ARPHA Preprints

PENSOFT

Project Report

Author-formatted document posted on 01/12/2025

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e180864

Final species and habitat distributions for current and future state

Sara Si-Moussi, Marianne Tzivanopoulos, Gabrielle Deschamps, Maxime Hoareau, Julien Renaud, Rémi Lemaire-Patin, Wilfried Thuiller

NATURA CONNECT

Species and habitat distributions for Europe

D3.2 Final species and habitat distributions for current and future state

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

Prepared under contract from the European Commission.

Project acronym:	NaturaConnect
Project full title:	NaturaConnect - Designing a resilient and coherent Trans-European Network for Nature and People
Grant agreement number:	101060429
Start of the project:	1 July 2022
Duration:	48 months
Project coordinators:	International Institute for Applied Systems Analysis (IIASA) and Martin-Luther Universität Halle-Wittenberg (MLU) naturaconnect.eu
Scientific coordinator:	Piero Visconti, PhD, IIASA
Type:	HORIZON Innovation Actions
Call:	HORIZON-CL6-2021-BIODIV-01

The contents of this material are the sole responsibility of the NaturaConnect consortium and do not necessarily reflect the opinion of the European Union. This report reflects the version finalised and submitted to the European Commission on 30.04.2025. Further changes to the report may be integrated following review from the European Commission.

Front cover: © Eloy Revilla, EBD-CSIC

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

Project reference number	**101060429**
Project title	NATURACONNECT - DESIGNING A RESILIENT AND COHERENT TRANS-EUROPEAN NETWORK FOR NATURE AND PEOPLE

Deliverable title	**Final species and habitat distributions for current and future state**
Deliverable number	D3.2
Contractual date of delivery	30.04.2025
Actual date of delivery	30.04.2025
Type of deliverable	DATA — data sets, microdata, etc
Dissemination level	PU - Public
Work package number	WP3
Institution leading work package	10 - CNRS
Task number	T3.3
Institution leading task	CNRS
Author(s)	Sara Si-Moussi, Marianne Tzivanopoulos, Maxime Hoareau, Julien Renaud, Rémi Lemaire-Patin & Wilfried Thuiller
Deliverable reviewers	Martin Jung (IIASA), Jeremy Dertien (MLU), Matea Osti (IIASA)
EC project officer	Christophe Coudun

Deliverable description	The datasets will be the final deliverable of task 3.3 and will include the following: Species distribution under present and future environmental conditions. Present and future potential distirbution of EUNIS habitat classes, but also primary and old-growth forests
Keywords	Species distribution models, habitat suitability, EUNIS habitats, invertebrates, plants, vertebrates, range maps.

Funded by the European Union
NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

Table of Contents

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

List of Figures

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

ARPHA Preprints Author-formatted document posted on 01/12/2025. DOI: https://doi.org/10.3897/arphapreprints.e180864

List of Tables

Funded by the European Union

NatureConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

ARPHA Preprints Author-formatted document posted on 01/12/2025. DOI: https://doi.org/10.3897/arphapreprints.e180864

Abbreviations

AWC	Available Water Capacity
DEM	Digital Elevation Model
EEA	European Environmental Agency
EU	European Union
EUNIS	European Nature Information System
EUNIS-ESy	European Nature Information System Expert System
ESDB	European Soil Database
ETC-BD	European Topic Center on Biological Diversity
EVA	European Vegetation Archive
GBIF	Global Biodiversity Information Facility
GCM	Global Circulation Model
GDD	Growing Degree Days
HDM	Habitat distribution models
IUCN	International Union for Conservation of Nature
LDAM	Label Distribution Aware Margin loss
MLP	Multi-layer Perceptrons
NFF	Nature Futures Framework
NN	Neural Networks
RCP	Representative Concentration Pathways
RF	Random Forest
LightGBM	Light Gradient Boosting Machine
LUCAS	Land Use/Cover Area frame Survey
SCD	Snow Cover Days
SDM	Species distribution models
SOC	Soil Organic Carbon
SWE	Snow Water Equivalent
TabNet	Tabular attention network
TSS	True Skill Statistic
XGBoost	Extreme Gradient Boosting
WP	Work Package

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

Executive summary

This deliverable outlines the creation of high-resolution (1km²) distribution maps for species and habitats across Europe, crucial for biodiversity conservation, policy compliance, and ecosystem management. Employing advanced Species Distribution Models (SDMs) and Habitat Distribution Models (HDMs), the task addressed plants, vertebrates, invertebrates, and all EUNIS Level 3 habitats.

Species distribution modeling involved machine learning algorithms, carefully selected environmental variables, and spatially comprehensive occurrence datasets from GBIF, EVA, and other databases. Ensemble modeling techniques, spatial block cross-validation, and pseudo-absence generation ensured robust, reliable predictions, validated with metrics like True Skill Statistic (TSS).

Habitat modeling similarly utilized environmental predictors (climate, topography, hydrography, geology, soil properties) alongside vegetation plot data from EVA and additional regional databases. Multi-class classification and ensemble forecasting methods provided high-quality predictive habitat maps validated externally and through cross-validation.

Current and future scenarios (2050) were developed under varying climate and land-use trajectories (SSP1-RCP2.6, SSP3-RCP7.0), incorporating model uncertainty and expert-informed constraints. These maps support targeted conservation planning, monitoring programs, and decision-making, guiding efforts to enhance Europe's protected area network and biodiversity management.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

1. Introduction

Understanding the spatial distribution of species and their habitats is a cornerstone of effective biodiversity conservation and ecosystem management. Furthermore, reporting on the distribution of species and habitats is a key component of European policy directives, such as the Birds and Habitats Directive. In Europe, where landscapes are shaped by centuries of human influence and ongoing environmental change, accurate and scalable tools for mapping species distributions are more critical than ever. Species Distribution Models (SDMs, Guisan & Thuiller 2005, Guisan et al. 2017) offer a data-driven approach to predict where species are most likely to occur based on environmental variables, species observations, and ecological theory. These models are widely used in conservation planning, climate change impact assessments, invasive species management, and the prioritization of protected areas (Pollock et al. 2020, Kass et al. 2024).

To translate species-level predictions into actionable conservation outcomes, it is equally important to understand the distribution of suitable habitats, the physical and ecological settings that support these species (Chytrý et al., 2020). Habitat distribution models (hereafter called HDM) complement SDMs by focusing on the spatial extent and suitability (Álvarez-Martínez et al., 2018) of specific habitat types, often defined through vegetation, land cover, or ecological function. These models help identify areas that provide critical resources or ecosystem conditions, offering a broader ecological perspective that includes both species needs and habitat availability. Together, species and habitat modeling form an integrated approach to biodiversity assessment, enabling more targeted and effective conservation planning.

In Task 3.3, the objectives were to predict the current and future distributions of species habitat directives and habitats in Europe at high resolution (1km) and to provide those distributions to the subsequent WPs. We extended the species list to encompass most European vertebrates, a large proportion of plant species, and a representative sample of invertebrate species. For the habitats, we modelled all EUNIS Level 3 habitats across Europe. For both species and habitats, we employed state-of-the art machine learning models to predict their distribution.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

2. Species Distribution Modelling

The species distribution modelling task focused on three taxonomic groups: plants, vertebrates and invertebrates. Leveraging Machine Learning algorithms, we tailored our modelling strategies and environmental variable selections (Table 1. List of environmental covariates used for Species Distribution Models) to the specific characteristics of each taxonomic group.

In the following, we present the general workflow for species distribution modelling, then we dive into the specifics of each group in dedicated sections. We note that the full methodological details and analysis of the results will be described in scientific manuscripts currently under preparation.

2.1 Overview of the general workflow

For the three taxonomic groups (plants, vertebrates, invertebrates), we aimed to model species distributions given a set of relevant and important environmental covariates, at 1km² resolution over Europe, and under current and future environmental conditions. Species distributions are shaped by environmental conditions and spatial accessibility—factors that extend beyond political boundaries. Effective modelling across Europe requires incorporating observations from ecologically connected regions like Russia, North Africa, and the Middle East (Figure 1). We included these areas to ensure accessible habitats and species climatic niches are represented, mitigating for example niche truncation effects (Barbet-Massin et al. 2010, Thuiller et al. 2004).

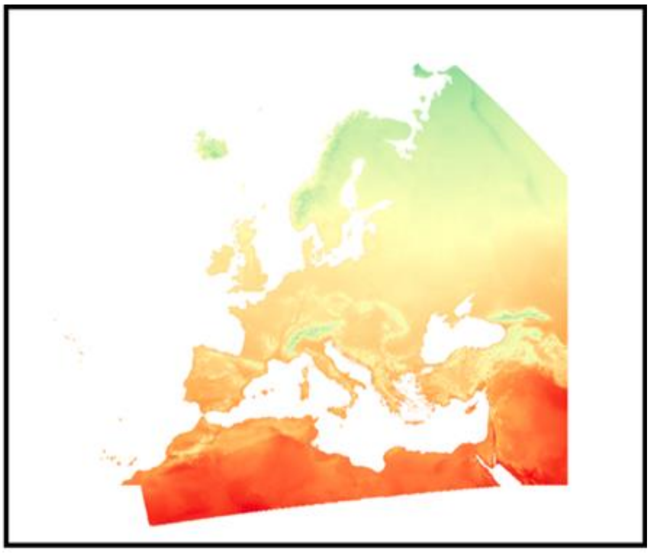

Figure 1. Spatial extent used to calibrate the species distribution models.

Funded by the European Union
NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

For the three species groups, we followed a common workflow in four steps, depicted in Figure

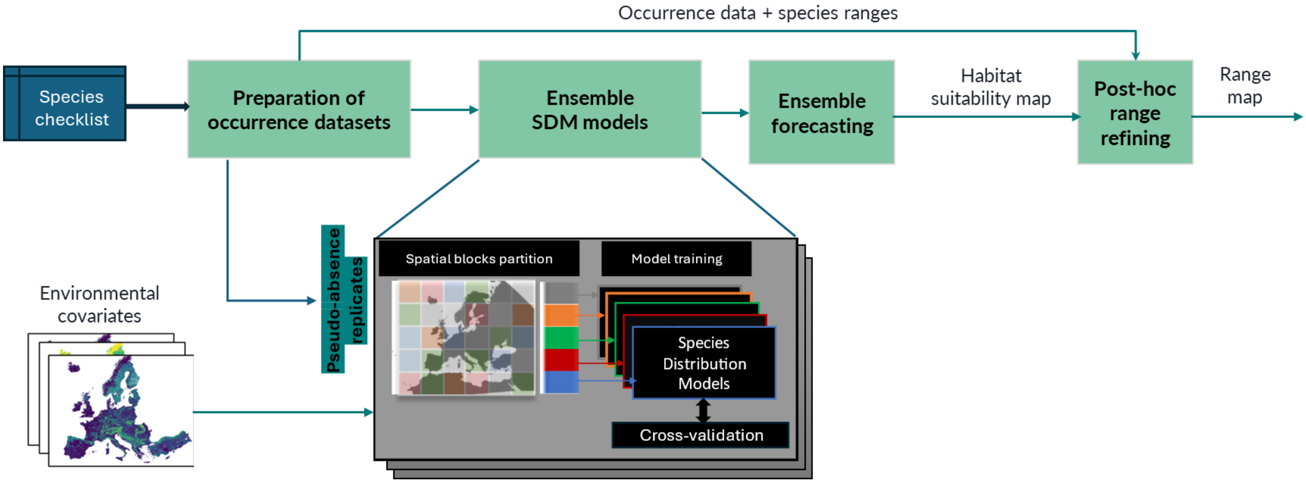

Figure 2 Overview of the general workflow for generating species distribution models (SDMs) to produce species range maps for current and future conditions

Environmental covariates

We selected environmental variables known to influence the distribution of our three taxonomic groups. The selection was made for each group based on literature (plants, vertebrates and invertebrates, Table 1), while we tried to maximize the consistency between the groups. We also made sure to select variables with a minimum correlation between each other (correlation < 0.70). To do so, we collected a set of rasters describing environmental conditions at 1km² resolution over Europe including climate (temperature, precipitation and snow cover trends) from the CHELSA climate database, terrain (slope, landforms, coast proximity) from EarthEnv, soil (physical and chemical properties) from SoilGrids. Additionally, we combined different data sources to produce hydrography layers) covering all our study area and describing water bodies and wetland extent and salinity (Corine land cover, CCI land cover, GLWD), as well as river density (EU-Hydro, HydroSHEDS). Finally, land systems and their use intensity were obtained from Sandstrom *et al.* 2023. The detailed set of predictors selected for each taxonomic group is summarized in Table 1.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

Table 1. List of environmental covariates used for Species Distribution Models

Category	Variable	Source	Time extent	V	P	I
Climate	Annual mean temperature (bio1)	CHELSA	1991-2020			
	Growing degree days (gdd5)					
	Temperature seasonality (bio4)					
	Annual precipitation amount (bio12)					
	Snow cover days (scd)					
	Snow water equivalent (swe)					
	Precipitation seasonality (bio15)					
Hydrology	Coastal area (coastal)	EU Hydro	N.A			
	River density	EU-Hydro, Hydro-SHEDS				
	Water bodies	Sandström et al 2023, GlobCover, GLWD				
	Wetlands					
	Water salinity					
Soil	Soil water pH (pH)	SoilGrids	N.A			
	Cation Exchange Capacity (CEC)					
	Soil organic carbon content in the fine earth fraction (soc)					
	Proportion of sand particles (> 0.05 mm) in the fine earth fraction (sand)					
	Proportion of silt particles (≥ 0.002 mm and ≤ 0.05 mm) in the fine earth fraction (silt)					
Geology	Geological class	GLiM	N.A			
Land Use	Land Cover (classes) in Europe	Sandström et al 2023	2018			
	Land Use Intensity (low-medium-high)					
	Land Cover (classes) outside Europe	CCI land cover 2023	2018			

Occurrence datasets

Starting from a species checklist, for each taxonomic group, occurrence datasets comprising geolocated species occurrences were collected over the study extent. Species presences were aggregated at the grid cell level to avoid pseudo-replication. We used a regular grid with 1km²cells encompassing the study area depicted in Figure 1.

SDMs need to compare the environmental conditions used by species to the available environment in the study area. In another words, since we only had information on species presences, we needed to generate pseudo-absences across Europe to confront where the species are found in respect to what is available. This is a common practice in the SDM domain, and there are various ways to generate them that depend on the sampling bias of the original data (Barbet-Massin et al. 2012). The pseudo-absence sampling strategy was thus

Funded by the European Union
NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

tailored to each taxonomic group (see further explanation for each taxonomic group). To account for the uncertainty associated with the pseudo-absence selection, this procedure was repeated five times. In another words, we have five sets of presence-pseudoabsence for each species of the lists.

Ensemble SDM models

For each occurrence dataset, we applied a spatial block partitioning of the presence/pseudo-absence observations into five subsets. We then performed a five-fold cross-validation, where different machine learning based SDM algorithms were trained on four subsets and tested on the fifth, rotating this process so that each subset was used once as the test set. This approach provides robust estimates of model performance and helps avoid overfitting to spatially clustered data (Roberts et al. 2017).

Combining pseudo-absence repetitions, cross-validation folds and algorithms, we obtained an ensemble of SDMs, each fitted with diverse algorithms and different observations. Only individual spatial distribution models achieving a cross-validated True Skill Statistic (TSS) score above a threshold of 0.4 were retained for the creation of ensemble models using the algorithms described in the next section.

Ensemble forecasting

Once calibrated and validated, SDMs can then be used for predicting the distribution of the species. Outputs are continuous between 0 and 1 and represent the suitability of the environment for the species, the higher the better. Hereafter, we name this continuous output: ‘species environmental suitability’. The advantage of using an ensemble approach is to provide an average across the best models (the ones passing the quality threshold) weighted by their predictive power, an uncertainty estimate across the predictions, and a committee averaging. Here are the following outputs that are delivered in D3.2:

- **Weighted Mean environmental suitability:** Weighted mean of suitability scores excluding low-performing models of the ensemble (TSS<0.4).
- **Uncertainty:** Coefficient of variation of suitability scores excluding low-performing models of the ensemble.
- **Committee averaging:** Average of binary suitability scores excluding low-performing models of the ensemble (TSS<0.4). Binary maps (presence and absence) are obtained for each prediction using a threshold that maximise its predictive ability on the test set.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

ARPHA Preprints Author-formatted document posted on 01/12/2025. DOI: https://doi.org/10.3897/arphapreprints.e180864

Post-hoc range constraints

Environmental suitability maps produced by SDMs often overestimate the suitable area a species could occupy, because they primarily reflect environmental suitability, not actual occupancy. To mitigate this over-prediction, we applied the Estar framework (Hoareau et al., in prep) to limit projections to areas with plausible current occupancy. This Bayesian method allows to constrain outputs from SDMs using either expert range maps like IUCN or BirdLife, or occurrence data from GBIF, or a combination of the two. These pseudo-ranges are then combined with model predictions using the permanence of ratio method (Journel 2002), with equal weighting, to produce constrained probability maps. Final binary maps were derived using a fixed 0.5 probability threshold.

Future scenarios

We projected future species distributions for 2035-2065 (display 2050) using climate and land-use scenarios based on SSP1-RCP2.6 and SSP3-RCP7.0 using 3 different GCMs (GFDL-ESM4, MRI-ESM2-0, UKESM1-0-LL). For Habitat Directive species, all GCMs were used for both SSPs, while for other species, only the three SSP1 GCMs and a single SSP3 GCM (GFDL-ESM4) were used. All other environmental variables were kept constant.

Table 2. Future climate and land-use change scenarios used for model predictions.

Species	**Climate scenario [SSP] [RCP]_[GCM]**	**Shared socio-economic pathways (SSP)**	**Representative Concentration Pathways (RCP)**	**Global Circulation Model (GCM) / Earth System Model (ESM)**	**Land use scenario (LU)**
ALL SPECIES	ssp126_gfdl-esm4	SSP1	RCP2.6	GFDL-ESM4	SSP1 default scenario (SSP1) *Dou et al 2023*
	ssp126_mri-esm2-0			MRI-ESM2-0	
	ssp126_ukesm1-0-ll			UKESM1-0-LL	
	ssp370_gfdl-esm4	SSP3	RCP7.0	GFDL-ESM4	SSP3 default scenario (SSP3) *VU Peter Verburg*
Habitat Directive species only	ssp370_mri-esm2-0			MRI-ESM2-0	
	ssp585_ukesm1-0-ll			UKESM1-0-LL	

Funded by the European Union
NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.
NATURA CONNECT

2.1. Detailed modelling methodology

2.1.1. Plant species distribution models

Species checklist

The species checklist was initially derived from all vascular plant species recorded in vegetation communities from the European Vegetation Archive (EVA, Chytrý et al., 2016). Taxonomic names were then harmonized and matched with the GBIF backbone taxonomy to ensure consistency across all data sources and compatibility with external databases.

Table 3. Taxonomic distribution of modeled plant species

Class	Magnoliopsida	Liliopsida	Polypodiopsida	Pinopsida	Lycopodiopsida	Gnetopsida	Ginkgoopsida
species	10 098	2 134	151	82	21	4	1
Article17	191/510	54/115	10/17	0 /1	4/7	/	/
IUCN	915/1 882	535/952	36/66	33/36	7/24	4/9	/

Occurrence data acquisition and curation

For vascular plants observation, the core dataset consisted of phytosociological plots from the European Vegetation Archive (EVA, Chytrý et al., 2016, *data request project 159*), spanning Europe and adjacent areas in North Africa and Western Asia, to better capture the full ecological niche of plant species at the continental scale. We integrated additional community-level datasets from the Conservatoire Botanique National Alpin (CBNA: 'Relevé exhaustif habitat', 'Relevé phytosociologique', 'Relevé exhaustif géogr.') covering alpine vegetation in France; the DivGrass database (Violle et al., 2015) for French grasslands, the French National Forest Inventory (IFN), and snowbed vegetation survey from the Iberian Peninsula (Illa et al., 2022). We further enriched the dataset with individual species occurrences from the Conservatoire Botanique National Alpin (CBNA) and the Global Biodiversity Information Facility (GBIF, *https://doi.org/10.15468/dl.cswgt4*). To ensure temporal consistency with environmental predictors, we retained only records collected between 1990 and 2022 with a spatial precision of at least 1 km, and excluded all observations located at sea, at institution centroids, or with unreliable coordinates.

All data were projected and aligned with the NaturaConnect 1 km² grid and duplicates across datasets were removed. For modeling, we retained only one presence per species per pixel to minimize spatial pseudo-replication. We also filtered the dataset to retain only species with

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

at least 10 distinct observations across the entire study area, ensuring sufficient data for robust model training.

After curation and taxonomic harmonization, the final dataset includes approximately 1.7 million vegetation plots and 12,468 vascular plant species, totalling over 21 million individual species observations.

Pseudo-absence selection

For each species, we generated a number of pseudo-absences equals to five times the number of presences, with a minimum of 10,000 points to ensure sufficient background contrast for model training. We implemented a stratified approach to pseudo-absence sampling based on the spatial distribution of botanical observation effort.

- 40% of pseudo-absences were sampled within pixels containing community survey data, with the probability of selection proportional to the number of community surveys recorded in the pixel (capped at 100 to limit over-representation of highly sampled areas).
- 40% were sampled within pixels with individual species occurrence data (e.g., GBIF or CBNA occurrences data), weighted by the number of different species observed in each pixel.
- The remaining 20% were drawn randomly from the rest of the study extent, to ensure environmental representativeness beyond sampled zones.

The strong imbalance between presences and pseudo-absences was explicitly addressed during model training. In particular, we ensured an equal weight between presence and pseudo-absence, ensuring a balanced contribution of both classes in the model's loss functions. This approach limits overfitting, even for rare species with very few occurrences.

Species Distribution Modelling

Species distribution models were calibrated using two machine learning algorithms: Random Forests (RF) and XGBoost. Both models were implemented in Python and tuned using grid search cross-validation, optimizing the F1 score. This method ensures the models are strongly regularized and do not overfit the data. Models were trained using up to 500 trees and evaluated using standard performance metrics, specifically by maximizing True Skill Statistics (TSS).

Model validation

The package BlockCV was used to create spatial blocks across the study area ranging from 10 km to 100 km (Valavi et al., 2018). Block size was flexible depending on the number and distribution of the species observations in space,

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

and for species with fewer than 50 presences in total we used random 5-fold cross-validation.

For each species, we built a total of 50 models (2 algorithms × 5 pseudo-absence sets × 5 cross-validation folds) and retained only those with a TSS ≥ 0.4 for ensemble predictions.

Post-hoc spatial constraints

We applied the Estar framework (Hoareau et al., in prep) to spatially constrain current predictions to derive realistic estimates. For plants, we used species occurrences from GBIF at 50 km resolution, along with vegetation community records from EVA, CBNA, DivGrass, IFN, and Snowbed survey, to estimate a coarse pseudo-range for each plant species. For each species, Estar then combined the ensemble prediction, committee averaging, together with the pseudo-range using the permanence of ratio (Journel 2002), with equal weighting, to produce constrained probability maps. Final binary maps were derived using a fixed 0.5 probability threshold.

2.1.2. Vertebrate species distribution models

Species checklist

We selected an initial pool of 1,259 species from the four terrestrial classes: Amphibia, Aves, Mammalia, Reptilia based on the TetraEU database of vertebrate species (Maiorano et al 2020). This species list covers a large proportion of species from the Habitats (Article 17) and Birds (Article 12) Directives as well as the IUCN Red List. Out of this list, 1210 species were retained at the end after data curation.

Table 4. Coverage of the vertebrate species of conservation interest across classes

Class	Amphibia	Aves	Mammalia	Reptilia
#species	117	529	296	268
#Article17	67 /75	/	89 / 142	92 /110
#Article12	/	422/532	/	/

Occurrence datasets

We downloaded species occurrence data from GBIF[i] for each species between January 4th and 9th 2023 including observations between 1980 and 2023 with a location uncertainty of

[i] Aves :
https://doi.org/10.15468/dl.6ptbvk ; https://doi.org/10.15468/dl.6am8ch ;
https://doi.org/10.15468/dl.6q3zpq ; https://doi.org/10.15468/dl.y8nfqw ;

Funded by the European Union
NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.
NATURA CONNECT

1km. Using the R package *CoordinateCleaner* (Zizka et al., 2019), duplicate observations were removed. We also collected data on species ranges from the IUCN for non-bird species and breeding ranges for birds from the BirdLife database.

Additionally, we gathered occurrence data at 10x10km from Article 12 and Article 17 reporting databases of the EU Habitats/Birds Directives for the final evaluation of the species environmental suitability maps (independent evaluation but at a coarser resolution).

To clean and curate the occurrence datasets from GBIF, they were combined with buffered IUCN range maps to filter out unrealistic records (Figure 3). Range polygons were rasterized and expanded using buffers sized to include 90% of outlier GBIF points within each taxonomic class. Species with a high proportion of observations outside their buffered range were manually reviewed and adjusted if necessary. In parallel, sampling effort maps were created by calculating observation density per pixel across target groups: Amphibians, Birds, Small Mammals, Large Mammals, Chiroptera, and Reptilia.

Occurrence datasets were then built using three types of points: certain presences (all GBIF records within the buffered range), certain absences (randomly sampled outside the buffered range), and uncertain absences (sampled within the range, weighted by sampling effort).

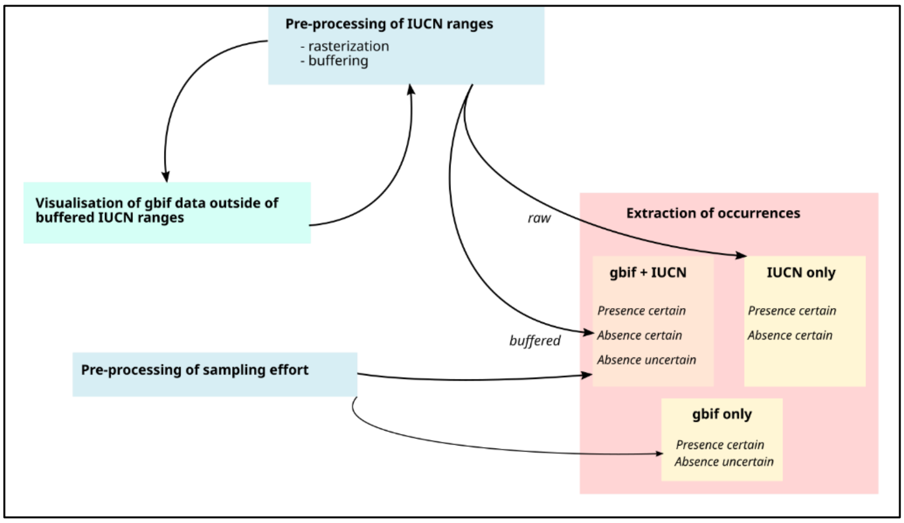

Figure 3. Workflow to generate occurrence datasets for vertebrate SDMs for different data availability settings.

For species lacking GBIF records (328 species) but with available expert maps, datasets were generated by sampling certain presence pixels within IUCN ranges

https://doi.org/10.15468/dl.gdta5n ; https://doi.org/10.15468/dl.tn2b3w ; https://doi.org/10.15468/dl.2bhy85
Amphibia/Mammalia/Reptilia: https://doi.org/10.15468/dl.rxfzz3

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

proportional to range size and certain absence pixels outside the range. Conversely, for species with GBIF records but no expert range (52 species), datasets included all observed presences and pseudo-absences. Species with not enough GBIF records (>20 records) and no IUCN ranges were modeled using habitat directives data (2 species) and dropped otherwise (49 species).

Species distribution models

We used three types of machine learning algorithms to model vertebrate species distributions: Random Forest (RF), extreme gradient boosting (XGBoost) and multi-layer perceptron neural networks (MLP). We used GridSearch via scikit-learn (Pedregosa et al 2011) to automatically tune the hyperparameters of each algorithm. For RF, we tuned tree depth and feature subsampling. For XGBoost, we additionally tuned regularization parameters. For MLP, we tested different architectures with varying depths (number of hidden layers) and widths (number of neurons). Finally, we used class-weighting to correct for imbalance of presence and absence observations particularly for rare species.

Post-hoc spatial constraints

We used Estar to constrain species prediction by using expert range maps only (IUCN, BirdLife). To avoid abrupt cut-offs at IUCN boundaries, we applied an exponential decay from the IUCN range boundaries to the buffer limits defined above for each taxonomic class (see Occurrence dataset).

2.1.3. Invertebrate species distribution models

Species checklist

Due to the large number of invertebrate species in Europe and the lack of occurrence data for many taxa, we selected a subset of taxonomic groups that are representative of European invertebrate diversity. We worked in collaboration with entomologists from the French National case study (the French Biodiversity Office, a stakeholder from Task 8.3) and other international experts to cover a wide range of taxonomic, trophic and functional groups including pollinators, predators, herbivores and parasitoids (**Error! Reference source not found.**).

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

Table 5: Invertebrate taxonomic groups included in the study, number of species and source of occurrence data per group. Subscript numbers correspond to the GBIF download DOI.

Group	Number of species	Source	Median N of observations
Anthophila (bees)	475	$GBIF_3$, SafeGuard	261
Araneae (spiders)	833	$GBIF_1$, observation.org	153
Carabidae (ground beetles)	442	$GBIF_1$, observation.org	173
Collembola (springtails)	188	$GBIF_1$, #GlobalCollembola, Edaphobase	75
Crassiclitellata (earthworms)	37	$GBIF_1$, Edaphobase, Vers2022, Phillips et al. 2021, MINOTAUR	123
Formicidae (ants)	108	$GBIF_1$, observation.org	110
Gastropoda (snails)	257	$GBIF_1$, observation.org	235
Ichneumonoidea (parasitoid wasps)	200	$GBIF_3$, observation.org	56
Lepidoptera (butterflies and moths)	3503	$GBIF_2$, observation.org	237 (moths), 866 (day butterflies)
Odonata (dragonflies)	115	$GBIF_2$, observation.org	1832
Orthoptera (crickets and grasshoppers)	276	$GBIF_2$, observation.org	148
Syrphidae (syrphid flies)	410	$GBIF_2$, SafeGuard	162

1. https://doi.org/10.15468/dl.wzfhjd
2. https://doi.org/10.15468/dl.trbz6u
3. https://doi.org/10.15468/dl.nmbycd

We constituted the invertebrate species checklist for the project based on the species of the above groups that had sufficient observations within our modelling extent (25 or more occurrence points, 25% of the occurrences at least in Europe). The GBIF taxonomic backbone at the time of the download was used as the taxonomy reference for the majority of species. For collembola, following expert advice, we used the checklist provided by collembola.org. For earthworms, we created a consensus taxonomic checklist based on the DriloBASE taxonomy

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

reference and the recent checklist by Misirlioğlu et al. (2023). This checklist was subsequently expertly validated by Mickael Hedde and Sylvain Gerard (INRAE), with additional expertise on specific taxa provided by Daniel Fernández Marchán (Complutense University of Madrid).

Occurrence data acquisition and curation

We used GBIF (https://www.gbif.org/) as the main source of species data (Table 4) with additional observations from scientific databases and citizen-science projects such as observation.org, SafeGuard (for bees and syrphid flies), Edaphobase (for springtails and earthworms), the MINOTAUR and Vers2022 projects (for earthworms), and datasets from Potapov et al. (2024) for springtails and Phillips et al. (2021) for earthworms. To ensure that species and environmental data were collected during similar time periods, we only kept human-made observations between the years 2000 and 2022. Species observations were included only when their precision was at least 1 km, to match the spatial resolution to the limits imposed by the availability of predictor data. We used the R package *CoordinateCleaner* (Zizka et al., 2019) to remove records with wrongly assigned coordinates at sea, institution headquarters or 0 coordinates. We also removed observations from gridded datasets with inaccurate resolution and experimental studies.

We used the NaturaConnect grid (Jung et al 2023) to align all the input data and removed duplicate observations from different source datasets through the R package *CoordinateCleaner* (Zizka et al., 2019). The total number of observations in each pixel by taxonomic group was used to calculate the sampling effort for that group. For the rest of the modelling, we only used the presence of species as one observation per pixel to avoid pseudo-duplication. At the end of the cleaning process and taxonomy checks, we kept 6,844 species with a number of species-specific observations ranging from 25 to 108,717.

Pseudo-absence selection

We generated pseudo-absence points representing the environmental conditions within the study extent. We applied a target-group approach, randomly selecting pseudo-absences for each species with the probability of sampling in a given pixel weighted by the sampling effort of the species' taxonomic group in that pixel. Following Barbet-Massin (2012), we selected 10.000 pseudo-absences for species with up to 5.000 observations, 20.000 for species with up to 10.000 observations,

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

30.000 for species with up to 20.000 observations, and for species with more than 20.000 observations we selected a number of pseudo-absences equal to the observations.

Species Distribution Modelling

We used the Biomod2 platform in R to calibrate the species distribution models (Guéguen et al., 2025). Four different statistical algorithms were used, GLMs, GAMs, Random Forests (RF) and XGBOOST. GLM and GAM formulas were tuned using a stepwise procedure and XGBOOST model parameters were optimized through Biomod2. For RF models a down-sampling procedure was used following (Valavi et al., 2022). For all modeling algorithms, presences and pseudo-absences were weighted equally to mitigate class imbalance.

Model validation

Spatial block cross-validation was used based the BlockCV package (Valavi et al., 2018), with block size adapted to species data availability, and random 5-fold cross-validation used for species with fewer than 50 presences. Overall, we generated for each species five sets of background points and five cross-validation folds per modelling algorithm, for a total of 100 models, to account for the random sampling of background points and cross-validation folds. We used the True Skill Statistic (TSS) and the Boyce Index as model validation metrics. Models were included in the ensemble only if they had a TSS score of at minimum 0.4. When all the models for a species had high TSS scores, we removed models that had a TSS score that was two standard deviations lower than the mean across all species models.

Post-hoc spatial constraints

Predicted current distributions were constrained using the Estar framework (Hoareau et al., in prep) to remove areas with projected suitable habitat for species that are not actually occupied. We used species observations from GBIF at 50 km resolution to generate a more comprehensive map of species occurrence and taxonomic group sampling effort at coarse scale. These observations were used to generate a pseudo-range of the species based on the ratio between the species frequency of observation and the background sampling effort of the taxonomic group. This pseudo-range was then combined with the committee averaging current projection and weighed equally to create a constrained spatial range. Binary constrained maps were built using a cutoff of 0.5 probability.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

2.3 Model Quality Assessment

Given the total number of species, algorithms and runs, we hereafter only provided a snapshot of model predictive accuracy across the different groups (Figure 4). We indeed represented the TSS values of the predictions (across the algorithms, pseudo-absence selections, and 5 –fold CV that were kept for the ensemble predictions, TSS >0.4) for the selected class of the broad class taxonomic groups. For most groups and classes, predictions reached very good performances with median TSS near 0.75 for plants, 0.9 for vertebrates and 0.75 for invertebrates (Figure 4). These high values support the use of the predictions for the subsequent WPs of the NaturaConnect project.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

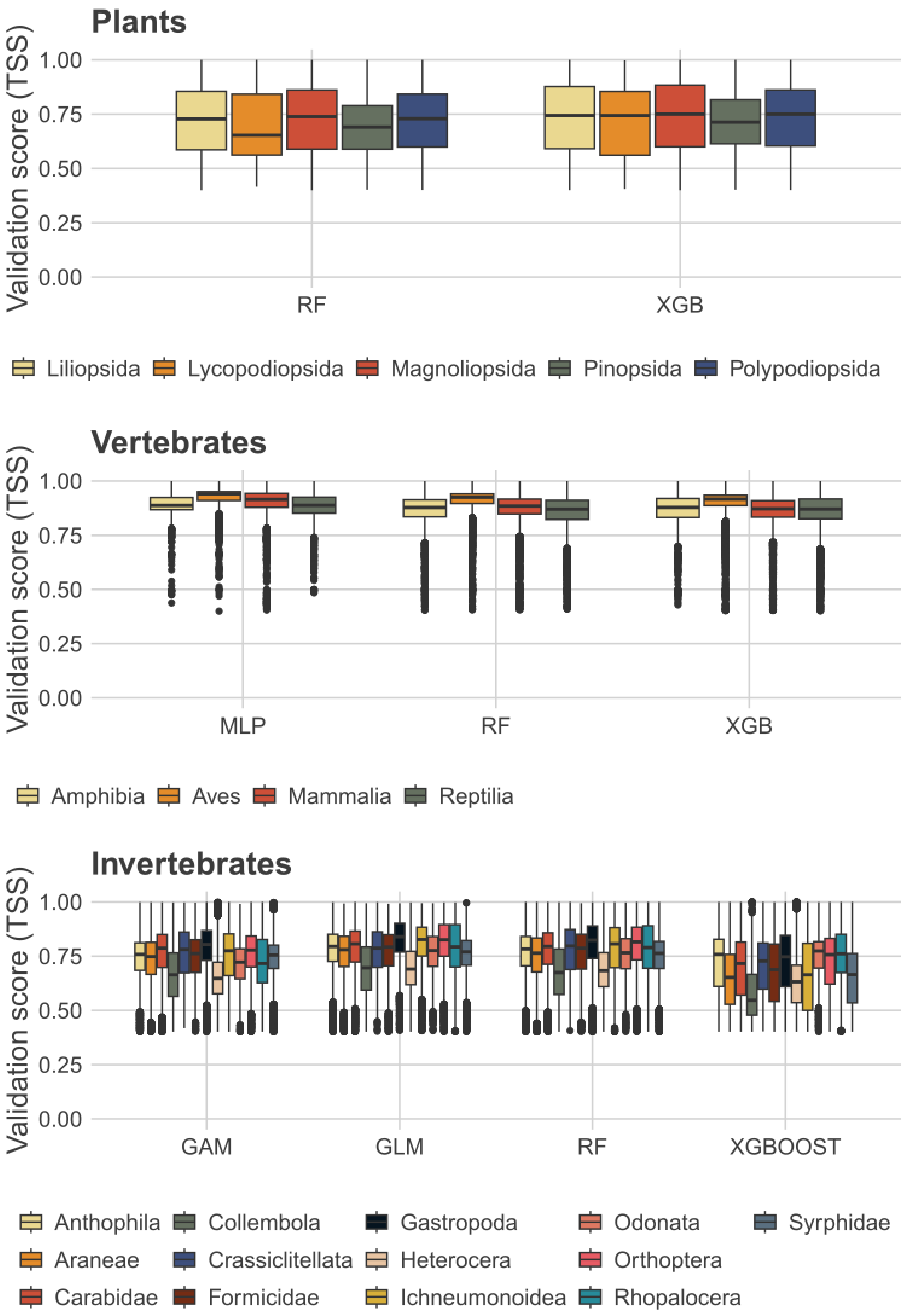

Figure 4. Single-model validation scores (TSS metric) for each kingdom, across different algorithms and taxonomic groups.

Funded by the European Union

NatureConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

3. Habitat distribution models

Similar to SDM for species, habitat distribution modelling (HDM) focuses on predicting the extent and suitability of habitats. As a thematic legend, we used the EUNIS habitat nomenclature (Chytrỳ, et al., 2020), which organises habitat classes into hierarchical levels. At EUNIS level 3, over 200 distinct habitat classes are identified, and grouped into nine broader formations at EUNIS level 1 including MA2 (Saltmarshes), N (Coastal habitats), Q (Wetlands), R (Grasslands), S (Shrublands - heathlands, scrub, and tundra), T (Forests), U (Sparsely vegetated habitats), and V (Man-made habitats).

3.1 Habitat observations

Training data

Vegetation observation plots[ii] from the European Vegetation Archive (EVA, https://euroveg.org/eva-database) served as ground truth data for training and testing the model across Europe. Each observation plot was translated to the EUNIS typology at level 3 based on its species composition using the EUNIS-ESy expert system (Chytrý et al 2020). Plots that were reported with no cover-abundance information for individual species were excluded. Further, plots smaller than 1 m^2 or larger than 1000 m^2, without geographical coordinates and those with reported uncertainty of the coordinates larger than 1 km were also excluded. The plots with a missing indication of location uncertainty were retained, assuming most of them were within 1 km from the indicated coordinates. The resulting dataset contained a total of about 1,2 million georeferenced plots.

Validation data

To validate the HDM predictions, we gather independent habitat-annotated vegetation plot databases, including 80k plots from the Landelijke Vegetatie Database (NL) (Hennekens et al 2018) and 180k forest plots from the French Forest Inventory (IFN) (IGN 2018), as well as regional habitat maps of Austria (AT) (Umweltbundesamt et al 2021).

[ii] EVA project # 217 – 2024-08-01: https://doi.org/10.58060/6n57-hh22

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

3.2 Environmental predictors

To train the models on the vegetation plots, we built a comprehensive database of ecologically meaningful environmental predictors at the highest resolution. Those variables had to be available across Europe at least at 1 km resolution. We selected a set of poorly correlated climatic variables, topography, hydrography, geology, and soil properties (Table). Additionally, we used land use maps to further refine habitat distribution maps.

Table 6. List of environmental predictors used for habitat distribution models

Type	Predictor	Data source
Climate	BIO1: Annual mean temperature (°C)	CHELSA V2.1 Resolution: 1km Temporal range: 1981-2010 Spatial range: Europe (Karger et al 2017)
	BIO4: Temperature seasonality	
	GDD5: Growing degree days heat sum above 5°C (°C)	
	BIO12: Annual sum of precipitations (kg.m²/yr)	
	BIO15: Precipitation seasonality	
	SCD: Snow covered days (day count)	
	SWE: Snow water equivalent (kg.m²/yr)	
Topography	EU DEM slope (degrees)	EU DEM v1.1 Resolution: 100m
	EU DEM aspect (degrees)	
	Landform classification based on the topographic position index. (10 classes)	EU-DEM v.1.1., upscaled to 100m, on SAGA-GIS. Resolution: 100m
Hydrography	Distance to inland water (m)	EU-Hydro river network Resolution: 100m
	Distance to sea coastline (m)	EU-Hydro coast-line Resolution: 100m
Geology	Dominant parent material class (27 classes)	European soil database (ESDB) Resolution: 1km (Panagos et al 2006)
	Depth to rock (m)	
Soil	Available Water Capacity (AWC) for the topsoil fine earth fraction	LUCAS topsoil physical properties Resolution: 500m
	Bulk density of the topsoil (kg/dm^3)	
	Coarse fragment (%) content in topsoil	

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

	Sand, silt, and clay proportions (%)	(Ballabio et al 2016)
	Soil acidity	LUCAS topsoil chemical properties Resolution: 500m (Ballabio et al 2019)
	Organic carbon content (g/kg)	
	Nitrogen content (g/kg)	
	Calcium carbonates (g/kg)	
	Cation exchange capacity (cmol/kg)	

3.3 Modeling framework

To handle the discrete nature of habitat classes, we employed a multi-class classification approach to accurately assign habitat types to specific locations based on the provided predictors (Figure 5). By jointly modelling all classes, multi-class models automatically consider associations between habitats and benefit rare habitats with limited observations. However, the co-occurrence of habitats from different formations (EUNIS level 1) can lead to habitat complexities or mosaics, making data ambiguous and potentially reducing the model's accuracy in predicting the correct habitat class.

To address the ambiguity, we leveraged the hierarchical structure of habitat classes by training separate multi-class models at level 3 for habitats within each EUNIS level 1 group. We employed an ensemble approach to address uncertainties arising from model choice and data sampling.

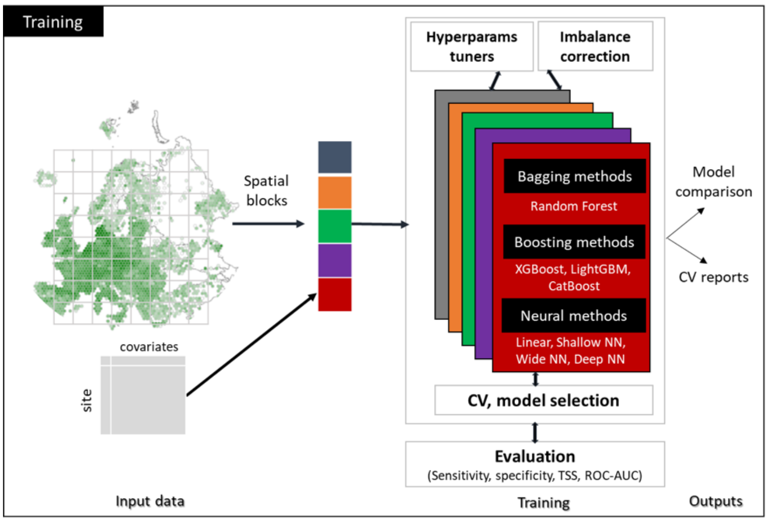

Figure 5. Habitat modeling framework overview for model training.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

3.3.1 Data partitioning

We considered the uncertainty due to training data sampling. Employing 5-fold cross-validation, we trained each algorithm with 20% of the observations hidden from each iteration. This process generated an ensemble of 5 classifiers per algorithm with access to distinct samples, thus accounting for data sampling uncertainty. The 5-fold partition of observations was obtained through a spatial block strategy that avoids overestimation of model performances due to data leakage between training and validation sets while allowing the models to capture local variability between nearby observations.

3.3.2 Machine learning models

Algorithm choice

To encompass this diversity, we created an ensemble of algorithms from different families including bagging decision trees, boosted decision trees and neural networks.

In the bagging family, we evaluated the Random Forest (Breiman et al 2001) algorithm whereas in the boosting family we evaluated three algorithms: XGBoost (Chen and Guestrin, 2016), CatBoost (Prokhorenkova et al., 2018) and LightGBM (Ke et al 2017). In the neural networks (LeCun et al., 2015) family, we evaluated Multi-Layer Peceptrons (MLP) architectures with varying widths and depths from zero (linear models), to few (shallow neural networks) or many hidden layers (deep neural networks), as well as tabular attention networks (TabNet) (Arik & Pfister et al 2021).

Training

We optimized the hyperparameters of each algorithm, and we applied appropriate imbalance correction techniques tailored to each algorithm to ensure optimal performances. Specifically, we used class weighted multi-class objectives for bagging and boosting trees whereas we trained neural networks with imbalance-aware optimization objectives: Focal Loss (Lin et al 2017) and Label Distribution Aware Margin loss (Cao et al 2019)).

Model selection criteria

We used macro-averaged class-wise precision, recall, and F1-score from cross-validation to select the best-performing model within each algorithm family.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

Precision measures how many predicted habitats were correct, recall assesses how well actual habitats were identified, and F1-score balances both metrics.

3.4 Generation of habitat maps

For a given scenario, the ensemble multi-class habitat models were used to predict the relative habitat probabilities given environmental layers. Habitat probabilities combined with a set of decision rules were then used to generate wall-to-wall habitat maps for Europe, following the workflow depicted in Figure 6.

3.4.1 Ensemble forecasting

After training, we obtained a set of models for each algorithm–data partition combination. By combining the relative habitat class probabilities predicted by each model, we generated two ensemble outputs:

(i) **Ensemble mean**: average habitat probability weighted by model quality (soft averaging);

(ii) **Consensus:** the proportion of models voting (attributing highest probability) for each habitat class (committee averaging).

3.4.2 Habitat mapping workflow

The habitat mapping workflow generates three types of output maps at 1km resolution:

1. **Habitat probability maps**: continuous maps of all EUNIS level 3 class probabilities (250+ maps).
2. **Dominant habitat maps**: categorical maps of the most likely (highest mean probability) habitats at level 3 within each formation (9 maps).
3. **Wall-to-wall habitat map:** categorical map of the top 3 EUNIS habitats at level 3 across all formations (previewed in **Error! Reference source not found.**), namely w all-to-wall habitat map. To generate this map, we apply cross walk rules (EEA, ETC-BD) to convert the land cover map into an EUNIS level 1 map. Afterwards, at each pixel, given its EUNIS 1 class, we map the dominant level 3 habitat class.

Dominant and wall-to-wall habitat maps are associated with confidence (ensemble mean probability of the most likely class) and consensus maps.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

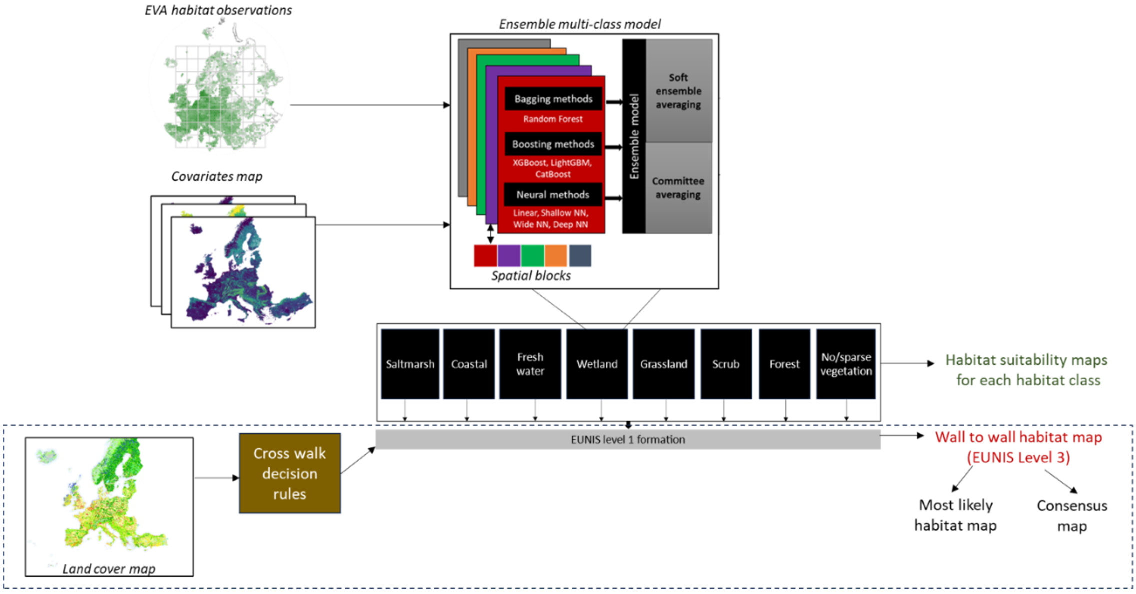

Figure 6. Workflow for the modeling and generation of habitat distribution maps.

3.5 Validation

The habitat maps showed strong predictive performance in cross-validation at the European scale (EVA), with most EUNIS level 3 classes achieving good F1-scores. Accuracy varied by habitat formation, with consistently high performance in abiotic-driven types like wetlands and coastal areas, and more variable results in structurally complex habitats like forests and grasslands. External validation in the Netherlands (NL), Austria (AT), Portugal (PT) and French forests (IFN) confirmed overall robustness, though performance dropped slightly, especially in Austria due to resolution mismatches. Trade-offs between precision and recall were observed across formations, and prediction improved when considering the top three habitat classes, especially for rare or narrowly distributed habitats.

Table 7. Evaluation scores for the habitat distribution modeling based on habitat formations.

Habitat formation	Dataset	F1-score	Precision	Recall
MA2 (Saltmarsh)	EVA	0.87 ± 0.11	0.82 ± 0.18	0.95 ± 0.06
	NL	0.76 ± 0.34	0.74 ± 0.33	0.80 ± 0.36
N (Coastal)	EVA	0.81 ± 0.09	0.75 ± 0.15	0.92 ± 0.07
	NL	0.82 ± 0.09	0.84 ± 0.06	0.83 ± 0.18
P (Freshwater)	EVA	0.87 ± 0.07	0.82 ± 0.12	0.94 ± 0.05
	NL	0.97 ± 0.02	0.96 ± 0.03	0.99 ± 0.01
Q	EVA	0.73 ± 0.13	0.67 ± 0.19	0.86 ± 0.09

Funded by the European Union
NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

(Wetland)	NL	0.95 ± 0.06	0.95 ± 0.08	0.96 ± 0.07
	AT	0.56 ± 0.46	0.55 ± 0.45	0.57 ± 0.46
R (Grassland)	EVA	0.66 ± 0.17	0.78 ± 0.15	0.59 ± 0.18
	NL	0.40 ± 0.32	0.57 ± 0.42	0.37 ± 0.33
	AT	0.37 ± 0.35	0.45 ± 0.39	0.37 ± 0.36
S (Scrub and Tundra)	EVA	0.83 ± 0.12	0.89 ± 0.06	0.79 ± 0.15
	NL	0.46 ± 0.39	0.49 ± 0.41	0.46 ± 0.41
	AT	0.31 ± 0.42	0.31 ± 0.42	0.31 ± 0.43
	PT	0.33 ± 0.39	0.41 ± 0.46	0.32 ± 0.41
T (Forest)	EVA	0.61 ± 0.20	0.77 ± 0.15	0.55 ± 0.23
	IFN	0.48 ± 0.32	0.78 ± 0.32	0.40 ± 0.33
	NL	0.38 ± 0.31	0.47 ± 0.36	0.37 ± 0.35
	AT	0.33 ± 0.39	0.34 ± 0.41	0.32 ± 0.39
	PT	0.39 ± 0.40	0.49 ± 0.41	0.36 ± 0.39
U (Sparsely vegetated)	EVA	0.94 ± 0.03	0.95 ± 0.03	0.94 ± 0.03
	NL	0.81 ± 0.22	0.96 ± 0.05	0.78 ± 0.31
	AT	0.35 ± 0.40	0.34 ± 0.39	0.36 ± 0.42

3.6 Habitat mapping under current and future conditions

Building on the previous workflow, we produced habitat probability and dominant habitat maps under current conditions using current climate (1990-2020) layers (Karger et al 2017) alongside other environmental layers. The current wall-to-wall map was then built using the land system map from (Dou et al 2021).

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

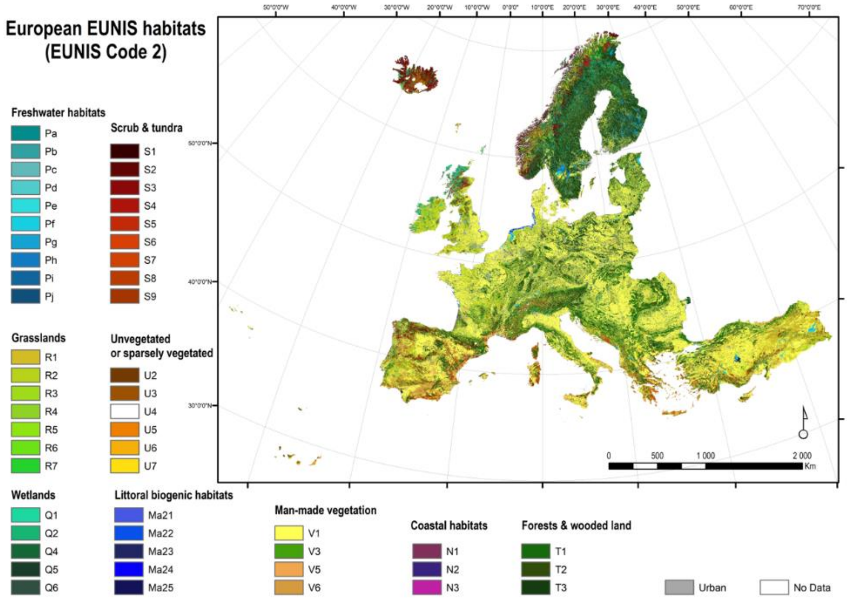

Figure 7. Wall-to-wall European map of EUNIS habitats. Each pixel represents the most likely EUNIS class at level 3. Due to the high number of classes, the color legend is provided at level 2

.

To forecast future habitat distributions, we assumed all environmental predictors remained constant except for climate. Habitat probability and dominant habitat maps were forecasted under two future climate scenarios: SSP1 (RCP 2.6) and SSP3 (RCP 7.0) across 3 Global Circulation Models (GCMs).

Finally, we produced a series of future wall-to-wall maps under varying land-use trajectories, including baseline socioeconomic pathways (SSP1 and SSP3) and nature-positive scenarios aligned with the Nature Futures Framework (NFF): Nature for Nature (NfN), Nature for Society (NfS), and Nature as Culture (NaC).

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

4. Applications of species and habitat maps

Species and habitat maps offer essential tools for advancing conservation and guiding ecological restoration efforts. Hereafter, we discuss some potential use cases and applications within the NaturaConnect project and beyond.

4.1 Potential users of habitat and species distribution maps

Species and habitat distribution maps are powerful tools that support informed decision-making across a range of sectors. Below are several examples of how different stakeholder groups can apply these models.

Local authorities and municipalities can use species and habitat distribution maps to gain an overview of the biodiversity currently present within their jurisdiction and how it may change under future climate and land use trajectories. This information is valuable for integrating nature into urban and regional planning, identifying priority areas for green infrastructures, and developing strategies for biodiversity conservation within city boundaries. These tools can help ensure that local development aligns with long-term ecological sustainability goals.

Protected area managers can apply these models to improve conservation planning at multiple scales. Protected area managers, for example, can incorporate future species assemblage projections into site-level conservation plans, ensuring management strategies remain relevant under climate change. Species distribution maps can also highlight species likely present in a protected area but not yet recorded, guiding targeted monitoring and inventory efforts (e.g. Guisan et al. 2006).

National-level authorities can use model outputs to update species and habitat lists of conservation concern, reflecting projected changes due to climate and land-use shifts. They can also identify species that are not currently threatened but are expected to experience range contraction, supporting the development of an "orange list" to enable proactive protection before population declines occur. Furthermore, national agencies can use these insights to adapt the design and effectiveness of protected area networks.

Non-governmental organizations (NGOs) engaged in conservation and restoration projects can benefit from distribution models to plan their work more effectively. These maps can inform site selection and guide the design of habitat restoration or species reintroduction efforts by identifying which species are best suited to current and future environmental conditions. For example, reforestation projects can be aligned with future climate projections to ensure long-term viability of the selected tree species. Distribution models also support monitoring and

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

impact assessment by highlighting areas where conservation interventions may yield the greatest ecological benefit.

Land-use sectors and developers—including agriculture, energy, transport, forestry, and urban planning—can use habitat and species distribution maps to incorporate biodiversity considerations into project planning and environmental impact assessments. These tools can help identify areas of high ecological value, or red zones, where development should be avoided due to potential negative impacts on sensitive habitats or species. Conversely, they can also highlight green zones where development is less likely to harm biodiversity, allowing planners to prioritize expansion in a way that minimizes environmental conflict.

4.2 Complementarity of habitat distribution maps

Habitat probability maps indicate the likelihood of occurrence for various habitat types, guiding targeted restoration and protection actions. Dominant habitat maps highlight the most likely habitat within each ecological group, making them especially useful for planning large-scale restoration initiatives. Wall-to-wall habitat maps offer detailed insights into the extent and condition of habitats, supporting the analysis of structural ecosystem connectivity. Finally, future habitat maps project how habitats may shift under different climate and land-use scenarios, enabling anticipatory planning, identification of at-risk areas, and prioritization of monitoring and intervention efforts.

4.3 Advancing ecological knowledge

Habitat and species maps across trophic levels (plants, invertebrates, vertebrates) provide a comprehensive characterization of ecosystems and can be combined to derive various community-level and ecosystem-level indicators.

Reconstructing ecological networks at large scale

With growing ecological knowledge about species' resource needs and typical interactions—such as trophic links, pollination, parasitism, and mutualism—we can use species and habitat distribution maps to spatially reconstruct ecological networks. By mapping where interacting species (e.g., host and parasite, plant and pollinator, predator and prey) are likely to co-occur, we can infer the presence and structure of interaction networks across regions and ecosystem types (O'Connor et al. 2020). Analyzing this spatial information under various scenarios would highlight entire networks may shift or fragment under changing environmental conditions (Albouy et al. 2014). This approach provides a powerful tool for assessing ecosystem functionality, identifying areas of high interaction richness, and prioritizing regions where conservation actions can help maintain or restore critical ecological functions.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

Mapping other facets of biodiversity

In addition to mapping ecological networks and their diversity, species maps can be combined with trait and phylogeny databases to derive measures of functional and phylogenetic diversity for both current and future conditions (Gaüzere et al. 2022). This could highlight areas that could lose or gain higher functions than expected by change in species richness (Gaüzere et al. 2023), for instance, areas that could become cradles or museums of life.

Quantifying the supply of ecosystem services

Species and habitat maps provide a valuable basis for quantifying ecosystem services by linking biodiversity patterns to the functions and benefits ecosystems provide (Schulp et al. 2014). For example, the presence and distribution of pollinators, seed dispersers, or pest-controlling species can be used to estimate the potential for regulating services like crop pollination or natural pest suppression (Kass et al. 2024). Similarly, habitat types mapped at fine resolution can be associated with supporting and regulating services such as carbon storage, water purification, or soil retention, based on known ecological functions of those habitats.

4.4 Supporting conservation

Assessing / updating the conservation status of species and habitats

Assessing the conservation status of species and habitats is essential for evaluating the effectiveness of current protection measures. Habitat and species maps can help determine how well existing protected area networks safeguard vulnerable biodiversity, in line with frameworks such as the IUCN Red List and the EU Habitats and Birds Directives (Pollock et al. 2020, Guisan et al. 2016). These tools also support the development of forward-looking assessments, such as an "orange list" of species that may not yet be threatened but face emerging risks from climate change, land-use change, or other pressures—enabling proactive conservation planning before declines become critical.

Assessing and optimizing habitat connectivity

High-resolution habitat maps are essential for assessing structural connectivity, as they reveal the location, fragmentation, and continuity of natural or semi-natural habitats, as well as potential barriers or gaps. Analyzing the spatial configuration of these habitats helps identify potential ecological corridors—stretches of habitat that facilitate movement between core areas—as well as bottlenecks or breaks in connectivity that may isolate populations or disrupt ecological flows.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

Building on structural data, functional connectivity can be inferred by integrating species-specific information such as dispersal distances, habitat preferences, and known movement patterns (Prima et al. 2024). This enables more realistic modeling of how different species interact with the landscape and highlights where conservation actions—such as habitat restoration, steppingstones, or green infrastructure—would be most effective in enhancing connectivity.

Spatial conservation planning

The modelled spatial distribution of species and habitats, in the form of spatial maps, will be integrated into the NaturaConnect Prioritzr software package alongside other data inputs to create variants of conservation planning options for implementing the TEN-N. These variants can be used by decision-makers to identify areas of conservation priority as potential additions to the current protected area network under consideration of a range of other values (e.g. areas that prioritise threatened species, while achieving an acceptable balance with other objectives such as impacts on other land-uses or recreation potential).

Funded by the European Union
NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

5. References

Albouy, C., Velez, Coll, M., Colloca, F., Le Loc'h, F., Mouillot, D., & Gravel, D. (2014). From projected species distribution to food-web structure under climate change. Global Change Biology, 20, 730-41.

Álvarez-Martínez, J. M., Jiménez-Alfaro, B., Barquín, J., Ondiviela, B., Recio, M., Silió-Calzada, A., & Juanes, J. A. (2018). Modelling the area of occupancy of habitat types with remote sensing. Methods in Ecology and Evolution, 9(3), 580-593.

Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6679-6687. https://doi.org/10.1609/aaai.v35i8.16826

Ballabio, C., Panagos, P., & Monatanarella, L. (2016). Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma, 261, 110-123.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., ... & Panagos, P. (2019). Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. Geoderma, 355, 113912.

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012) Selecting pseudo-absence for species distribution models: how, where and how many? Methods in Ecology and Evolution, 3, 327-338

Barbet-Massin, M., Thuiller, W., & Jiguet, F. (2014) How much do we overestimate local extinction rates when restricting the range of occurrence data in climate suitability models? Ecography. 33(5), 378-386

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems, 32.

Chen, T. & Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *In* KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785 – 794, https://doi.org/10.1145/2939672.2939785.

Chytrý, M., Tichý, L., Hennekens, S. M., Knollová, I., Janssen, J. A. M., Rodwell, J. S., Peterka, T., Marcenò, C., Landucci, F., Danihelka, J., Hájek, M., Dengler, J., Novák, P., Zukal, D., Jiménez-Alfaro, B., Mucina, L., Abdulhak, S., Abramova, L., Aćić, S., … Schaminée, J. H. J. (2021). EUNIS-ESy: Expert system for automatic classification of European vegetation plots to EUNIS habitats (v2021-06-01) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4812736

Chytrý M., Hennekens S.M., Jiménez-Alfaro B., Knollová I., Dengler J., Jansen F., Landucci F., Schaminée J.H.J., Aćić S., Agrillo E., Ambarlı D., Angelini P., Apostolova I., Attorre F., Berg C., Bergmeier E., Biurrun I., Botta-Dukát Z., Brisse H., Campos J.A., Carlón L., ... Yamalov S. 2016. European Vegetation Archive (EVA): an integrated database of European vegetation plots. *Applied Vegetation Science* 19: 173–180. https://doi.org/10.1111/avsc.12191

Copernicus Land Monitoring Service, European Environment Agency (EEA). (2016). EU-DEM v1.1.

Dou, Y., Cosentino, F., Malek, Z., Maiorano, L., Thuiller, W., & Verburg, P. H. (2021). A new European land systems representation accounting for landscape characteristics. Landscape Ecology, 36, 2215-2234.

European Environment Agency (EEA). (2019). EU-Hydro – River Network Database.

Gaüzere, P., O'Connor, L., Botella, C., Poggiato, G., Münkemüller, T., Pollock, L.J. Brose, U., Maiorano, L., Harfoot, M.H. and Thuiller, W. (2022) The diversity of interactions complements functional and phylogenetic facets of biodiversity. Current Biology, 32(9), 2093-2100

Gaüzère, P., Botella, C., Poggiato, G., O'Connor, L.M.J., Di Marco, M., Dragonetti, C., Maiorano, L., Renaud, J. & Thuiller, W. (2023) Dissimilarity of vertebrate trophic interactions reveals spatial uniqueness but functional redundancy across Europe. Current Biology, 33(23), 5263-5271.e3

Guéguen M, Blancheteau H, Thuiller W (2025). *biomod2: Ensemble Platform for Species Distribution Modeling*. R package version 4.3-2-2, https://biomodhub.github.io/biomod2/.

Guisan, A. and Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. Ecology Letters 8, 993-1009

Guisan, A., O. Broennimann, R. Engler, N. G. Yoccoz, M. Vust, N. E. Zimmermann, et A. Lehmann. (2006). Using niche-based models to improve the sampling of rare species. Conservation Biology, 20, 501-11

Guisan, A, Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., et al. (2013). Predicting species distributions for conservation decisions. Ecology Letters, 16, 1424-35.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

Guisan, A., Thuiller, W. and Zimmermann, N.E. (2017) Habitat Suitability and Distribution Models - with Applications in R. Ecology, Conservation and Biodiversity series. Cambridge University Press.

Hennekens, S. NDFF Dutch vegetation database. Dutch National Database of Flora and Fauna (NDFF) https://doi.org/10.15468/KSQXEP (2018).

Hoareau, M., Si-Moussi, S. & Thuiller, W. A Bayesian framework to spatially constrain habitat suitability maps from species distribution models. In prep.

IGN – Inventaire Forestier National Français. Données brutes, Campagnes annuelles 2005 et suivantes. (2022).

Illa, E., Pérez-Haase, A., Brufau, R., & Font, X. (2022). Living on the edge: Plant diversity in the Iberian chionophilous vegetation. *Applied Vegetation Science*, *25*(4), e12701.

Journel, A.G. (2002). Combining Knowledge from Diverse Sources: An Alternative to Traditional Data Independence Hypotheses. Mathematical Geology, 34, 573–596

Jung, M. (2023). Standardized reference grids for spatial analyses at various grain sizes (1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7971151

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., ... & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. Scientific data, 4(1), 1-20.

Kass, J.M., Fukaya, K., Thuiller, W. & Mori, A.S. (2024) Biodiversity modeling advances will improve predictions of nature's contributions to people. Trends in Ecology & Evolution, 39, 338–348

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

Misirlioğlu, M., Reynolds, J., Stojanović, M., Trakić, T., Sekulić, J., James, S., ... & Brown, G. (2023). Earthworms (Clitellata, Megadrili) of the world: an updated checklist of valid species and families, with notes on their distribution.

O'Connor, L.M.J, Pollock, L.J., Braga, J., Ficetola, G.F., Maiorano, L., Martinez-Almoyna, C., Montemaggiori, A., Ohlmann, M. & Thuiller, W. (2020) Unveiling the food webs of tetrapods across Europe through the prism of the Eltonian niche. Journal of Biogeography, 47, 181–192

Panagos, P. (2006). The European soil database. GEO: connexion, 5(7), 32-33.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Phillips, H. R., Bach, E. M., Bartz, M. L., Bennett, J. M., Beugnon, R., Briones, M. J., ... & Webster, E. R. (2021). Global data on earthworm abundance, biomass, diversity and corresponding environmental properties. *Scientific data*, *8*(1), 136.

Pollock, L.J., O'Connor, L.M.J., Mokany, K., Rosauer, D.F., Talluto, M.V. & Thuiller, W. (2020) Protecting biodiversity (in all its complexity): new models and methods. Trends in Ecology and Evolution, 35(12), 1119-1128

Potapov, A. M., Chen, T. W., Striuchkova, A. V., Alatalo, J. M., Alexandre, D., Arbea, J., ... & Scheu, S. (2024). Global fine-resolution data on springtail abundance and community structure. *Scientific data*, *11*(1), 22.

Prima, M.C., Renaud, J., Witté, I., Suarez, L., Rouveyrol,Fernando, M.,Sacchi, A., Cosentino, F., Santini, L., Maiorano, L., Moreira, F., Dertien, J., Fernandez, N., & Thuiller, W. (2024) A comprehensive framework to assess multi-species landscape connectivity for conservation planning. Methods in Ecology and Evolution, 15(12), 2385-2399

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

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

Sandström, Evelina; Namasivayam, Anandi; Oostdijk, Saskia; Scherpenhuijzen, Niek; Debonne, Niels; Verburg, Peter, 2023, "Preliminary land system map for Europe", https://doi.org/10.34894/THARMK, DataverseNL, V5

Schulp, C.J.E., Thuiller W., Verburg, P.H. (2014) Wild food in Europe: a synthesis of knowledge and data of terrestrial wild food as an ecosystem service. Ecological Economics. 105, 292-305

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

NATURA CONNECT

Thuiller, W., Brotons L., Araújo M.B., Lavorel S. (2004) Effects of restricting environmental range of data to project current and future species distributions. Ecography, 27, 165-172

Umweltbundesamt. MAES/EUNIS habitat map Austria 10m. 56.1 MBytes PANGAEA https://doi.org/10.1594/PANGAEA.934147 (2021).

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Biorxiv*, 357798.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological monographs*, *92*(1), e01486.

Violle, C., Choler, P., Borgy, B., Garnier, E., Amiaud, B., Debarros, G., ... & Viovy, N. (2015). Vegetation ecology meets ecosystem science: Permanent grasslands as a functional biogeography case study. *Science of the Total Environment*, *534*, 43-51.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., ... & Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*(5), 744-751.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

ARPHA Preprints Author-formatted document posted on 01/12/2025. DOI: https://doi.org/10.3897/arphapreprints.e180864

6. Data availability

Habitat maps can be found on Zenodo, including high resolution habitat maps for current conditions (100m) here: https://doi.org/10.5281/zenodo.11108225; as well as current and future habitat maps at 1km resolution here: https://doi.org/10.5281/zenodo.15307414

In the long term, species data will all be available on the EBV data portal. Vertebrate maps can already be previewed and accessed via the portal at: https://doi.org/10.25829/wpfn43

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.

ARPHA Preprints Author-formatted document posted on 01/12/2025. DOI: https://doi.org/10.3897/arphapreprints.e180864

More information about the project:

NaturaConnect has 22 partner institutions: International Institute for Applied System Analysis (project lead; Austria); German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig (project co-lead; Germany); Associacao Biopolis (Portugal); BirdLife Europe (Netherlands); Birdlife International (United Kingdom); Centre National De La Recherche Scientifique (France); Doñana Research Station - Agencia Estatal Consejo Superior De Ivestigaciones Cientificas (Spain); Europarc Federation (Germany); Finnish Environment Institute (Finland); Humboldt-University of Berlin (Germany); Institute for European Environmental Policy (Belgium); Netherlands Environmental Assessment Agency (Netherlands); Rewilding Europe (Netherlands); University of Evora (Portugal); University of Helsinki (Finland); University of Natural Resources and Life Sciences, Vienna (Austria); University of Rome La Sapienza (Italy); University of Warsaw (Poland); Vrie University of Amsterdam (Netherlands); WWF Central and Eastern Europe (Austria); WWF Romania and WWF Hungary.

NaturaConnect aims to design and develop a blueprint for a truly coherent **Trans-European Nature Network** (TEN-N) of conserved areas that protect at least 30% of land in the European Union, with at least one third of it under strict protection. Our project unites universities and research institutes, government bodies and non-governmental organizations, working together with key stakeholders to create targeted knowledge and tools, and build the capacity needed to support European Union Member States in realizing an ecologically representative, resilient and well-connected network of conserved areas across Europe.

https://naturaconnect.eu/

Funded by the European Union

NaturaConnect receives funding under the European Union’s Horizon Europe research and innovation programme under grant agreement number 101060429.

Funded by the European Union

NaturaConnect receives funding under the European Union's Horizon Europe research and innovation programme under grant agreement number 101060429.