Supplemental Information

Melissa Chapman[1,2], Martin Jung[2], David Leclère[3], Carl Boettiger[4], Andrey L. D.Augustynczik[3], Mykola Gusti[3], Leopold Ringwald[3], and Piero Visconti[2]

[1]National Center for Ecological Analysis and Synthesis, Santa Barbara, CA, USA
[2]Biodiversity, Ecology and Conservation Research Group, International Institute for Applied Systems Analysis (IIASA), Vienna,Austria
[3]Integrated Biosphere Futures Research Group, International Institute for Applied Systems Analysis (IIASA), Vienna,Austria
[4]Department of Environmental Science, Policy, and Management, University of California Berkeley, Berkeley, CA, USA

Meeting European conservation and restoration targets under future land-use demands

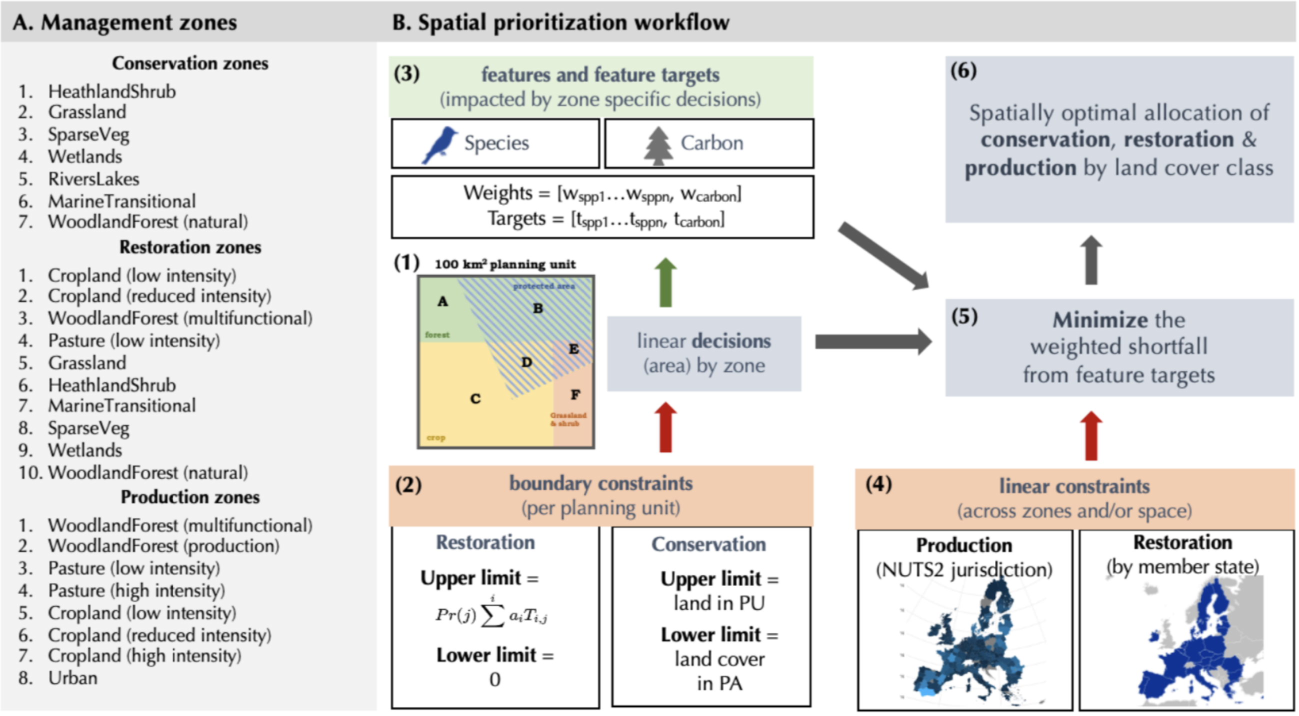
A. Management zones
Conservation zones
1. HeathlandShrub
2. Grassland
3. SparseVeg
4. Wetlands
5. RiversLakes
6. MarineTransitional
7. WoodlandForest (natural)
Restoration zones
1. Cropland (low intensity)
2. Cropland (reduced intensity)
3. WoodlandForest (multifunctional)
4. Pasture (low intensity)
5. Grassland
6. HeathlandShrub
7. MarineTransitional
8. SparseVeg
9. Wetlands
10. WoodlandForest (natural)
Production zones
1. WoodlandForest (multifunctional)
2. WoodlandForest (production)
3. Pasture (low intensity)
4. Pasture (high intensity)
5. Cropland (low intensity)
6. Cropland (reduced intensity)
7. Cropland (high intensity)
8. Urban
B. Spatial prioritization workflow
(3) features and feature targets
(impacted by zone specific decisions)
Species
Carbon
Weights = [$w_{spp1}...w_{sppn}$, w_{carbon}]
Targets = [$t_{spp1}...t_{sppn}$, t_{carbon}]
(1) 100 km² planning unit
protected area
A
B
forest
D
E
C
F
Grassland & shrub
crop
linear decisions (area) by zone
(2) boundary constraints
(per planning unit)
Restoration
Upper limit =
$Pr(j)\sum^{i} a_i T_{i,j}$
Lower limit =
0
Conservation
Upper limit =
land in PU
Lower limit =
land cover
in PA
(6)
Spatially optimal allocation of
conservation, restoration &
production by land cover class
(5) Minimize the
weighted shortfall
from feature targets
(4) linear constraints
(across zones and/or space)
Production
(NUTS2 jurisdiction)
Restoration
(by member state)

Figure 1. Schematic diagram of the prioritization analyses. (A) List of management zones, whose spatial allocation is optimized under a suite of different scenarios (Table S1) in the analyses. (B) For each species, we set an extinction-risk informed target to be met by conserving or restoring habitat types for the species within their potential range. For carbon, the aim was to maximize the amount of carbon stored in conserved or restored areas. Depending on the scenario variants, species were weighted in importance differently relative to carbon to explore the implications of putting more emphasis on different objectives. The area allocation of a planning unit to a given management zone was bounded depending on the planning unit and zone (B2). In addition, the optimization included constraints on the area under restoration and how much area needed to be under production (grazing, farming, timber harvesting, B4) in 2030. The result of this constrained optimization is a series of maps identifying priorities for conservation, restoration, and production of food and timber products (B6).

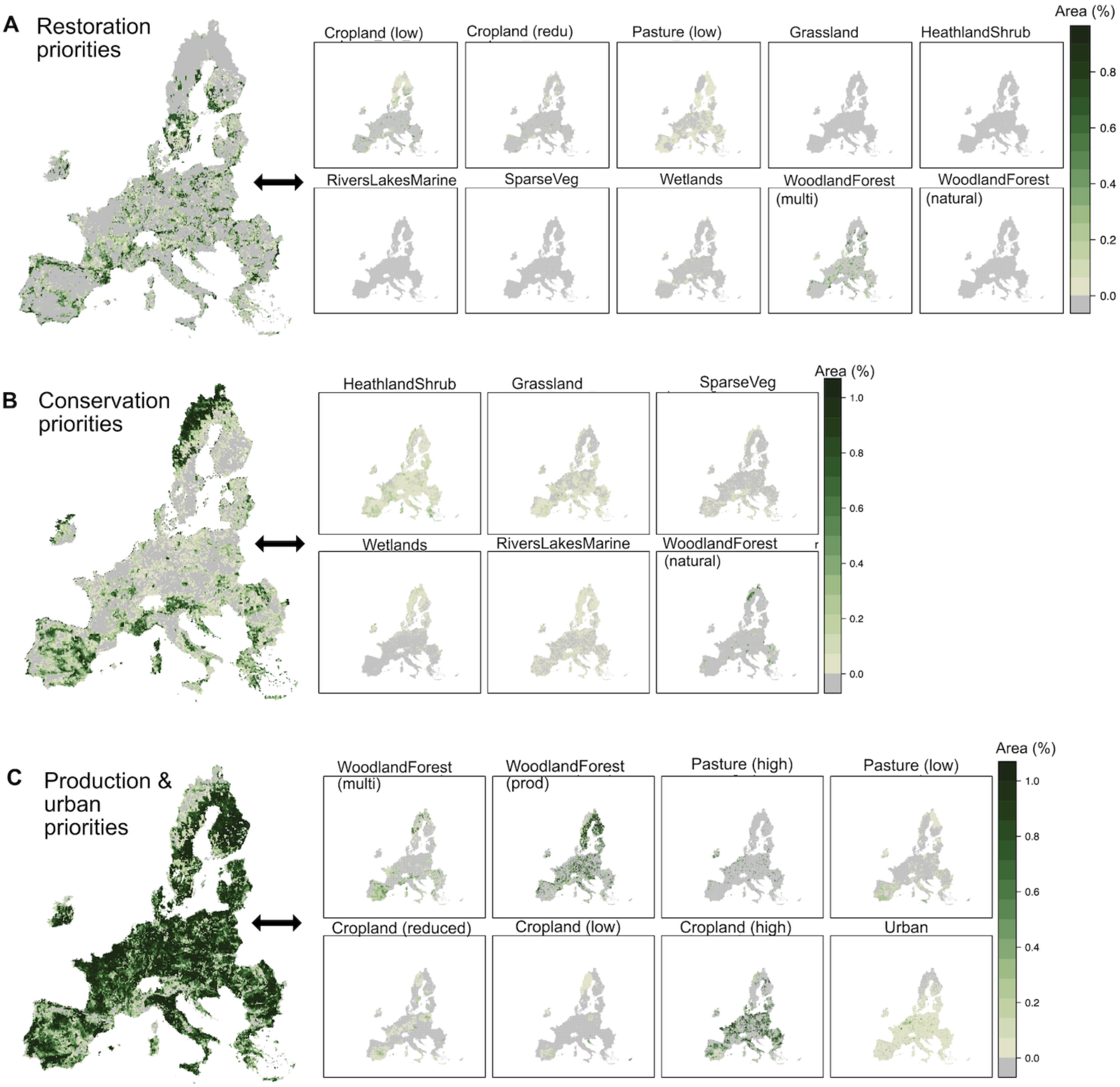

Figure 2. **Optimal allocation of conservation, restoration, and production across land-cover types for one scenario** (carbon target weight set to equal the sum of all species targets and restoration budgets defined at the member state scale). (A) The optimal allocation of conservation (maintained natural land) across the EU and the breakdown of conservation by land cover type. (B) The optimal allocation of restoration across the EU and the breakdown of restoration by land cover type. (C) Production area targets, while constant throughout scenarios at the sub-national (NUTS2) level, vary in their spatial distribution as the result of the conservation and restoration priorities of that given scenario. Urban areas remain constant throughout all scenario solutions and are set to match 2018 urban area distributions in each planning unit.

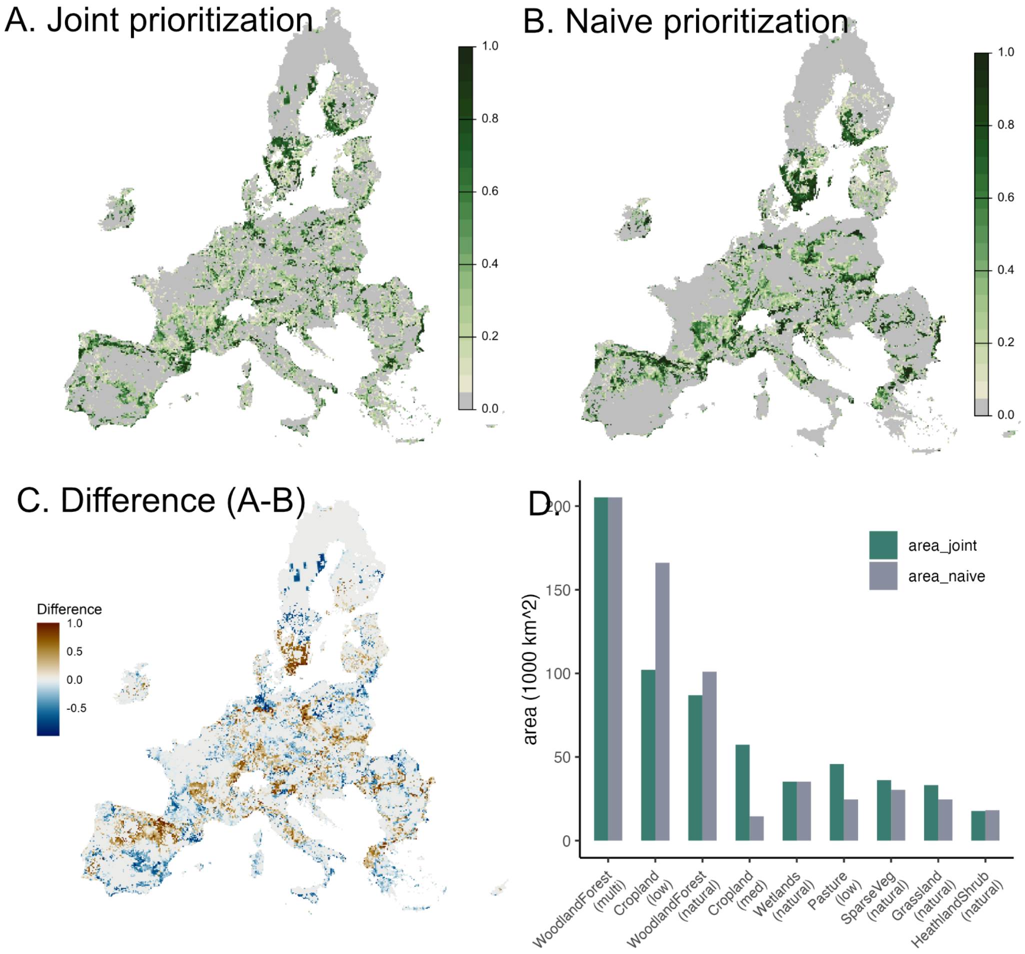

Figure 3. Implications of jointly optimizing the allocation of restoration and conservation. Using the same planning unit constraints and restoration transition matrix from figure 2, we compare solutions of restoration prioritization when (A) jointly optimized to meet restoration targets in the context of the optimal allocation of conservation and production lands (same as figure 2) and (B) only considering restoration priorities. (C) Significant spatial differences and (D) land type differences in restoration priority emerge between the two solutions.

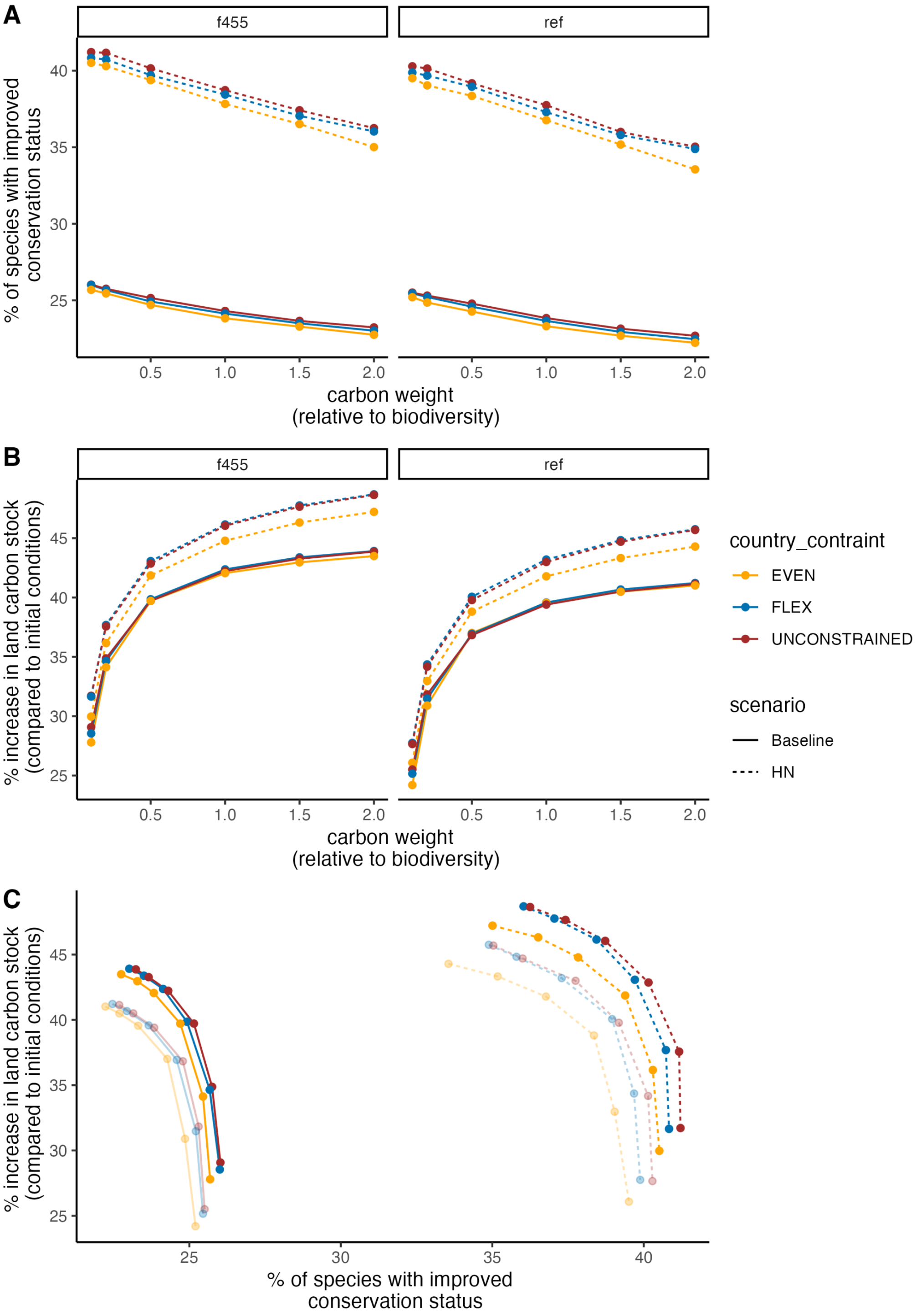

Figure 4. Each point represents a different weighting carbon relative to biodiversity objectives (A) Contributions of solutions to improving the conservation status of species have a linear response to the weighting of species targets relative to carbon targets across all scenarios. (B) By contrast, carbon stock increases are non-linearly with carbon weight. (C) The carbon and biodiversity impacts of all scenarios show that objective weighting and restoration scenarios have a larger impact than burden sharing on the expected 2030 outcomes.

Table 1. We consider 120 different solution scenarios (Figure 1A) below and throughout the manuscript.

Restoration scenario	**Burden Sharing**	**Carbon Weight**	**Production constraints**
Baseline	Even	10 different weighting (0.1-2)	F455
Baseline	Flex (restoration up to 25% per MS)	10 different weighting (0.1-2)	F455
Baseline	Unconstrained	10 different weighting (0.1-2)	F455
Baseline	Even	10 different weighting (0.1-2)	REF
Baseline	Flex (restoration up to 25% per MS)	10 different weighting (0.1-2)	REF
Baseline	Unconstrained	10 different weighting (0.1-2)	REF
High Nature	Even	10 different weighting (0.1-2)	F455
High Nature	Flex (restoration up to 25% per MS)	10 different weighting (0.1-2)	F455
High Nature	Unconstrained	10 different weighting (0.1-2)	F455
High Nature	Even	10 different weighting (0.1-2)	REF
High Nature	Flex (restoration up to 25% per MS)	10 different weighting (0.1-2)	REF
High Nature	Unconstrained	10 different weighting (0.1-2)	REF

Table 2. **Cropland carbon adjustments by intensity level.** These values are used as multipliers to refine the spatial carbon estimates which are not intensity specific. The mean IPCC default value for different production intensities are mapped to zone classification and normalized to 1.

Level	Temperature regime	Zone classification	IPCC default
full	all	high	1
reduced	cool temperate	mid	0.98
reduced	cool temperate	mid	1.04
reduced	warm temperate	mid	0.99
reduced	warm temperate	mid	1.05
no-till	cool temperate	low	1.03
no-till	cool temperate	low	1.09
no-till	warm temperate	low	1.04
no-till	warm temperate	low	1.1

Table 3. **Pasture carbon adjustments by intensity level.** These values are used as multipliers to refine the spatial carbon estimates which are not intensity specific. The mean IPCC default value per zone classification is normalized to 1.

	Level	Temperature regime	Zone classification	IPCC Default
Management (FMG)	Nominally managed (non – degraded)	all	High intensity	1
Management (FMG)	High intensity grazing	all	High intensity	0.9
Management (FMG)	Severly degraded	all	High intensity	0.7
Management (FMG)	Improved	temperate	Low intensity	1.14

Table 4. **Forestry carbon adjustments by intensity level.** These values are calculated using outputs from the G4M model. We average values across all countries for simplicity and use as a multiplier to spatially explicit carbon maps.

Intensity	**Value**	**Emission factor**
production	194.70	0.40
primary	484.72	1
multi	339.70	0.70

Table 5. Mapping EEA threats onto crop production intensity zones. Threat codes align with species included in the prioritization, allowing for the differentiation of biodiversity contributions of different intensities of production.

Code	Pressure/threat	Crop (high)	Crop (redu)	Crop (low)
A01	Conversion into agricultural land	3	2	1
A02	Conversion from one type of agricultural land use to another	2	1	0
A03	Conversion from mixed farming/ agroforestry to specialised production	2	1	0
A04	Changes in terrain and surface of agricultural areas	2	1	0
A05	Removal of small landscape features for agricultural land parcel consolidation	1	1	0
A07	Abandonment of management/use of other agricultural/agroforestry systems	1	1	0
A08	Mowing or cutting of grasslands	1	1	0
A15	Tillage practices (e.g. ploughing) in agriculture	2	1	0
A16	Other soil management practices in agriculture	2	1	0
A17	Harvesting of crops and cutting of croplands	2	1	0
A18	Irrigation of agricultural land	2	1	1
A19	Application of natural fertilisers on agricultural land	2	1	1
A20	Application of synthetic (mineral) fertilisers on agricultural land	2	1	1
A21	Use of plant protection chemicals in agriculture	2	1	1
A22	Use of physical plant protection in agriculture	2	1	1
A23	Use of other pest control methods in agriculture (excluding tillage)	1	1	2
A24	Waste management practices in agriculture	1	1	1
A25	Agricultural activities generating point source pollution to surface/ground waters	3	2	1
A26	Agricultural activities generating diffuse pollution to surface or ground waters	3	2	1
A29	Agricultural activities generating soil pollution	3	2	1
A30	Active abstractions from groundwater/ surface water for agriculture	3	2	1
A31	Drainage for use as agricultural land	3	2	1
A34	Introduction and spread of new crops (including GMOs)	2	1	0
A35	Agricultural crops for renewable energy production	3	2	1

Table 6. Mapping EEA threats onto pasture production intensity zones. Threat codes align with species included in the prioritization, allowing for the differentiation of biodiversity contributions of different intensities of production.

Code	Pressure/threat	PastureLow	PastureHigh
A01	Conversion into agricultural land	1	2
A02	Conversion from one type of agricultural land use to another	0	1
A03	Conversion from mixed farming/ agroforestry systems to specialised production	0	1
A04	Changes in terrain and surface of agricultural areas	1	2
A05	Removal of small landscape features for agricultural land parcel consolidation	0	1
A06	Abandonment of grassland management	0	1
A07	Abandonment of management/use of other agricultural and agroforestry systems	0	1
A08	Mowing or cutting of grasslands	1	2
A09	Intensive grazing or overgrazing by livestock	1	2
A10	Extensive grazing or undergrazing by livestock	0	0
A13	Reseeding of grasslands and other semi-natural habitats	1	1
A14	Livestock farming (without grazing)	1	2

Table 7. Mapping EEA threats onto forestry production intensity zones. Threat codes align with species included in the prioritization, allowing for the differentiation of biodiversity contributions of different intensities of production.

Code	Pressure/threat	Forestry (multi)	Forestry (production)
B01	Conversion to forest from other land uses, or afforestation (excluding drainage)	0	0
B02	Conversion to other types of forests including monocultures	1	1
B03	Replanting with or introducing non-native or non-typical species	1	1
B04	Abandonment of traditional forest management	1	2
B05	Logging without replanting or natural regrowth	1	1
B06	Logging (excluding clear cutting) of individual trees	1	0
B07	Removal of dead and dying trees, including debris	1	2
B08	Removal of old trees (excluding dead or dying trees)	1	1
B09	Clear-cutting, removal of all trees	0	1
B10	Illegal logging	1	1
B11	Cork extraction and forest exploitation excluding logging	0	0
B12	Thinning of tree layer	1	1
B13	Burning for forestry	0	0
B14	Suppression of fire for forestry	0	0
B15	Forest management reducing old growth forests	0	1
B16	Wood transport	1	1
B17	Tillage practices in forestry and other soil management practices in forestry	0	1
B18	Application of natural fertilisers	1	1
B19	Application of synthetic fertilisers in forestry, including liming of forest soils	1	2
B20	Use of plant protection chemicals in forestry	1	2
B21	Use of physical plant protection in forestry, excluding tree layer thinning	1	1
B22	Use of other pest control methods in forestry	1	1
B23	Forestry activities generating pollution to surface or ground waters	1	2
B24	Forestry activities generating air pollution	1	1
B25	Forestry activities generating marine pollution	1	2
B26	Forestry activities generating soil pollution	1	2
B27	Modification of hydrological conditions	1	1
B28	Forests for renewable energy production	1	1
B29	Other forestry activities, excluding those relating to agro-forestry	1	2

Supplemental Methods

1 Data

1.1 Biodiversity data

1.1.1 Biodiversity data collation

Openly available biodiversity data sources in Europe are heterogeneous in type, format and purpose; and to be able to use them in an integrated SDM type of approach, a considerable amount of data harmonization and format control is necessary.

Throughout we followed the taxonomic "backbone" of GBIF and codified functions to harmonize and match taxonomic names from different data sources to the GBIF taxonomy backbone of 2021, (GBIF Secretariat (2021)). We primarily focused throughout on terrestrial species listed in the EU Article 12 (Birds directive) and Article 17 (Habitats directive), or which are assessed by European Redlist of species. A complete list of all included species can be found at https://github.com/milliechapman/EU-restoration-prioritization/blob/main/figures/updated/spp_list.csv.

Firstly, we obtained presence-only records from GBIF for all animal and plant species in the database ((GBIF Secretariat (2021)). We excluded fossil specimens, and those with invalid spatial coordinates, and applied standard data pre-preprocessing steps for unstructured citizen science data using the 'CoordinateCleaner' R package[1]. We removed duplicate points (those within a 2-km distance within the same year) and highly uncertain records.

Additionally, we collated taxonomic group specific data for bird and plant species. For birds we made use of eBird data[2], which we processed similarly as GBIF records above. Potential absence data were inferred from sites where full communities were assessed and for which the focal species had never been recorded. To further limit the number of total absence sites, we first took eBird sites where the species had been recorded, and spatially buffered these by 200 km, excluding any sites within this buffer zone. From the remaining potential absence sites, we randomly selected an equal number of absences as presence sites in which the species was recorded, up to a maximum of 500 per species. For plant species, we also obtained presence-absence data from comprehensive vegetation plot inventories collated and made available through the SPlotOpen database[3]. We filtered these data to Europe, and the representative subset species, as well as excluding any observations prior the year 2000 as above, and to records with a positional uncertainty of less than 2 km. We inferred absence data similarly to eBird but using a lower maximum of 100 absences at maximum, because of the smaller size of this dataset.

We further obtained polygonal global, European, and Mediterranean species ranges from the IUCN Red List version 2021-2 (IUCN 2021) and from BirdLife International (BirdLife International and Handbook of the Birds of the World 2020). These data were filtered to only include areas where species were recorded as extant, possibly extant, or possibly extinct, and included all seasonal occurrences. Where existing for a given species, we further compiled habitat preference (land-cover and elevation) and threat information using data from the IUCN Red List (IUCN 2021) and (EEA preferences). Those estimates were linked to species-specific priors (see below).

Finally, we also obtained the 2020 spatial distribution data for birds listed in the Article 12 Birds Directive and for animals and plants listed in the Article 17 Habitats Directive, excluding sensitive species. Although these data do include population estimates recorded at the (sometimes sub-) Member State level, for this work we used these data only as occurrence data recorded as presence-only atlas data on a 10-km grid across Europe. For the course of this work, we treated the species information reported at various sites (polygons) in the Natura2000 network as presence-absence information, recognizing that surveys in some sites might be incomplete or outdated and not all species are necessarily recorded during. A R-package was created to format these data for the modelling (https://github.com/iiasa/rN2000). We supplemented these data with species checklists for Important Bird Areas (IBAs,[4]) across Europe, adding presence-absence records per species for the polygon area of the IBAs where the species was recorded or not.

1.1.2 Predictor Variables

We prepared a series of environmental predictors related to topography, soil conditions, climate and land cover.

In both planning and species distribution modeling, we make use of land cover and land-use data that is consistent with the European Ecosystem accounting framework such as the Mapping and Assessment of Ecosystems and their Services (MAES) system. For the current distribution of land cover, we used data from the (https://land.copernicus.eu/pan-european/corine-land-cover/clc2018). For mapping the potential natural distribution of a species (see below) we additionally considered data on the potential distribution of land cover[5] matched to the same legend. The thematic legend of the Corine land-cover data was then recategorized into different MAES categories through a crosswalk, however, differed from the MAES categories as we split the class Pasture (2.3.1) from other Grasslands as considered by the MAES Grassland class.

In addition to land cover and land use we also considered data on topography of European landscapes making use of the EU DEM ver1.1 (https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1). We considered the mean elevation (in m), the topographic position index (TPI,[6]) and the aspect of the topography, which we transformed into eastness and northness estimates through a sinusoidal and cosinus transformation respectively. This transformation is necessary to avoid circularities in units (0 degree being closer to 359 degrees) caused by the degrees-based characterization of an aspect layer.

We used long-term average climatic conditions in Europe. Specifically, we leveraged downscaled bioclimatic ERA5 indicators over the last 40 years (1979 to 2018) from the European Copernicus program. These climatic indicators represent Essential Climate Variables (ECV) such as the surface energy, drought or moisture all of which are known to be important factors in delineating the range and environmental niche. Specifically, we made use of the BIOCLIM data BIO01 to BIO19, average aridity and cloud cover, the annual sum of frost days, potential evaporation and volumetric soil water as well as different characterizations of the number for growing degree days and the start, end and length of the growing season[7]. For those parts of European member states which are missing in the European downscaled Copernicus Climate products (Such as the Spanish Canary islands) we used the average values of the global rather than the downscaled climate product instead.

For both current and potential species projections of species distributions, we considered only variables related to land cover as well as temporally static variables that are unlikely to change in a future world such as for instance altitude. On the other hand, for predictions of the potential natural distribution of a species we made use of all predictors excluding those related to land cover and land use. We added species-group-specific sets of predictors to the modeling to be included. For example, for bird species, we additionally included a layer depicting the Euclidean distance to the ocean from each land grid cell, given the importance of marine waters to many onshore nesting grassland and wetland species. For plant species we included spatial-explicit predictors related to groundwater and soil conditions, specifically data on the Ph value and Calcium Carbonate content of groundwater resources as well as estimates of the depth to groundwater in meters from[8]. In particular, groundwater Ph and Calcium carbonate do not cover small islands (Madeira, Canary Islands) as well as Cyprus, which is why we filled any remaining missing values of the predictors with a spatial prediction from a random forest model, using spatial proximity and climatic variables as covariates[5]. We also included data from a thematic layer of a European soil lithology classification system owing to the importance of different soil types to growth and niche space of plant species.

Most observational species occurrence points in Europe are known to be biased towards areas with higher accessibility and wealth, with critically Eastern European member states having a comparably lower density of records compared to western and northern European member states. Besides the integration of multiple datasets, we attempted to control for such sampling biases through a model-based control following[9]. To do so, we first took the presence-only and presence-absence localities of all biodiversity sources considered in this work (see above) and rasterizing them for the target background. This resulted in a counted number of points for each grid cell, which we then aggregated overall as the total sum of all occurrences. Furthermore, we prepared data on the accessibility of land ([10]) and the human population density per grid cell using data from the GHSL product for Europe ([11]). The biased background grid was then created by first calculating an adjusted log transformation of each individual layer through the following equation ($log(1+x-min(x)*w)$), where w stands for a numerical weight reflective of the direction of bias (-1 for accessibility, 1 for all others), and afterwards the individual layers normalized to a range of 0 to 1 and averaged.

Finally, all layers were aggregated from their original resolution to a 10-km (Lamberts Equal Area projection) grain size determined by the background modelling domain layer. We did so by either calculating the proportion of grid cells in each coarser grain if these binary data, by calculating the bilinear resampled average of all values within the coarser grain, or – in the case of multinomial categorical data – calculating the mode of all finer grained classes. For the modelling all continuous variables were standardized to the mean and divided by their standard deviation, thus ensuring compatibility in terms of units. All covariates were matched against the modelling extent and made consistent with a NUTS2 representation of European countries. Any missing data at the pixel level not filled or extrapolated at this stage was filled with missing data across all covariates. All calculations and variable preparations were done using GDAL and R packages such as "Raster" or "sf".

1.1.3 Current distribution of a species

We estimated the current distribution of species using an integrated species distribution modelling (iSDM) approach where different best-available data sources (occurrence, preference, expert information) are integrated into one joint prediction using different types of linear and non-linear modelling approaches[12].

We collated for each species the available suitable data (see above), separating between different types, namely presence-only, presence-absence and polygon presence-only data, the latter of which is used in the form of spatial-explicit offsets (e.g. expert range). For species occurring in Natura 2000 or Important Bird area (IBA) sites, we assumed that the species occur in all Natura 2000 sites in which a presence was indicated and that biophysical conditions are relatively homogeneous within Natura 2000 sites (which, given the small size of many sites and the SDM modelling grain size of 10km is a reasonable assumption).

We sampled at random across all Natura 2000 and IBA sites presence point estimates up to a number of two-times as many as there are other occurrence observations (see above), to broadly characterize the environmental conditions prevalent in those sites. We furthermore created an equal number of absence point data which we sampled at random across all sites excluding the ones where the species has been recorded as present (thus resulting in species-specific contrasts).

Because of computational reasons and to further reduce sampling biases, we applied thinning on point data across all datasets for species with more than 200 records. The process of spatial thinning removes occurrence points at random from areas that are oversampled, for example because of sampling or spatial biases in the database[13]. Notably thinning only removes points from grid cells where there are multiple and never removes all points from any grid cell completely. For presence-only records from GBIF – usually the largest data source by size – we first applied a bias thinning, where we preferentially removed observations from 10km grid cells considered as biased (based on occurrence information across all species). In addition, and across all point occurrence datasets, we also removed at random observations until a minimum number of 10 points at maximum has been reached. This approach ensures that presence and absence information (where existing) are relatively homogeneously distributed in density across the European land area, thus representing average conditions representing the suitable species habitats across the modelling period.

Whenever presence-only atlas or expert-delineated information on the occurrence of species existed, such as for example from the global, Mediterranean or European IUCN assessments, the Amphibian and Reptile atlas[14], the Atlas Hymenoptera for bumblebee species (http://www.atlashymenoptera.net/default.aspx) or for polygon information from the EU Habitats directive or Bird directive data, we included this information as spatial offsets. For IUCN we only used those parts of the range where the species is permanently resident or which are part of its breeding distribution. Spatial offsets can be included in species distribution model as spatial priors, thus increase the probability of any given grid cell to be suitable for a species[15]. We first binarized the different range estimates and then calculated the Euclidean distance from the boundary of the range to all other grid cells in the modelling background, after which we applied a negative exponential kernel with a average dispersal distance estimates to account for the decreasing suitable value of a grid cell[16]. The resulting distance layer was then rescaled to a range from 0 (furthest away) to 1 (within the range). Notably we used different distance transforms depending on whether bird or non-bird species were estimated, using either infinite and an average 20-km distance transformations for non-bird species respectively. All offsets created in this way were log-transformed before adding them to any model using presence-only information and in the case multiple offsets were supplied, these were combined first via simple multiplication.

To avoid overprediction into novel areas, the predictions were spatially constrained by a broad environmental zoning layer[?] and the expanded offsets highlighted above. This was done by removing broad zones in which there are no contemporary occurrence points to avoid, for instance, extrapolations from a Mediterranean into boreal climatic zones, while also allowing modest extrapolations within similar environmental conditions. It should be noted that this zoning was only included for current projections and not for any potential distribution. We also tested for collinearity between included variables, removing those that were highly collinear (Pearson's r > 0.7) unless they were known to be of particular importance to a species (e.g. have a set prior, see below).

We estimated the potential distribution of the species through an ensemble modelling approach (stacked SDM,[17, 18]) using state-of-the-art machine learning and Bayesian algorithms that complement each other's strengths. Model structure and response were determined based on data type, with Poisson Process models being fitted for presence-only datasets and logistic regressions for presence-absence data. We fitted tree-based regressions using the XGBoost modelling approach[19]. XGBoost makes use of gradient descent boosting, supports variable regularization and also non-linear tree-based regressions. XGBoost Models were fitted using a total of 10000 boosting iterations, a learning rate of 0.001 and Gamma parameter of 4 (larger is more conservative) for regularization. We also used another gradient descent boosting algorithm (GDB) available from the 'mboost' R-package[20]. GDB models makes use of non-linear baselearners (splines) for additive inference similar to the popular Generalized Additive Models (GAMs), however in contrast to GAMs it also supports variable regularization directly through boosting and additional baselearners (see below on priors). Here models were fitted using a total of 2500 boosting iterations and a learning rate of 0.001 per iteration. Bayesian regularized regressions were fitted using the 'BoomSpikeSlab' R-package[21] and 10 000 MCMC iterations and four MCMC chains. Lastly we used regularized linear regression models fitted with the "glmnet" package[22]. In a Poisson-Process modelling framing these type of regressions are statistically equivalent to the popular maxent package[23].

Models were only fitted for those species for which at least – after thinning - 20 data points were available, assuming that species with fewer records have not been sampled comprehensively enough to make inferences about their current distribution. Further, we made use of simple rules to avoid fitting overly complex models for a limited number of observations. Only linear models (boosted and non-boosted) were fitted for species with fewer than 100 observational points and for species with point observations fewer than 1.5 times the minimum data size of 20, we did only fit linear Bayesian Poisson Process models and not use any non-linear or boosted approaches to avoid overfitting to limited datasets[24]. Linear regressions, compared to non-linear ones, usually fare better when the goal is extrapolation and are less prone to model overfitting. In case only presence-only information from GBIF was available for a species, we furthermore included spatial effects as covariate using polynomial-transformed coordinates[16].

Integrating prior information on species habitat and elevational preferences and distances to known occurrences can improve range estimates. We obtained information on the susceptibility of species to certain habitat and elevational preferences from the EEA habitat preference database and IUCN[25]. Preferences to certain land-cover types in the IUCN habitat preferences or respectably Corine land-cover categories were remapped to MAES categories. Priors can help to stabilize and avoid mapping unrealistic response functions to certain covariates[26]. Elevational preferences were included as specific threshold transforms on the raw elevation data and were used instead of the raw elevation data instead. This approach thus creates two separate discrete elevational bounds in which a species might or might not exist. We specified monotonically constrained baselearners for both XGBoost and GDB[27], which are prior constraints placed on the linear and non-linear effects to follow certain directions. Here we used priors assuming either increasing, in case the habitat was highly preferred or suitable, or alternatively positive constraints when the species was just known to occur in the habitat. Previous studies have shown that the use of such monotonicity constraints in SDMs can results in more ecological plausible response functions[27]. For Bayesian Poisson-Process models we used Zellner-style spike and slab priors with two parameters, a coefficient for a Gaussian prior on the mean coefficient of the covariate and a inclusion probability which states the probability by which a certain variable is to be used and thus avoided to be regularized out from the model[28]. For habitats preferred by a species we used mean coefficients of 3 and an inclusion probability of 1, for suitable habitats we used a coefficient of 1 and 0.5 respectively and for occasionally occupied habitats we used 0.1 in both cases. Similarly, for BART models we specified priors as transition probabilities for the variable so that the regression tree is forced - with a certain probability, here 0.75 – to generate a split for a given variable.

On the full point occurrence dataset, we then applied a spatial block cross validation scheme using the blockCV R-package[18][29]. Specifically, we created three spatial folds of training and testing data to evaluate each of the three algorithms on. However for species with very few occurrence records overall (< 50) and where the creation of spatial folds failed (owing to points being too close in distance), we instead implemented default randomized folds where 25% of data was removed at random. All predictors were scaled prior to model fitting by subtracting the mean from each value and dividing by their standard deviation to ensure comparable unit scales. We included among the final predictors also the bias variable (see above), which was controlled during the prediction[9], thus helping to reduce some of the spatial biases in available occurrence datasets. Ensemble of different datasets per species were integrated and thus separate models were estimated for each spatial block and for each data type[30].

Each separate model prediction was binarized using a 0.05 percentile threshold and then validated using the withheld data to obtain an estimate of the True-Skill-Statistic (TSS). We used the TSS values to create a weighted mean ensemble of all predictions[12]. Individual predictions from different models were first normalized owing to the different units (relative rate of occurrence compared to relative occurrence probability). We then thresholded all ensembles using a 5% minimum percentile threshold on all observed data points (across datasets), thus creating a conservative estimate of where suitable habitat for a species might or might not persist in Europe[24]. We used percentile-based thresholds[31] opposed to approaches maximizing any performance metric since they can be applied across different predictions and dataset types (presence-only and presence-absence). Furthermore it assumes that the least suitable habitat at which the species is known to occur is the minimum suitability value for the species, while allowing for some flexibility so that outliers do not bias the threshold. For each species the validation statistics, the predicted suitability and the thresholded map is then retained. All SDMs were fitted using the integrated species distribution modelling framework ibis.iSDM coded for R[12].

1.1.4 Potential distribution of a species

The goal of this modelling is to obtain a depiction of the potential distribution of the species (sensu[5]). We understand potential in this context as the contemporary climatic, soil and natural vegetation conditions that would allow a species to persist in an area. Critically, and opposed to the mapping of current suitable habitat, this approach considers only contemporary differences in climate and soil, and not any land-cover or land-use, aligning with the concept of potential natural vegetation of Europe[32]. The predictions from the species distribution models used here thus aim to depict where a species might exist under contemporary conditions, while also allowing modest inter- and extrapolation from its current distribution.

While for the current estimation of species distribution (see above) we considered each biodiversity data type separately in a model, for the potential distribution of the species we merged datasets with presence-only data and presence-absence data, adding pseudo-absence points to the former[33, 34]. This is a widely applied approach for SDM mapping, which however is scale dependent and can result in an overestimation of the niche[?][26], yet in this context is acceptable given that our aim is to map the widest possible potential distribution of species (although we modestly constrain the prediction, see below). Although there can be benefits in modelling these datasets jointly for more constrained predictions[?], our aim is to identify and characterize the maximum potential extent of the environmental niche of a species given contemporary conditions. We first combined all cleaned and filtered point occurrence data into one joint dataset, removing duplicates per grid cell in the process. For each presence-only dataset we created pseudo-absence points randomly distributed within the modelling domain, but spatially weighted them so that pseudo-absences preferentially fall into areas with high bias defined by human population density.

We used a similar ensemble modelling approach as for the current estimates of species ranges (see above), however used binomial distributed responses throughout, adding pseudo-absence points for presence-only datasets. For validating and thresholding the ensemble models we evaluated the predictions in terms of their accuracy through the F1 score, which is calculated as the ratio of the model precision (true positives) and the recall (also known as sensitivity). We specifically chose the F1 score for evaluation since it maximizes correct predictions and thus can help to ensure that most training occurrence points are retained. The final ensemble prediction was then created as a weighted mean of the nine different F1 scores (3 spatial blocked subsets per algorithm). All modelling was done in the integrated species distribution modelling framework ibis.iSDM coded for R[12].

1.2 Land cover data

1.2.1 Potential distribution of natural land cover

For the identification of potentially restorable land we followed the concept of potentially natural vegetation (sensu[5]). To map potential land cover, we first assembled a Europe-wide database on the distribution of habitats in Europe where we followed the thematic legend of the MAES habitat classification system at level 2, while ignoring any strictly anthropogenic habitats (e.g. Urban, Cropland, Pasture) as well as Rivers and Lakes. Different data sources on the distribution of habitats differ in terms of their geographic spread and biases. In order to not rely on any single data source of European ecosystems we integrated habitat data from three different sources collated for Europe:

We took habitat information from the European Habitat Directive which gives the occurrence of all EUNIS habitats listed in the Article 17 of the habitats directive at a 10km resolution. We used a crosswalk developed by the European Environment Agency and Biodiversity Topic Centre to translate the EUNIS types to Corine CLC and subsequently to MAES level 2 categories[35]. We further made use of point occurrence datasets for key habitats in the new EEA suitability predictions[36]. Here the habitat categories were reclassified into the respective MAES types, following the CLC to MAES crosswalk. Finally, we prepared point occurrence data from the openly available land-cover and land-use database LUCAS[37]. The LUCAS database contains stratified and repeated survey records of local land-cover and land-use types for Europe[37] with the latest one being available for the year 2018. We took the LUCAS survey records and reclassified the land-cover type (”$lc1_table$”) to the natural MAES ecosystem categories, discarding all anthropogenic created habitats (Cropland, Urban, ...) in the process.

As predictors for the potential habitat modelling, we considered data on the potential distribution of land cover ([38],[32]) as well as long-term average climatic conditions in Europe where we used downscaled bioclimatic ERA5 indicators over the last 40 years (1979 to 2018) from the European Copernicus program. These climatic indicators represent Essential Climate Variables (ECV) such as the surface energy, drought or moisture all of which are known to be important factors in delineating the range and environmental niche. Specifically, we made use of the BIOCLIM data BIO01 to BIO19, average aridity and cloud cover, the annual sum of frost days, potential evaporation and volumetric soil water as well as different characterizations of the number of growing degree days and the start, end and length of the growing season. We also included a predictor that quantified the Euclidean distance to the ocean from each terrestrial grid cell, given the importance of some wetland habitats to brackish water and coastal conditions. We prepared data on groundwater and soil conditions, specifically data on the Ph value and Calcium Carbonate content of groundwater resources as well as estimates of the depth to groundwater in meters[8,39]. We also included data from a thematic layer of a European soil lithology classification system[40] owing to the importance of difference soil types. The individual lithology classes were included as factorial combinations in the modeling.

We estimated the potential distribution of the habitat by relating presence-only habitats with pseudo-absence points[33,34]. This is a widely applied approach usually for species distribution modelling, which although it can result in “overconfident” extrapolations, is in this context desirable since our aim is to map the potential natural distribution of a habitat. Although it is possible to create predictions of potential natural habitat as multi-nominal problem, e.g. where each class has a different and exclusive probability to potentially occur[5], we decided instead to estimate the distribution of the habitat separately, since in many areas of Europe there is a potential for more than one habitat to potentially occur under natural conditions especially when succession trajectories are unknown. Thus, for each habitat dataset we created an equal balanced number of pseudo-absence points randomly distributed within the modelling domain. We furthermore rasterized the 10-km Article 17 data and applied an Euclidean distance transform to them, e.g. there is a monotonically decreasing probability of a habitat type to occur outside the Article 17 reporting data. The resulting layer was then included as an additional covariate.

We used non-linear and tree-based Bayesian Regression Trees (BART) for projecting the probability of any potential habitat. The BART algorithm has the benefit of being able to quantify complex non-linear interactions between variables as well as being able to consider prior information in a Bayesian framework[41]. For the regression trees we used a logistic model formulation of the response by assuming the habitat presence and pseudo-absences to be Bernoulli distributed, e.g.

$y_{habitat}$ $Pr(y = 1|x|)$.

We fitted the BART models with 500 tree and 50 burn-in iterations across four MCMC chains through the use of the ‘dbarts’ R-package[41]. From the resulting posterior of the fitted model and for each grid cell, we summarized the median and lower (5%) and upper (95%) percentile of the posterior, thus allowing us to spatially represent the uncertainty of each individual habitat type prediction. The resulting predictions thus contain an estimate of the probability of a potential occurrence of a given natural habitat for each 10km grid cell.

1.3 Carbon data

1.3.1 Current distribution of carbon

For current carbon stocks we used data on above-ground, below-ground and soil organic carbon at risk from land-use change from[31]. These data were created by selecting and integrating best available carbon maps for different vegetation classes. For more detail on the integration and handling of individual data layers see[31]. We used the carbon data at the original resolution of 100m and intersected them with the current distribution of land cover according to the Corine landcover data for 2018, which we reclassified to the MAES legend. The intersected individual carbon estimates were then aggregated (arithmetic mean) to a grain size of 10km used for the prioritization. All data are in units of tC/ha and for the analysis we combined the current carbon layers by calculating the combined sum of above- & below-ground and soil organic carbon for Europe and included it as an additional feature in the prioritization.

1.3.2 Potential carbon

To spatially allocate specific restoration priorities, we needed to identify areas with high carbon sequestration potential. Here we used an approach that combined the different techniques from[42] and[43]. First, we created a regular sampling grid at 1km for each current MAES ecosystem type reclassified from Corine 2018 (see above) and extracted the fraction of the respective land cover type. We then extracted estimates of current reference carbon data (in tC/ha) from the above, below and soil organic carbon data layers from[31] as well as from the European-specific JRC Biomass map[?]For each of the different types of carbon products (below, above and soil carbon) we then calculated a consensus estimate (arithmetic mean) per gridded 1km point. We further extracted estimates on whether a given point locality was situated in a peat land as considered by the European peatland map[?] or land covered by primary forest[44]

From the resulting extracted estimates, we then selected for each natural land cover type (e.g. Grassland, Heathland, Marine inlets, Sparsely vegetated land, Wetland and Forest and Woodland) a total of 10000 reference points for modelling training. We ensured that (a) the respective land cover type currently covers at least 50% of a given 1km grid cell, (b) average carbon density estimates are in the largest 95% percentile of values, (c) the reference points were preferentially sampled in remaining European peat and primary forest sites for the Wetland, Marine Inlets and Forest & Woodland classes, (d) points were geographically representative by covering each European biogeographical regions (adjusted for area) and (e) that extracted mean carbon density estimates were corrected for the fraction of non-natural land contained within them. Instead of using a single reference value for the carbon contained in non-natural systems[42], we calculated the average estimate of all non-natural land cover types in MAES (e.g. cropland and urban).

As above for potential species distributions and land cover, we then subjected these reference estimates to a spatial extrapolation approach. Here we followed an approach set out by[43] and estimated potential carbon density as $CDensPo = f(S,T,C)$, where potential carbon density is predicted as a function of soil, S, topographic, T, and climatic, C, factors. We used the same predictors as for potential land cover and species occurrences. Finally, we corrected each estimate by the amount of potentially occurring natural land cover.

References

1. Zizka, A. *et al.* <span style="font-variant:small-caps;">CoordinateCleaner</span> : Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* **10**, 744–751, DOI: 10.1111/2041-210X.13152 (2019).

2. of Ornithology, C. L. eBird Basic Dataset. Version: EBD_relmar-2021. (2021).

3. Sabatini, F. M. & et al. sPlotOpen – An environmentally balanced, open-access, global dataset of vegetation plots. DOI: 10.1111/geb.13346 (2021).

4. Donald, P. F. *et al.* Important Bird and Biodiversity Areas (IBAs): the development and characteristics of a global inventory of key sites for biodiversity. *Bird Conserv. Int.* **29**, 177–198, DOI: 10.1017/S0959270918000102 (2019).

5. Hengl, T. *et al.* Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential. *PeerJ* **6**, DOI: 10.7717/peerj.5457 (2018).

6. De Reu, J. *et al.* Application of the topographic position index to heterogeneous landscapes. *Geomorphology* **186**, 39–49, DOI: 10.1016/j.geomorph.2012.12.015 (2013).

7. Copernicus Climate Change Service. Downscaled bioclimatic indicators for selected regions from 1979 to 2018 derived from reanalysis, DOI: 10.24381/CDS.FE90A594 (2021).

8. Hájek, M. *et al.* A European map of groundwater pH and calcium. *Earth Syst. Sci. Data* **13**, 1089–1105, DOI: 10.5194/essd-13-1089-2021 (2021).

9. Warton, D. I., Renner, I. W. & Ramp, D. Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology. *PLoS ONE* **8**, e79168, DOI: 10.1371/journal.pone.0079168 (2013).

10. Weiss, D. J. *et al.* A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336, DOI: 10.1038/nature25181 (2018).

11. Pesaresi, M. *et al.* A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **6**, 2102–2131, DOI: 10.1109/JSTARS.2013.2271445 (2013).

12. Jung, M. An integrated species distribution modelling framework for heterogeneous biodiversity data. *Ecol. Informatics* **76**, 102127, DOI: 10.1016/j.ecoinf.2023.102127 (2023).

13. Steen, V. A., Tingley, M. W., Paton, P. W. C. & Elphick, C. S. Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods Ecol. Evol.* **12**, 216–226, DOI: 10.1111/2041-210X.13525 (2021).

14. Sillero, N. *et al.* Updated distribution and biogeography of amphibians and reptiles of Europe. *Amphibia-Reptilia* **35**, 1–31, DOI: 10.1163/15685381-00002935 (2014).

15. Merow, C., Allen, J. M., Aiello-Lammens, M., Silander, J. A. & Fortin, M. Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information. *Glob. Ecol. Biogeogr.* **25**, 1022–1036, DOI: 10.1111/geb.12453 (2016).

16. Domisch, S., Wilson, A. M. & Jetz, W. Model-based integration of observed and expert-based information for assessing the geographic and environmental distribution of freshwater species. *Ecography* **39**, 1078–1088, DOI: 10.1111/ecog.01925 (2016).

17. Biber, M. F., Voskamp, A., Niamir, A., Hickler, T. & Hof, C. A comparison of macroecological and stacked species distribution models to predict future global terrestrial vertebrate richness. *J. Biogeogr.* **47**, 114–129, DOI: 10.1111/jbi.13696 (2020).

18. Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J. & Elith, J. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecol. Monogr.* **92**, DOI: 10.1002/ecm.1486 (2022).

19. Chen, Y. *et al.* Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. *Ecol. Indic.* **135**, 108545, DOI: 10.1016/j.ecolind.2022.108545 (2022).

20. Hofner, B., Mayr, A., Robinzonov, N. & Schmid, M. Model-based boosting in R: a hands-on tutorial using the R package mboost. *Comput. Stat.* **29**, 3–35, DOI: 10.1007/s00180-012-0382-5 (2014).

21. Scott, S. L. BoomSpikeSlab: MCMC for Spike and Slab Regression (2022).

22. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, DOI: 10.18637/jss.v033.i01 (2010).

23. Phillips, S. J. & Dudík, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161–175, DOI: 10.1111/j.0906-7590.2008.5203.x (2008).

24. Merow, C., Smith, M. J. & Silander, J. A. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* **36**, 1058–1069, DOI: 10.1111/j.1600-0587.2013.07872.x (2013).

25. Agency, E. E. Linkages of species and habitat types to MAES ecosystems (2017).

26. Hannemann, H., Willis, K. J. & Macias-Fauria, M. The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble modelling: Unstable response functions in SDMs. *Glob. Ecol. Biogeogr.* **25**, 26–35, DOI: 10.1111/geb.12381 (2016).

27. Hofner, B., Hothorn, T., Kneib, T. & Schmid, M. A Framework for Unbiased Model Selection Based on Boosting. *J. Comput. Graph. Stat.* **20**, 956–971, DOI: 10.1198/jcgs.2011.09220 (2011).

28. Cui, W. & George, E. I. Empirical Bayes vs. fully Bayes variable selection. *J. Stat. Plan. Inference* **138**, 888–900, DOI: 10.1016/j.jspi.2007.02.011 (2008).

29. Meyer, H., Reudenbach, C., Wöllauer, S. & Nauss, T. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecol. Model.* **411**, 108815, DOI: 10.1016/j.ecolmodel.2019.108815 (2019).

30. Fletcher, R. J. *et al.* A practical guide for combining data to model species distributions. *Ecology* e02710, DOI: 10.1002/ecy.2710 (2019).

31. Jung, M. *et al.* Areas of global importance for conserving terrestrial biodiversity, carbon and water. *Nat. Ecol. & Evol.* **5**, 1499–1509, DOI: 10.1038/s41559-021-01528-7 (2021).

32. Bohn, U. Map of the Natural Vegetation of Europe. Tech. Rep. (2004).

33. Guisan, A. & Thuiller, W. Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* **8**, 993–1009, DOI: 10.1111/j.1461-0248.2005.00792.x (2005).

34. Barbet-Massin, M., Jiguet, F., Albert, C. H. & Thuiller, W. Selecting pseudo-absences for species distribution models: how, where and how many?: *How to use pseudo-absences in niche modelling? Methods Ecol. Evol.* **3**, 327–338, DOI: 10.1111/j.2041-210X.2011.00172.x (2012).

35. on Biological Diversity, E. T. C. Crosswalk between EUNIS habitats classification and Corine land cover.

36. Agency, E. E. EUNIS habitat classification (2019).

37. d'Andrimont, R. *et al.* Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union. *Sci. Data* **7**, 352, DOI: 10.1038/s41597-020-00675-z (2020).

38. Hengl, T., Jung, M. & Visconti, P. Potential distribution of land cover classes (Potential Natural Vegetation) at 250 m spatial resolution, DOI: 10.5281/ZENODO.3631254 (2020).

39. Fan, Y., Li, H. & Miguez-Macho, G. Global Patterns of Groundwater Table Depth. *Science* **339**, 940–943, DOI: 10.1126/science.1229881 (2013).

40. European Commission. Joint Research Centre. Institute for Environment and Sustainability. *Soils of the European Union.* (Publications Office, LU, 2008).

41. Carlson, C. J. embarcadero: Species distribution modelling with Bayesian additive regression trees in <span style="font-variant:small-caps;">r</span>. *Methods Ecol. Evol.* **11**, 850–858, DOI: 10.1111/2041-210X.13389 (2020).

42. Strassburg, B. B. N. *et al.* Global priority areas for ecosystem restoration. *Nature* **586**, 724–729, DOI: 10.1038/s41586-020-2784-9 (2020).

43. Walker, W. S. *et al.* The global potential for increased storage of carbon on land. *Proc. Natl. Acad. Sci.* **119**, e2111312119, DOI: 10.1073/pnas.2111312119 (2022).

44. Sabatini, F. M. *et al.* European primary forest database v2.0. *Sci. Data* **8**, 220, DOI: 10.1038/s41597-021-00988-7 (2021).