below-ground biodiversity. *J. Biogeogr.* **41**, 1204–1214 (2014). [doi:10.1111/jbi.12284](https://doi.org/10.1111/jbi.12284)
20. P. Lavelle, C. Lattaud, D. Trigo, I. Barois, Mutualism and biodiversity in soils. *Plant Soil* **170**, 23–33 (1995). [doi:10.1007/BF02183052](https://doi.org/10.1007/BF02183052)

21. R. R. Dunn, D. Agosti, A. N. Andersen, X. Arnan, C. A. Bruhl, X. Cerdá, A. M. Ellison, B. L. Fisher, M. C. Fitzpatrick, H. Gibb, N. J. Gotelli, A. D. Gove, B. Guenard, M. Janda, M. Kaspari, E. J. Laurent, J.-P. Lessard, J. T. Longino, J. D. Majer, S. B. Menke, T. P. McGlynn, C. L. Parr, S. M. Philpott, M. Pfeiffer, J. Retana, A. V. Suarez, H. L. Vasconcelos, M. D. Weiser, N. J. Sanders, Climatic drivers of hemispheric asymmetry in global patterns of ant species richness. *Ecol. Lett.* **12**, 324–333 (2009). [doi:10.1111/j.1461-0248.2009.01291.x](https://doi.org/10.1111/j.1461-0248.2009.01291.x) [Medline](#)
22. H. Kreft, W. Jetz, Global patterns and determinants of vascular plant diversity. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 5925–5930 (2007). [doi:10.1073/pnas.0608361104](https://doi.org/10.1073/pnas.0608361104) [Medline](#)
23. C. Fragoso, P. Lavelle, Earthworm communities of tropical rain forests. *Soil Biol. Biochem.* **24**, 1397–1408 (1992). [doi:10.1016/0038-0717\(92\)90124-G](https://doi.org/10.1016/0038-0717(92)90124-G)
24. K. J. Gaston, T. M. Blackburn, *Pattern and Process in Macroecology* (Blackwell, 2000).
25. J. Davison, M. Moora, M. Öpik, A. Adholeya, L. Ainsaar, A. Bâ, S. Burla, A. G. Diedhiou, I. Hiiesalu, T. Jairus, N. C. Johnson, A. Kane, K. Koorem, M. Kochar, C. Ndiaye, M. Pärtel, Ü. Reier, Ü. Saks, R. Singh, M. Vasar, M. Zobel, Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science* **349**, 970–973 (2015). [doi:10.1126/science.aab1161](https://doi.org/10.1126/science.aab1161) [Medline](#)
26. M. Maraun, H. Schatz, S. Scheu, Awesome or ordinary? Global diversity patterns of oribatid mites. *Ecography* **30**, 209–216 (2007). [doi:10.1111/j.0906-7590.2007.04994.x](https://doi.org/10.1111/j.0906-7590.2007.04994.x)
27. T. Decaëns, D. Porco, S. W. James, G. G. Brown, V. Chassany, F. Dubs, L. Dupont, E. Lapiéd, R. Rougerie, J. Rossi, V. Roy, DNA barcoding reveals diversity patterns of earthworm communities in remote tropical forests of French Guiana. *Soil Biol. Biochem.* **92**, 171–183 (2016). [doi:10.1016/j.soilbio.2015.10.009](https://doi.org/10.1016/j.soilbio.2015.10.009)
28. D. C. Coleman, D. A. Crossley, P. F. Hendrix, *Fundamentals of Soil Ecology* (Elsevier, ed. 2, 2004), pp. 271–298.
29. J. W. Spaak, J. M. Baert, D. J. Baird, N. Eisenhauer, L. Maltby, F. Pomati, V. Radchuk, J. R. Rohr, P. J. Van den Brink, F. De Laender, Shifts of community composition and population density substantially affect ecosystem function despite invariant richness. *Ecol. Lett.* **20**, 1315–1324 (2017). [doi:10.1111/ele.12828](https://doi.org/10.1111/ele.12828) [Medline](#)

30. N. Eisenhauer, J. Schlaghamerský, P. B. Reich, L. E. Frelich, The wave towards a new steady state: Effects of earthworm invasion on soil microbial functions. *Biol. Invasions* **13**, 2191–2196 (2011). [doi:10.1007/s10530-011-0053-4](https://doi.org/10.1007/s10530-011-0053-4)
31. M. A. Drumond, A. Q. Guimarães, H. R. El Bizri, L. C. Giovanetti, D. G. Sepúlveda, R. P. Martins, Life history, distribution and abundance of the giant earthworm *Rhinodrilus alatus* RIGHI 1971: Conservation and management implications. *Braz. J. Biol.* **73**, 699–708 (2013). [doi:10.1590/S1519-69842013000400004](https://doi.org/10.1590/S1519-69842013000400004) [Medline](#)
32. L. Santini, N. J. B. Isaac, L. Maiorano, G. F. Ficetola, M. A. J. Huijbregts, C. Carbone, W. Thuiller, Global drivers of population density in terrestrial vertebrates. *Glob. Ecol. Biogeogr.* **27**, 968–979 (2018). [doi:10.1111/geb.12758](https://doi.org/10.1111/geb.12758)
33. D. Song, K. Pan, A. Tariq, F. Sun, Z. Li, X. Sun, L. Zhang, O. A. Olusanya, X. Wu, Large-scale patterns of distribution and diversity of terrestrial nematodes. *Appl. Soil Ecol.* **114**, 161–169 (2017). [doi:10.1016/j.apsoil.2017.02.013](https://doi.org/10.1016/j.apsoil.2017.02.013)
34. Intergovernmental Panel on Climate Change, *Climate Change 2014 Synthesis Report Summary for Policymakers* (2014); www.ipcc.ch/site/assets/uploads/2018/02/AR5_SYR_FINAL_SPM.pdf.
35. D. K. Hackenberger, B. K. Hackenberger, Earthworm community structure in grassland habitats differentiated by climate type during two consecutive seasons. *Eur. J. Soil Biol.* **61**, 27–34 (2014). [doi:10.1016/j.ejsobi.2014.01.001](https://doi.org/10.1016/j.ejsobi.2014.01.001)
36. M. A. Bradford, G. F. Veen, A. Bonis, E. M. Bradford, A. T. Classen, J. H. C. Cornelissen, T. W. Crowther, J. R. De Long, G. T. Freschet, P. Kardol, M. Manrubia-Freixa, D. S. Maynard, G. S. Newman, R. S. P. Logtestijn, M. Viketoft, D. A. Wardle, W. R. Wieder, S. A. Wood, W. H. van der Putten, A test of the hierarchical model of litter decomposition. *Nat. Ecol. Evol.* **1**, 1836–1845 (2017). [doi:10.1038/s41559-017-0367-4](https://doi.org/10.1038/s41559-017-0367-4) [Medline](#)
37. A. Rice, P. Šmarda, M. Novosolov, M. Drori, L. Glick, N. Sabath, S. Meiri, J. Belmaker, I. Mayrose, The global biogeography of polyploid plants. *Nat. Ecol. Evol.* **3**, 265–273 (2019). [doi:10.1038/s41559-018-0787-9](https://doi.org/10.1038/s41559-018-0787-9) [Medline](#)

38. A. Shade, R. R. Dunn, S. A. Blowes, P. Keil, B. J. M. Bohannan, M. Herrmann, K. Küsel, J. T. Lennon, N. J. Sanders, D. Storch, J. Chase, Macroecology to unite all life, large and small. *Trends Ecol. Evol.* **33**, 731–744 (2018). [doi:10.1016/j.tree.2018.08.005](https://doi.org/10.1016/j.tree.2018.08.005) [Medline](#)
39. J. M. Anderson, J. S. I. Ingram, Eds., *Tropical Soil Biology and Fertility: A Handbook of Methods* (Cambridge Univ. Press, ed. 2, 1993), pp. 88–91.
40. ISO, “Soil quality - Sampling of soil invertebrates - Part 1: Hand-sorting and extraction of earthworms” (ISO 23611-1:2018); www.iso.org/standard/70449.html.
41. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona, Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012). [doi:10.1038/nmeth.2019](https://doi.org/10.1038/nmeth.2019) [Medline](#)
42. J. Koricheva, J. Gurevitch, K. Mengersen, Eds., *Handbook of Meta-Analysis in Ecology and Evolution* (Princeton Univ. Press, 2013).
43. M. D. Bartlett, M. J. I. Briones, R. Neilson, O. Schmidt, D. Spurgeon, R. E. Creamer, A critical review of current methods in earthworm ecology: From individuals to populations. *Eur. J. Soil Biol.* **46**, 67–73 (2010). [doi:10.1016/j.ejsobi.2009.11.006](https://doi.org/10.1016/j.ejsobi.2009.11.006)
44. M. J. Crawley, *The R Book* (Wiley, 2012).
45. M. B. Bouché, Strategies lombriciennes. *Ecol. Bull.* **25**, 122–132 (1977).
46. G. G. Brown, How do earthworms affect microfloral and faunal community diversity? *Plant Soil* **170**, 209–231 (1995). [doi:10.1007/BF02183068](https://doi.org/10.1007/BF02183068)
47. J. Seeber, G. U. H. Seeber, R. Langel, S. Scheu, E. Meyer, The effect of macro-invertebrates and plant litter of different quality on the release of N from litter to plant on alpine pastureland. *Biol. Fertil. Soils* **44**, 783–790 (2008). [doi:10.1007/s00374-008-0282-6](https://doi.org/10.1007/s00374-008-0282-6)
48. M. Blouin, Y. Zuily-Fodil, A. T. Pham-Thi, D. Laffray, G. Reversat, A. Pando, J. Tondoh, P. Lavelle, Belowground organism activities affect plant aboveground phenotype, inducing plant tolerance to parasites. *Ecol. Lett.* **8**, 202–208 (2005). [doi:10.1111/j.1461-0248.2004.00711.x](https://doi.org/10.1111/j.1461-0248.2004.00711.x)

49. J. Boyer, G. Reversat, P. Lavelle, A. Chabanne, Interactions between earthworms and plant-parasitic nematodes. *Eur. J. Soil Biol.* **59**, 43–47 (2013). [doi:10.1016/j.ejsobi.2013.10.004](https://doi.org/10.1016/j.ejsobi.2013.10.004)
50. G. Loranger-Merciris, Y.-M. Cabidoche, B. Deloné, P. Quénéhervé, H. Ozier-Lafontaine, How earthworm activities affect banana plant response to nematodes parasitism. *Appl. Soil Ecol.* **52**, 1–8 (2012). [doi:10.1016/j.apsoil.2011.10.003](https://doi.org/10.1016/j.apsoil.2011.10.003)
51. G. G. Brown, C. A. Edwards, L. Brussaard, in *Earthworm Ecology*, C. A. Edwards, Ed. (CRC Press, ed. 2, 2004), pp. 13–49.
52. M. B. Bouché, F. Al-Addan, Earthworms, water infiltration and soil stability: Some new assessments. *Soil Biol. Biochem.* **29**, 441–452 (1997). [doi:10.1016/S0038-0717\(96\)00272-6](https://doi.org/10.1016/S0038-0717(96)00272-6)
53. M. Joschko, H. Diestel, O. Larink, Assessment of earthworm burrowing efficiency in compacted soil with a combination of morphological and soil physical measurements. *Biol. Fertil. Soils* **8**, 191–196 (1989). [doi:10.1007/BF00266478](https://doi.org/10.1007/BF00266478)
54. T. Hengl, J. Mendes de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, B. Kempen, SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE* **12**, e0169748 (2017). [doi:10.1371/journal.pone.0169748](https://doi.org/10.1371/journal.pone.0169748) [Medline](#)
55. D. N. Karger, O. Conrad, J. Böhrner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, H. P. Linder, M. Kessler, Climatologies at high resolution for the Earth's land surface areas. *Sci. Data* **4**, 170122 (2017). [doi:10.1038/sdata.2017.122](https://doi.org/10.1038/sdata.2017.122) [Medline](#)
56. D. K. Hall, G. A. Riggs, MODIS/Terra Snow Cover Monthly L3 Global 0.05Deg CMG, Version 6 (NASA National Snow and Ice Data Center Distributed Active Archive Center, 2015). [doi:10.5067/MODIS/MOD10CM.006](https://doi.org/10.5067/MODIS/MOD10CM.006)
57. R. J. Zomer, A. Trabucco, D. A. Bossio, L. V. Verchot, Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agric. Ecosyst. Environ.* **126**, 67–80 (2008). [doi:10.1016/j.agee.2008.01.014](https://doi.org/10.1016/j.agee.2008.01.014)

58. R. J. Zomer, D. A. Bossio, A. Trabucco, L. Yuanjie, D. C. Gupta, V. P. Singh, Trees and Water: Smallholder Agroforestry on Irrigated Lands in Northern India. *IWMI Res. Rep.* **122**, 45 (2007).
59. J. Danielson, D. Gesch, “Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010)” (2011); <https://pubs.er.usgs.gov/publication/ofr20111073>.
60. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015). [doi:10.18637/jss.v067.i01](https://doi.org/10.18637/jss.v067.i01)
61. A. F. Zuur, E. N. Ieno, C. S. Elphick, A protocol for data exploration to avoid common statistical problems. *Methods Ecol. Evol.* **1**, 3–14 (2010). [doi:10.1111/j.2041-210X.2009.00001.x](https://doi.org/10.1111/j.2041-210X.2009.00001.x)
62. N. Eisenhauer, A. Stefanski, N. A. Fisichelli, K. Rice, R. Rich, P. B. Reich, Warming shifts ‘worming’: Effects of experimental warming on invasive earthworms in northern North America. *Sci. Rep.* **4**, 6890 (2014). [Medline](#)
63. M. Nieminen, E. Ketoja, J. Mikola, J. Terhivuo, T. Siren, V. Nuutinen, Local land use effects and regional environmental limits on earthworm communities in Finnish arable landscapes. *Ecol. Appl.* **21**, 3162–3177 (2011). [doi:10.1890/10-1801.1](https://doi.org/10.1890/10-1801.1)
64. A. F. Zuur, E. N. Ieno, A. A. Saveliev, *Mixed Effects Models and Extensions in Ecology with R* (Springer, 2009).
65. S. Dray, A.-B. Dufour, The ade4 Package: Implementing the Duality Diagram for Ecologists. *J. Stat. Softw.* **22**, 1–20 (2007). [doi:10.18637/jss.v022.i04](https://doi.org/10.18637/jss.v022.i04)
66. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001). [doi:10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)
67. A. Liaw, M. Wiener, Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
68. U. Grömping, Variable importance assessment in regression: Linear regression versus random forest. *Am. Stat.* **63**, 308–319 (2009). [doi:10.1198/tast.2009.08199](https://doi.org/10.1198/tast.2009.08199)
69. C. Strobl, A. L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007). [doi:10.1186/1471-2105-8-25](https://doi.org/10.1186/1471-2105-8-25) [Medline](#)

70. B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F. A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* **10**, 213 (2009). [doi:10.1186/1471-2105-10-213](https://doi.org/10.1186/1471-2105-10-213) [Medline](#)
71. G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning* (Springer, 2013).
72. MuMIn, Multi-Model Inference. R Package version 1.42.1 (2018); <https://CRAN.R-project.org/package=MuMIn>.
73. N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, R. Moore, Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017). [doi:10.1016/j.rse.2017.06.031](https://doi.org/10.1016/j.rse.2017.06.031)
74. R: *A Language and Environment for Statistical Computing* (2016).
75. P. Lavelle, A. V. Spain, *Soil Ecology* (Springer, 2001).
76. P. A. Sanchez, S. Ahamed, F. Carré, A. E. Hartemink, J. Hempel, J. Huising, P. Lagacherie, A. B. McBratney, N. J. McKenzie, Mde. L. Mendonça-Santos, B. Minasny, L. Montanarella, P. Okoth, C. A. Palm, J. D. Sachs, K. D. Shepherd, T.-G. Vågen, B. Vanlauwe, M. G. Walsh, L. A. Winowiecki, G.-L. Zhang, Digital soil map of the world. *Science* **325**, 680–681 (2009). [doi:10.1126/science.1175084](https://doi.org/10.1126/science.1175084) [Medline](#)