BirdWatch

The Diversity of Life from Space

D4.1 Data, algorithms, and workflows for SDM

Deliverable for the Horizon Europe Project BirdWatch

Version 1.0

Funded by
the European Union

Legal Disclaimer

This document reflects only the views of the author(s). Neither the European Global Navigation Satellite Systems Agency (EUSPA) nor the European Commission is in any way responsible for any use that may be made of the information it contains. The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The below referenced consortium members shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law.

This document and the information contained within may not be copied, used, or disclosed, entirely or partially, outside of the BirdWatch consortium without prior permission of the project partners in written form.

Funded by
the European Union

Document Information

GA Number	101082634		**Type of Action**	Horizon-IA
Full Title	BirdWatch - a Copernicus-based service for the improvement of habitat suitability of farmland birds via satellite-enabled monitoring, evaluation and optimisation of CAP greening measures			
Project Acronym	BirdWatch			
Start Date	February 1st, 2023		**Duration**	36 Months
Project URL	https://birdwatch-europe.org/			
Deliverable	D4.1: Data, algorithms, and workflows for SDM			
Work Package	WP 4100			
Project Month of Delivery	**Contractual**	M12	**Actual**	M12
Nature	Report	**Dissemination Level**		PUB
Lead Beneficiary	University of Potsdam (UP)			
Responsible Author	Levin Wiedenroth - UP			
Contributions from	Levin Wiedenroth - UP Damaris Zurell - UP Emma Underwood - UP			

Funded by
the European Union

History of Changes

Version	Issue Date	Stage	Description	Comments	Contributor
1.0	23.01.2024	Draft	First version of D4.1		Levin Wiedenroth Damaris Zurell Emma Underwood

Funded by
the European Union

Table of Contents

1. Introduction
2. Species distribution models
2.1 Nested SDMs
2.1.1 Variable pre-selection
2.1.2 SDM algorithms
2.1.3 SDM ensemble and predictions
2.1.4 Joint species distribution models
2.2 SDM evaluation
3. Data
3.1 Occurrence data filtering and preparation
3.2 Standardised bird occurrence surveys
3.2.1 European Breeding Bird Atlas
3.2.2 Dachverband Deutscher Avifaunisten
3.2.3 Museum of Nature South Tyrol
3.2.4 Lithuanian Ornithological Society
3.3 Opportunistic citizen science data
3.3.1 Natuurpunt
3.3.2 eBird
3.4 Absence data
3.5 Climate data
3.6 Habitat data
4. References

Funded by
the European Union

1. Introduction

This is the first deliverable of work package 4000, which focuses on species distribution models (SDMs). These models use adequate statistical and machine learning methods to relate species occurrence records to prevailing environmental conditions. The fitted species-environment relationships can then aid the identification of suitable habitats for species by predicting the habitat suitability in space if geographic information of environmental layers is available.

As this deliverable is the foundation for the upcoming deliverables in work package 4000 it describes the work flows implemented to build, train, and test the species distribution models. It further focuses on the species occurrence data collection and data preparation needed for the models, describing the different filtering steps to ensure high data quality. Lastly, it names and describes the different datasets that contribute occurrence data and environmental data.

Funded by
the European Union

2. Species distribution models

Species distribution models (SDM) are the most widely used modelling tool in ecology (Guisan et al., 2017). They predict habitat suitability over space and time and require comparably simple data inputs, mainly information about species occurrence in the form of presence-only or presence-absence data and geographic information of environmental variables (Fig 1). The main modelling steps are conceptualisation, data preparation, model fitting, model evaluation, and prediction (Zurell et al., 2020). Conceptualisation involves gathering all necessary species information to build a preliminary understanding of species ecology (D2.2) and carefully planning the data preparation and modelling workflow (this deliverable). Data preparation relates both to species data (cf. section 3) and environmental data (D3.1). It also involves matching all data at a common spatial resolution and extent, and spatially thinning the species data to avoid any problems due to spatial autocorrelation. Model fitting describes the actual calibration of the species-environment relationship and any preliminary modelling steps such as variable pre-selection to avoid problems of multicollinearity. A multitude of statistical and machine learning algorithms are available to fit SDM, each with different strengths and weaknesses (Elith et al., 2006; Valavi et al., 2022). Considering these different algorithms is important to account for uncertainty in the modelling process (Araújo & New, 2007; Thuiller et al., 2019) and it is generally recommended to use ensembles of at least three different algorithms that are as unrelated as possible (IUCN 2021). Model evaluation is crucial for ascertaining the predictive performance of the model when making predictions to independent data and different times and places. Especially, when the models are being used for guiding management decisions, high predictive performance is of utmost importance. Typical procedures for model evaluation include x-fold cross validation, where the data are split into x folds and the model is recalibrated on (x-1) folds and then predictions are evaluated on the hold-out fold. More recently, it has been recommended to use block cross-validation where the folds are structured in geographic or environmental space to better assess predictive performance when extrapolating (Roberts et al., 2017). Finally, after successful evaluation of the model, the SDMs can be projected into geographic space by predicting them to geographic layers of environmental information. The predicted habitat suitability ranges from 0, very low suitability, to 1, perfect suitability.

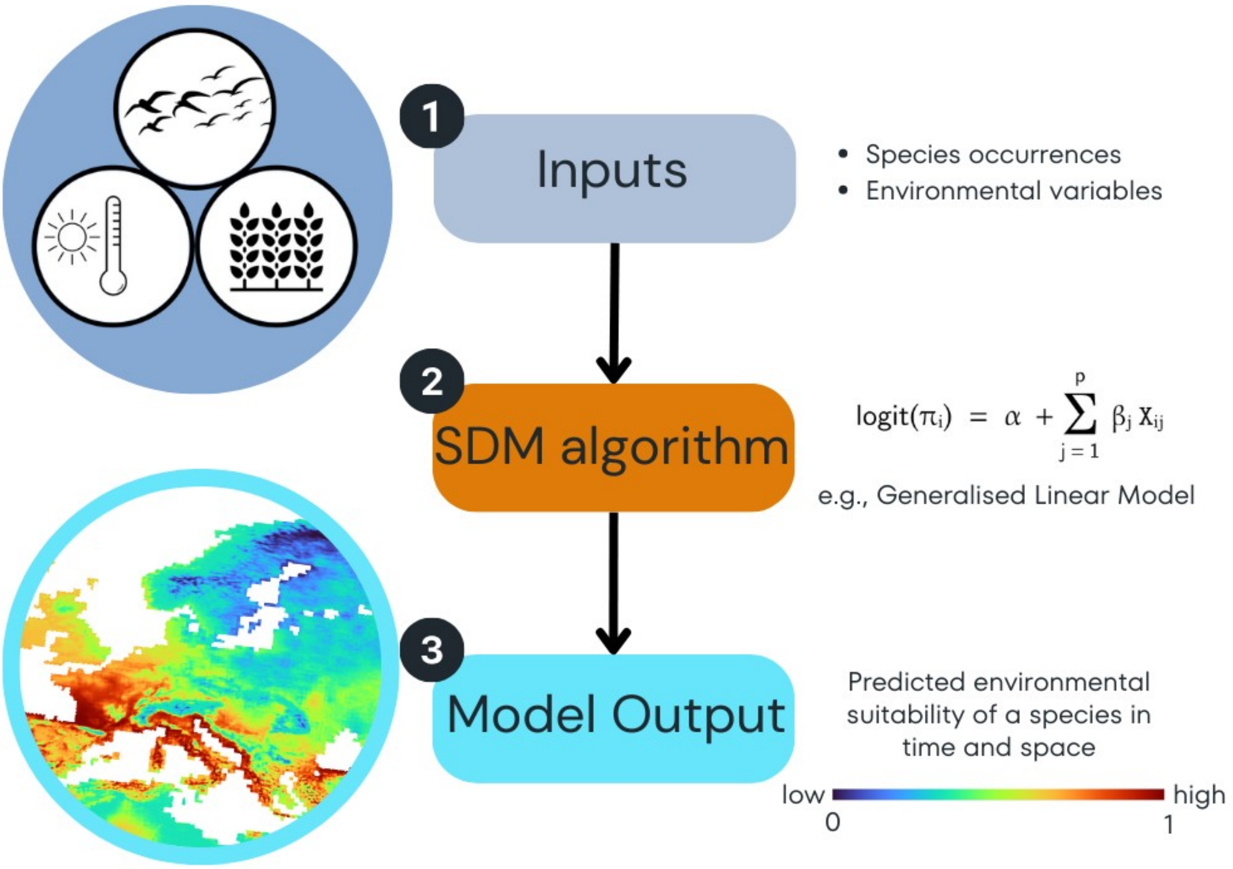
1
Inputs
• Species occurrences
• Environmental variables
2
SDM algorithm
$\mathrm{logit}(\pi_i) = \alpha + \sum_{j=1}^{p} \beta_j x_{ij}$
e.g., Generalised Linear Model
3
Model Output
Predicted environmental suitability of a species in time and space
low
high
0
1

Figure 1: SDM concept. First, it has to be determined which environmental conditions a species experiences. Second, a species-environment relationship is identified using a model algorithm. Third, this relationship is used to predict habitat suitability. Based on the habitat suitability a species potential distribution can be determined.

Generally, SDMs are used to determine the environmental conditions that limit species distributions. Thereby, it is important to consider that ecology is highly hierarchical and species distributions are limited by different environmental conditions at different spatial scales. At large spatial scales (large extents, coarse grain), climate is the main driver of most species distributions, where as at smaller spatial scales (smaller extent, fine grain), other factors like land use and habitat availability are more important determinants of fine-scale distribution of species (Guisan & Thuiller, 2005; Fig. 2).

In BirdWatch, the goal is to fit SDMs for different farmland bird species at fine spatial resolution for the different test regions and assess how different land uses and land use intensities affect species distributions. Initially, these SDMs will be calibrated separately for each test region while at later stages (WP4400), we will also test cross-region transferability in order to assess whether these models could be applied to the entirety of Europe. A major challenge here is that we find long environmental gradients in Europe and our test regions are situated in different climate zones in Europe. To ensure cross-predictability of the models, we thus need to consider both large-scale climatic conditions that limit the distribution of species across Europe and fine-scale land use and habitat conditions that limit the distribution of species within the study regions. This challenge was not anticipated in the project proposal and required slight adaptations in the planned workflow. Specifically, we adapted the SDM workflow to use so-called nested SDMs that we further describe below.

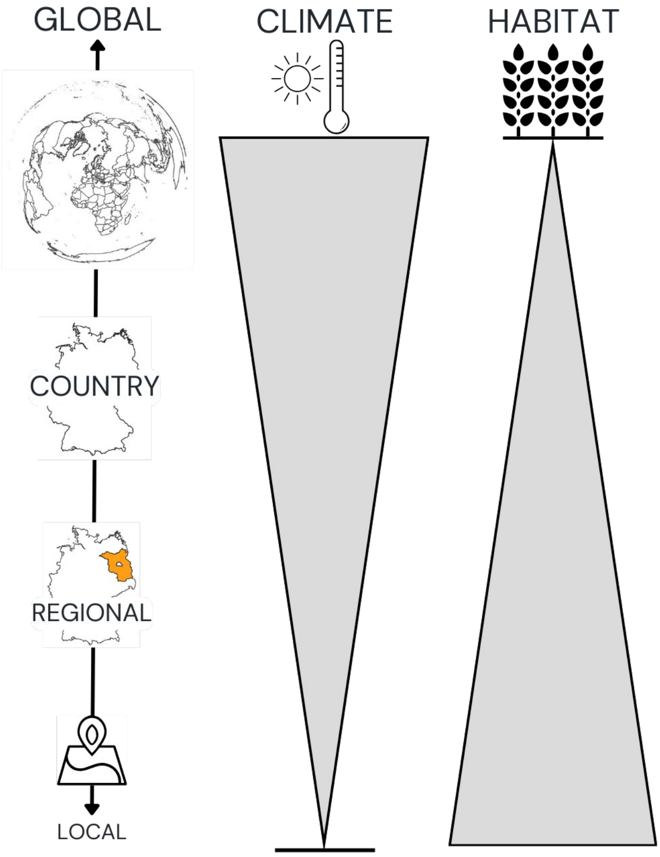

Figure 2: Impacts of scale on drivers of species distribution. The two drivers shown are climate and habitat. The bars indicate the impact of each driver depending on the scale considered. The vertical axis on the left represents the scale (global, country, regional, and local).

2.1 Nested SDMs

Nested SDMs account for the scale dependency in factors limiting species distributions. Climate is often the main determinant of large-scale species range limits while land use and other habitat factors determine the fine-scale distribution of species. Here, we adopted such a nested SDM design that allows to delineate for each species the areas of climatic suitability across Europe and

Funded by
the European Union

within climatically suitable areas describe the land use and land use intensity levels the species prefers.

Fig. 3 summarises the conceptual workflow of our nested SDMs (Adde et al., 2023; Pearson et al., 2004). First, we will fit coarse-grain SDMs to climatic data and then use the predicted climatic suitability as a predictor within the fine-grain SDMs additional to predictor variables related to land use and land use intensity. For the coarse-grain SDMs we use data from the European breeding bird atlas at 50 km spatial resolution whereas for the fine-grain SDMs we use bird occurrence data from different regional sources (cf. section 3). At both scales, SDMs are fitted using an ensemble approach with five different algorithms that differ in their model flexibility or complexity and in their extrapolation behaviour (generalised linear models, GLM; generalised additive models, GAM; random forest, RF; boosted regression trees, BRT; Maxent). First, for each species the coarse-grain SDMs will be fitted to coarse-grain climate data using these five different SDM algorithms, and then an ensemble will be constructed using the mean predicted climatic suitability over all algorithms. Second, for each species fine-grain SDMs will be constructed using the same five SDM algorithms and using the ensemble climate prediction and the land use and habitat predictors derived from earth observation (WP3000) as predictor variables. The resulting fine-grain predictions of species distribution will also be summarised within an ensemble approach and the resulting predictions and their associated uncertainty will then feed into WP5000. Below, we describe the different SDM algorithms and ensemble approaches in more detail.

Funded by
the European Union

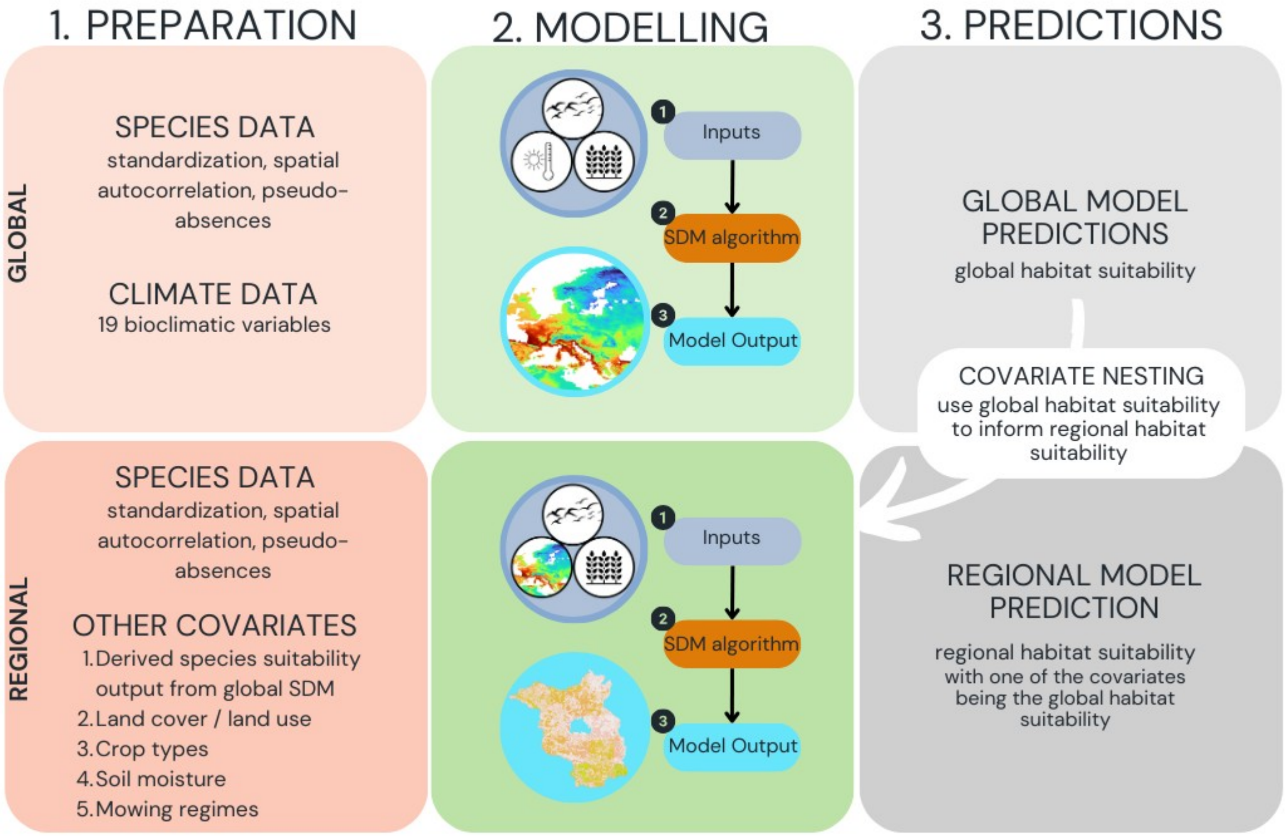
1. PREPARATION
2. MODELLING
3. PREDICTIONS
GLOBAL
SPECIES DATA
standardization, spatial autocorrelation, pseudo-absences
CLIMATE DATA
19 bioclimatic variables
1
Inputs
2
SDM algorithm
3
Model Output
GLOBAL MODEL PREDICTIONS
global habitat suitability
COVARIATE NESTING
use global habitat suitability to inform regional habitat suitability
REGIONAL
SPECIES DATA
standardization, spatial autocorrelation, pseudo-absences
OTHER COVARIATES
1.Derived species suitability output from global SDM
2.Land cover / land use
3.Crop types
4.Soil moisture
5.Mowing regimes
1
Inputs
2
SDM algorithm
3
Model Output
REGIONAL MODEL PREDICTION
regional habitat suitability with one of the covariates being the global habitat suitability

Figure 3: Nested SDM framework. The covariate nesting method is presented in which the habitat suitability output from a global SDM is used as one of the inputs for a regional SDM. The three panels at the top represent the workflow of the global SDMs, whereas the bottom three represent the workflow from the regional SDM. From left to right are the three main work steps: data preparation, model building and testing, and model predictions.

2.1.1 Variable pre-selection

At each spatial scale considered, we will pre-select weakly correlated variables as candidate predictors in our SDM. This step is important to avoid or reduce problems of multicollinearity (highly correlated predictor variables) that can lead to inflated errors in the models, hinder the correct identification of the most relevant predictors in the model and reduce extrapolation

Funded by
the European Union

ability. We will use the simple yet highly effective method "select07" where we first check for linear correlations between all potential pairs of predictor variables and then remove from pairs with an absolute correlation coefficient greater than 0.7 the less important variable (Dormann et al., 2013). Variable importance will be assessed using univariate models within a five-fold cross-validation design. For this, we will fit for each predictor variable a univariate generalised linear model (GAM) using a logit link function and 4 degrees of freedom on four of five folds, cross-predict the model to the hold-out fold and calculate the explained deviance over all folds. The explained deviance then serves as univariate variable importance in the select07 method.

2.1.2 SDM algorithms

We will use five different algorithms to fit SDMs: generalised linear models (GLM), generalised additive models (GAM), random forest (RF), boosted regression trees (BRT), and Maxent. GLMs and GAMs belong to regression-based methods while RF, BRTs and Maxent belong to non-parametric machine learning techniques.

GLMs are parametric regression models that use the logit link function to fit linear or higher polynomials to establish the species-environment relationship. We will use linear and quadratic terms (second-order polynomials) in GLMs and use an AIC-based stepwise variable selection. In the fine-grain model, climatic suitability is forced as predictor and will not be removed during the stepwise variable selection.

GAMs are semi-parametric regression methods that use data-defined, non-parametric smoothing functions to fit non-linear species-environment relationships. GAMs do not fit the response function to all data points at once, but use a moving-window approach to fit a local smoother to a proportion of the data. Small window sizes will yield highly flexible response shapes while large window sizes will produce less flexible response shapes that are closer to a parametric GLM. Here, we will use cubic smoothing splines with up to 10 degrees of freedom.

Both RFs and BRTs are based on regression trees and use model averaging to reduce overfitting problems of simple regression trees that are sensitive to local optima and noise in the data. Specifically, RFs use a bagging procedure for averaging the outputs of a multitude of different regression trees. Bagging stands for “bootstrap aggregation”. Basically, this procedure fits many regression trees to bootstrapped samples of the training data and then averages the results. An important feature of RFs are the out-of-bag samples, which means that the prediction/fit for a specific data point is only derived from averaging trees that did not include this data point during tree growing. Thus, the output of RFs is essentially cross-validated. RFs estimate variable importance by a permutation procedure, which measures for each variable the drop in mean accuracy when this variable is permutated.

BRTs use boosting as an alternative averaging approach for improving the predictive performance of tree-based methods. BRTs iteratively fit relatively simple trees by putting emphasis on observations fitted poorly by the previous trees, i.e. by fitting the new tree to the residuals of the previous tree. The final BRT can be thought of as a linear combination of all trees, similar to a regression model where each term is a single tree (Elith et al., 2008). Thereby each tree is shrunk by the learning rate (the shrinkage parameter, typically <1), which determines how much weight is given to single trees. Similarly to RFs, only a subset of the data (the bag fraction) is used for fitting consecutive trees (but in contrast to RFs, the subsets are sampled without replacement and thus constitute real data splits). This bag fraction should typically range 0.5-0.75 (Elith et al., 2008). The tree complexity controls the interaction depth; 1 means only tree stumps (with two terminal nodes) are fitted, 2 means a model with up to two-way interactions etc. We will optimise the number of trees based on the decrease in deviance when validating the trees on the out-of-bag fraction (Elith et al., 2008).

Maxent is a popular machine learning method that aims to minimise the relative entropy between the probability density of presences and the probability density of the environment estimated in environmental space. The density of available background data in environmental space can be regarded as the null model that assumes that the species will occupy environmental conditions proportional to their relative availability in the landscape (Guisan et al., 2017). Maxent allows fitting very complex, highly non-linear response shapes (Merow et al., 2013), defined by so-called feature classes. Maxent currently recognises six feature classes: linear, product, quadratic, hinge, threshold and categorical. If not otherwise specified by the user and if the data contain more than 80 presences, then Maxent will by default use all feature classes in model fitting, for fewer presences it will automatically determine the number of features based on the number of presences. During model fitting, Maxent will select features based on regularization, trading off likelihood and model complexity to avoid overfitting.

All of the SDM algorithms used require some form of absence or background data to contrast the species presences. If standardised survey data exist that provide both presence and absence information, these data can be directly used in the models. By contrast, if only opportunistic (e.g. citizen science) data are available that provide presence-only data, then background data or pseudo-absence data need to be derived prior to model building. In such cases, we will generate pseudo-absence or background data using guidelines provided by Barbet-Massin et al., 2012. All resulting data sets will be spatially thinned to avoid problems of spatial autocorrelations. Specifically, we will thin to a distance twice as long as the spatial resolution.

2.1.3 SDM ensemble and predictions

In ensembles, predictions can be combined or averaged in different ways (Thuiller et al., 2009). Here, we will use different ensemble approaches for the coarse-grain and the fine-grain models. For the coarse-grain climatic model, we will calculate simple averages of predictions using the arithmetic mean of the predictions from the single SDM algorithms. This yields a single ensemble prediction of continuous climatic suitability values for each species that can be fed into the fine-scale SDMs as predictor.

The fine-grain SDMs will serve as input to the BirdWatch optimisation algorithm (WP5000) where we want to optimise the management decisions under uncertainty. To this end, we will generate different predictions from the fine-grain SDMs based on two different ensemble approaches and also report uncertainty in SDM predictions. As the first option, we will derive simple averages of continuous habitat preference values for each species along with the standard deviation of the predictions from the single SDM algorithms. As a second alternative option, we will derive a committee average by first converting our continuous to binary predictions for each SDM algorithm. Binary predictions represent predictions of potential presence and potential absence and will be achieved by applying a simple threshold to the continuous habitat preference values using the maxTSS approach that aims to identify the threshold that maximised the true skill statistic TSS (or sum of true positive and true negative rate). Then, these binary predictions are summed up and divided by the maximum number of algorithms used yielding a prediction of how many SDM algorithms agree that the species should be present at a specific location.

2.1.4 Joint species distribution models

In the original BirdWatch proposal, we additionally suggested fitting joint species distribution models (JSDMs) to improve predictions of local species assemblages. JSDMs simultaneously fit species-environment relationships for many species together with their residual co-variation. This residual covariation can be indicative of interspecific interactions, spatial effects, and missing environmental predictors.

After careful consideration, we decided against using JSDMs in the BirdWatch model. First, due to the hierarchical nature of our species data, we had to adopt a nested SDM approach and thus a more complex SDM workflow than originally anticipated. Second, it has been shown repeatedly that JSDM do not improve predictive accuracy over more classical SDMs (Wilkinson et al., 2019; Zurell et al., 2019) and need to be interpreted with great caution (König et al., 2021; Poggiato et al., 2021). Thus, their implementation would not yield improved predictions of habitat preference

Funded by
the European Union

for the different farmland birds. Last, getting high-quality bird distribution data for all our farmland birds proved challenging. Specifically, standardised survey data with presences and absences are not available for all test regions (cf. section 3) and we thus needed to adopt modelling workflows to presence-only data that require pseudo-absence or background data generation. Current JSDMs algorithms require presence-absence data and no JSDM methods have been developed yet to deal with presence-only data.

For these various reasons, we decided against using JSDMs in BirdWatch but are at the same time very confident that the exclusion of JSDMs strengthens rather than weakens our approach.

2.2 SDM evaluation

Within WP4200 and WP4300 we will use five-fold cross-validation to evaluate model predictive performance while in WP4400 we will cross-predict the models to different test regions and validate the predictive performance against the occurrence data from those regions. In either case, we will use a suite of different performance measures to receive a broad view of model predictive accuracy. Specifically, we will use the threshold-dependent measures sensitivity, specificity and TSS (true skill statistic) and the threshold-independent measure AUC (area under the ROC, receiver operating characteristic, curve). Sensitivity is the true positive rate, meaning the number of correctly predicted presences, and specificity the true negative rate, the number of correctly predictive absences. Both range between 0 and 1. TSS is a composite measure defined as sensitivity + specificity - 1, and ranges between -1 and +1. All three measures require binary predictions, thus the continuous SDM prediction first needs to be converted into a binary presence-absence prediction. As indicated in section 2.1, we will use the maxTSS approach to identify an optimal threshold. AUC is calculated by quantifying sensitivity (true positive rate) and specificity (true negative rate) for many potential thresholds along the entire range of predicted occurrence probabilities. Then, 1-specificity (false positive rate) is plotted on the x-axis against sensitivity on the y axis and the AUC is then defined as the integral under this curve. AUC ranges

Funded by
the European Union

between 0 and 1 with 1 indicating perfect discrimination, 0.5 indicating random prediction and 0 indicating mirror-inverted predictions. If we would take a random presence and a random absence from our observations and make predictions, then AUC can be interpreted as the chance of assigning a higher predicted occurrence probability to the presence compared to the absence point. Typically, we regard AUC>0.7 as indicating fair predictions.

3. Data

The data we gathered for this deliverable will be the foundation for building the SDMs for the different regions and thus for the upcoming deliverables of work package 4000. To build the SDMs we require two types of data: 1. Species occurrence data, in the form of presence-absence or presence-only data and 2. Environmental data, such as land cover or crop types, as well as data that describes the climatic conditions experienced by the species.

Species occurrence data give information on where a species is present and can also be used to make an informed guess on where a species is absent. Besides the name of the species, occurrence data always include the coordinate location and timestamp of where the species was observed, usually giving the year and month of observation. Additional information, like the geographic accuracy or the displayed behaviour are also commonly included.

We use two different types of species occurrence data: standardised bird survey data and opportunistic citizen science data. Standardised bird occurrence surveys provide high quality data at the cost of much higher collection effort, planning, and preparation. Not every region selected in the BirdWatch project has such datasets available. As an alternative, there are opportunistic citizen science datasets. Their data are collected in a less structured way, and are therefore of lower quality, but the availability and spatial coverage is much higher. For our regions, we used standardised data where obtainable and opportunistic citizen science data to infill gaps. Further information on standardised bird occurrence surveys and opportunistic citizen science data is given under section 3.2 and 3.3 respectively.

Environmental data are the second type of data needed for SDMs to form species-environment relationships. The data provide information on the environmental conditions each species experiences. We use coarse-grain climate data at a resolution of 50 x 50 km and European extent and fine-grain regional data describing the habitat requirements of the species at a resolution of 200 x 200 m. Further information on the environmental data is given under section 3.5 and 3.6.

3.1 Occurrence data filtering and preparation

To identify suitable high quality occurrence data, regardless of data source type (standardised or opportunistic), we filtered the available data systematically. First, we determined if the data were verified by experts in bird identification. The verification was for example based on photographic evidence, the observer's skill level in bird identification, or on the proximity to other observations of the same species. If an occurrence did not pass this validation phase, or validation did not take place at all, we did not include that record. To ensure the presences represented highly suitable habitats we only included presences that represented breeding bird occurrences. To identify breeding presences, we filtered by breeding period and behaviour displayed during observation. For Germany and Flanders the breeding season ranges from March to June, for South Tyrol from April to July, and for Lithuania from March to July. The breeding seasons were determined based on expert opinions of ornithologists active within the BirdWatch regions. Behaviours that ensure or indicate breeding birds were, for example, a bird actively breeding, building a nest, or displaying territorial behaviours. Next, we filtered by the geographic accuracy of the observation. For many observations, the exact location of the bird is not known, for example, when it was identified only by sound. In those cases, an accuracy measure in metres is given, and we only included presences within 100 metres. This distance was based on the resolution of the regional environmental data, which is 200 x 200 m and the home range size of our study species (D2.2). The selected accuracy ensures that the observation is associated with the correct environment for the given location. Lastly, the occurrence data had to align with the spatial extent and resolution of the environmental data (grid of 200 x 200 m). To ensure an environmental grid cell was not overrepresented by species occurrences, we removed duplicates where more than one species record lay within a grid cell.

We ran these filtering steps for several different years per region to identify the most suitable year. The year chosen was not necessarily the year with the most occurrences in total but in which all species had comparably high numbers of occurrences to build the models, with preference for more recent years. This resulted in the year 2022 for Germany, South Tyrol, and Lithuania. For Flanders, we had to decide on the year before we were able to filter the data to allow the environmental data compilation to commence. Within the BirdWatch consortium, we decided on the year 2018, which provides good amounts of occurrence data for most species. Table 1 gives an overview of the available high-quality data for the region of Flanders, South Tyrol, and Lithuania after applying the aforementioned filtering steps. For Germany, these numbers correspond to the maximum available occurrence points.

In case a species was rare and regional datasets provided only a few high-quality occurrence points, we checked additional data availability in a global dataset (eBird). However, due to quality criteria in our filtering steps this yielded no additional data points. Thus, some species or even entire regions (South Tyrol) had only a few occurrence points after filtering. Within work packages 4200 and 4300 we will test the effect of different filtering steps on model building and performance. Depending on the result, we might not have to filter the occurrence data as rigorously, which would increase the number of occurrences for rare species and might have positive effects on their model performance. If no solutions can be found, it is possible that some species cannot be modelled for a region due to insufficient amounts of data.

Table 1: Number of species occurrence data (before spatial thinning) for the four regions of BirdWatch.

	GERMANY	FLANDERS	SOUTH TYROL	LITHUANIA
Eurasian skylark	16004	2025	36	1163
Meadow pipit	722	547	0	156
Yellowhammer	10605	1026	48	610
Red-backed shrike	1068	8	31	76
Black-tailed godwit	15	330	0	3
Eurasian tree sparrow	4708	73	13	87
Whinchat	304	1	10	400
Eurasian turtle dove	219	118	0	13
European starling	15096	729	10	353
Northern lapwing	638	1472	0	155

3.2 Standardised bird occurrence surveys

Standardised surveys in biology and more specifically ecology collect data that are highly comparable and informative. Data collection is based on standardised protocols and along predefined routes or transects. The observers, often volunteers, are skilled in bird recognition and can identify birds based on sight as well as sound. Complete checklists are used when an area is surveyed, meaning that all birds present are documented. The advantage of listing all observed species is that locations where a species was not detected can then be interpreted as an absence with high certainty.

Survey locations are strategically distributed across the area to serve as representative samples, ensuring coverage of the entire area without the need to sample every individual section. To minimise biases and have the data as comparable as possible between sites and years the same sampling procedures and methods are used every time. To reduce observer bias, which is caused by differences in the ability to identify birds, observers have to pass a test that ensures they are proficient in bird identification. Sample effort, the amount of time spent to survey a single location with higher effort increasing the likelihood to observe species, is kept the same for all locations ensuring a low effort bias. Spatial bias, when areas are over or under sampled, is accounted for by placing the survey areas in a way that represents the entire study area with its different habitats. This also accounts for accessibility bias, that some areas are sampled more or less often, because of their accessibility. To reduce seasonal bias, the surveys are always held during the same time of the year. For birds, this period is often during their breeding season, since a breeding habitat is a good indication of habitat suitability. Data are always collected at the same time of day to minimise temporal bias in bird detectability because of differences in activity during the day. To reduce weather bias, which is caused by different weather conditions affecting the likelihood of observing birds, the surveys are only held on days with comparably good weather conditions while rainy or very windy days are excluded. At the end of a survey period, all data are checked for potential errors, such as typos and then cleaned by the organisation that coordinated the surveys. Data not meeting the standards of the survey are discarded, resulting in the final data being highly comparable with a low bias.

Within BirdWatch we obtained standardised bird occurrence survey data from the European Breeding Bird Atlas (EBBA2), the Dachverband Deutscher Avifaunsiten (DDA), the Museum of Nature South Tyrol, and the Lithuanian Ornithological Society (LOD). An overview of the data sources for the different regions is provided in figure 4 while the data will be described in more detail in the following sections.

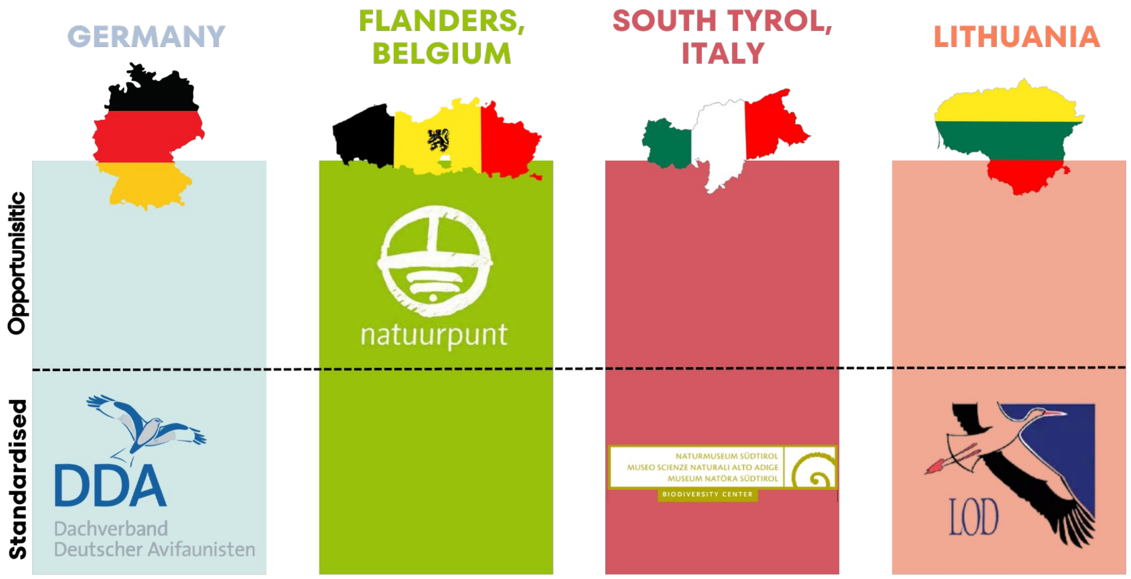

Figure 4: Data sources. Overview on the different data sources that contributed occurrence data for their respective region. The data sources were separated by opportunistic citizen science data (top row) and standardised surveys (bottom row). The data for Lithuania was provided by the Lithuanian Ornithological Society (LOD).

3.2.1 European Breeding Bird Atlas

EBBA2 is one of the largest European standardised bird survey projects that ever took place (EBCC, 2022; Keller et al., 2020). The goal of EBBA2 was to get an overview of the distribution of breeding birds for the whole of Europe. The project was led by the European Bird Census Council and the atlas was published in 2020. Organisations from 48 different countries were involved in the data collection, with around 120,000 fieldworkers contributing to the project, most of them on a voluntary basis. Recorded were breeding species occurrences, whereas birds occurring only temporarily, for example, when migrating through an area or only staying over winter, were excluded. The breeding occurrences were further divided into categories: possible breeding, probable breeding, and confirmed breeding. These categories were introduced by the European Ornithological Atlas Committee and are widely used in European national atlases and online platforms. An observation was classified as "possible breeding" when a bird was observed during the breeding season in a possible nesting habitat or when a singing male or a male that used breeding calls was present during the breeding season. Under "probable breeding", were occurrences where a pair was observed in a breeding habitat, an individual displayed courtship behaviour, or individuals were seen building a nest. Lastly, an occurrence was considered as "confirmed breeding" when a used nest or eggshells were found, recently fledged young was observed, or a nest with eggs or young in it was observed. For the species distribution models, all categories were included as presences. The occurrence data were mainly collected over a time span of 5 years, from 2013 to 2017. Four different data sources were used in EBBA2: 1. atlas data (which were collected during national or regional projects), 2. monitoring data (general monitoring as well as species-specific monitoring), 3. casual observations (often stemming from online platforms), and 4. surveys conducted for EBBA2. The usage of different data sources for EBBA2 was considered necessary and advantageous because the different situations in each country and data availability could be considered This ensured a complete coverage of the European continent, with a spatial resolution of 50 km.

3.2.2 Dachverband Deutscher Avifaunisten

The DDA is a German avifaunistic umbrella organisation for birding clubs and individual birders (https://www.dda-web.de/). One of their main tasks is the coordination of several bird occurrence data surveys. The Monitoring of Common Breeding Birds (MCB) provides data on the species relevant to the BirdWatch project and will be used to build the SDMs for the region of Germany. The MCB started in 1989 and was updated in 2004 to make the data collection more standardised. The goal of MCB is to have comparable yearly occurrence data on breeding birds to be able to infer population trends. To achieve this goal, more than 2,600, 1 x 1 km sample plots were distributed across Germany, based on a stratified random sampling approach. The first stratum describes the “environmental regions” of Germany, i.e., regions with similar abiotic attributes like climate, slope and elevation, and soil. The second stratum refers to land-cover types, e.g., arable land, settlements, forests, and other special habitats like heathlands, mires and bare soils. With these two categories, the whole of Germany can be represented. Volunteers with sufficient knowledge of bird identification survey the sample plots by walking a 3 km long route during the morning hours on days with suitable weather conditions. Each sample plot is surveyed four times per year between the 10th of March and the 20th of June. During each survey, the observer notes down all birds they hear or see and determines territories for each species based on the observations. The summarised territory maps are then the final output of the MCB survey. We will use these simplified territory maps and derive bird occurrence data at the target spatial resolution of 200 m.

3.2.3 Museum of Nature South Tyrol

The Museum of Nature South Tyrol provided standardised occurrence data for the South Tyrol region. They collected the data as part of different monitoring and research projects. The surveys were carried out by bird experts skilled in bird recognition. For their data collection, they used standardised protocols for point counts. The point count data collection was conducted by walking along transects and stopping every 200 m. At each stop, 10 minutes were spent by the observer to document all birds heard or seen. The transects went along mountain areas as well as open areas. The point count surveys were conducted two to three times a year during the breeding season. The surveys were done early during the day (5:30 to 11:30 am) on days with reasonable weather conditions, e.g., no rain or strong winds. The point count data came with two accuracy classes, more and less than 100 metres. During their surveys they only collected data on seven out of the 10 species relevant for BirdWatch. The other three species do not breed in the region of South Tyrol; this information is based on expert opinion and backed up by the data. These three species are the Black-tailed Godwit (*Limosa limosa*), Northern Lapwing (*Vanellus vanellus*), and Meadow Pipit (*Anthus pratensis*).

3.2.4 Lithuanian Ornithological Society

We received occurrence data for Lithuania from the LOD (https://www.birdlife.lt/). It was collected as part of the "Common Bird Population Abundance Monitoring project" (http://www.ipgs.lt/), which was initiated in 1985 to provide data on bird abundance and change (Kurlavicius, 2004). In their project, they used the point count method and a standardised protocol to collect the data. The point data were collected by walking along approximately 10 km long routes that consisted of 20 stops, roughly 500 m apart from one another. Locations for routes were selected based on a random stratification approach. This ensured that the sample locations represented the whole of Lithuania. During each of the twenty stops, a skilled person in bird recognition via appearance and sound documented all birds observed within five minutes. This was done twice a year during the breeding season. Surveys were always conducted in the morning on days with suitable weather conditions, without strong winds or rain. Each observation was classified by how far away from the observation point the bird was observed. This resulted in three categories: 1) the bird was within 50 m of the observation location, 2) the bird was within 50 - 100 m or 3) the bird was further away than 100 m. For the data to be included in our models, we only selected observations that fell into accuracy category 1 or 2.

3.3 Opportunistic citizen science data

Opportunistic citizen science data are data collected by volunteers in an informal and spontaneous manner. Thus, unlike in standardised surveys, the data collection is based on volunteers contributing observations taken during everyday activities without prior planning or coordination. An example of such opportunistic data is someone walking through a park, observing a bird, and contributing this sighting to a database. This type of data comes with many advantages but also several disadvantages.

One major advantage is the tremendous amount of data that is generated in this way. This can be attributed to the simplicity of data collection and the amount of people that can contribute data. Another reason is its cost-effectiveness. Much less organisation and planning has to go into collecting data with volunteers using their own resources to make observations. Because so many people participate in the data collection, a wide geographic and temporal coverage can be achieved spanning over entire countries and many years. Data are collected much more frequently, which can reduce problems of imperfect detection and give a more complete picture of suitable species habitats.

On the downside, the lack of structure during data collection leads to lower quality in the data and decreases comparability between the different observation events. After data collection, it is often difficult to implement measures that ensure data quality if certain information is missing. For example, if no information is given on how long the observer spent at a location or which behaviour the species displayed during observation, it is not possible to acquire this information post hoc. The lack of standards and missing structure for data collection introduce many sources of bias into the data. People contributing vary in their ability to identify birds, introducing an observer bias. The fact that not all species observed have to be reported also adds observer bias. Observers can spend as much or as little time observing a bird, which can cause an effort bias. Another bias common in opportunistic data is spatial bias because some areas are visited more often than others. This is also caused by high accessibility bias, meaning that more accessible areas are visited more often, and thus more observations come from these areas. Observations can be made during any time of the day and during any weather conditions, introducing temporal and weather biases into the data. In order to minimise these disadvantages, organisations that collect opportunistic citizen science data often provide a protocol to people in order to make the different data entries more comparable with one another. It is then up to the observer to include as much meta-information as possible. Further, these organisations often have quality controls in place. For example, bird experts or algorithms check data entries and flag observations that seem out of place. An unlikely observation could be a species that is very far away from its usual range or an observation of an unusually high number of individuals for a given species. In addition, we implemented several filtering steps described under the section "3.1 Occurrence data filtering and preparation" to ensure that all occurrence points met a similar level of quality.

Natuurpunt and eBird were identified to contribute opportunistic citizen science data to the BirdWatch project. In the following sections, we describe these two organisations and their data. An overview of which data sources were used for each region is found in figure 4.

3.3.1 Natuurpunt

Natuurpunt (https://www.natuurpunt.be/) is an NGO from Flanders and provides localised occurrence data for that region. It is the largest nature conservation organisation in Belgium, which hosts an opportunistic citizen science platform via the website https://waarnemingen.be/. Everyone can contribute their casual bird observations to this opportunistic citizen science platform. To ensure species are correctly identified, all data must be validated. There are five different validation categories: 1) definitely correct: the observer provided evidence, for example, photos, or the observation was made by a bird expert directly; 2) very likely correct: the observation of a species is in close proximity with the observation of the same species made by others; 3) possibly correct: the species identification was made by automatic image recognition; 4) not possible to verify: none of the aforementioned validation techniques could be applied to the observation; and 5) not checked yet: the observation has not been validated yet. To build the SDMs for Flanders, we only used data from the validation categories 1 and 2.

Natuurpunt further ensures data quality through a protocol that has to be filled out by data observers when uploading data. It is obligatory to always provide information on the date when the observation took place, which species was observed, and where the species was observed in the form of coordinates. Additional information can be provided if available. This includes information on the accuracy of the coordinates in metres and information on the behaviour the species displayed while being observed.

3.3.2 eBird

eBird is a large global opportunistic citizen science project which documents bird distributions, abundance, habitat use, and population trends (https://ebird.org/home). We use data from their Basic Dataset (eBird Basic Dataset, 2023). The minimum requirements to contribute an observation to eBird are to provide the scientific name of the species, the date of the observation, and the location with coordinates. eBird provides observers with a protocol to provide additional information. This includes the behaviour displayed during the observation, indicating if the birds are breeding or not. If the information was given, we only included data that had the breeding category C3 (probable) or C4 (confirmed). Behaviours from category C3 included, for example, courtship or territorial defence behaviours, and from C4, when the bird was sitting in a nest or displayed a distraction behaviour. In addition, eBird ensures that all data are validated. In the first instance, the data are validated by an algorithm that determines if the observation shows unusual patterns. If the algorithm flags an observation, the observer is contacted to provide additional information for validation. Then, a reviewer from eBird is consulted and makes the final decision on the validation status.

3.4 Absence data

To form a reliable species-environment relationship, most SDM algorithms need absence or background data in addition to presence data. These data complement the presence data by providing information on where species do not occur and help to infer unsuitable environmental conditions. True absence data are rarely available for a given location because of the amount of effort that would have to be invested to minimise imperfect detection. There are common alternative approaches that can be used to infer so-called pseudo-absences or background data, indicating locations where species likely did not occur and how many should be selected for SDM construction (Barbet-Massin et al., 2012).

Full species checklists, which include all observed and identified species during one period of data collection at a location and not just individual species, can be used to make an informed estimate on absence locations. In this case, an absence location is equal to a location where the study species was not listed as present. For all our standardised survey datasets, we will use this approach to define absences.

For data where no full checklists were used, which is the case for our opportunistic citizen science data, random background points have to be sampled to train the models. Our study regions determine the spatial extent from which random points can be selected because outside of the study regions, we do not have environmental data to match the points.

3.5 Climate data

Climate is considered the main driver of a species distribution at coarse scales (McGill, 2010) and thus determines broad areas of suitability. We obtained climate data from the CHELSA (Climatologies at high resolution for the earth's land surface areas) dataset (Karger et al., 2017, 2021). As parameters we used maximum mean monthly temperature, minimum mean monthly temperature, and mean monthly precipitation per year. The data came at a 1 km resolution, which we aggregated to a 50 km resolution to match the EBBA2 dataset. As changes in climate can have lagged effects on habitat occupancy of birds, we included climate data from the EBBA 2 collection period and the previous year (Albright et al., 2011). The years we selected climate data from were 2012 - 2017.

In our next step, we averaged each of the three climate variables for every month of every year across 50 x 50 km cells over the six years. Based on these climate variables we calculated 19 bioclimatic variables - i.e., bio1: mean annual temperature; bio2: mean diurnal range; bio3: isothermality; bio4: temperature seasonality; bio5: max temperature of the warmest month; bio6: min temperature of the coldest month; bio7: temperature annual range; bio8: mean temperature of the wettest quarter; bio9: mean temperature of the driest quarter; bio10: mean temperature of the warmest quarter; bio11: mean temperature of the coldest quarter; bio12: total annual precipitation; bio13: precipitation of the wettest month; bio 14: precipitation of the driest month; bio15: precipitation seasonality; bio16: precipitation of the wettest quarter; bio17: precipitation of the driest quarter; bio18: precipitation of the warmest quarter; and bio19: precipitation of the coldest quarter. These bioclimatic variables are commonly used in species distribution modelling (Booth et al., 2014).

3.6 Habitat data

Unlike climate data, habitat data plays a crucial role in shaping species distribution at finer scales (McGill, 2010) and determines where a species occurs. The data describing the species habitats are provided by our BirdWatch partners at a resolution of 200 x 200 m as part of work package 3000. The habitat parameters were grouped into landscape features: crop type, land cover, and remote sensing. The habitat data will be provided for each region for the year deemed most suitable based on the number of occurrences. For more detailed information regarding the habitat parameters, see deliverable “3.1. Database of geospatial data”.

4. References

Adde, A., Rey, P.-L., Brun, P., Külling, N., Fopp, F., Altermatt, F., Broennimann, O., Lehmann, A., Petitpierre, B., Zimmermann, N. E., Pellissier, L., & Guisan, A. (2023). N-SDM: A high-performance computing pipeline for Nested Species Distribution Modelling. *Ecography*, *2023*(6), e06540. https://doi.org/10.1111/ecog.06540

Albright, T., Pidgeon, A., Rittenhouse, C., Clayton, M., Flather, C., Culbert, P., & Radeloff, V. (2011). *Heat waves measured with MODIS land surface temperature data predict changes in avian community structure*. https://doi.org/10.1016/J.RSE.2010.08.024

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, *22*(1), 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). bioclim: The first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, *20*(1), 1–9. https://doi.org/10.1111/ddi.12144

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

EBCC (2022). *European Breeding Bird Atlas 2 website.* European Bird Census Council. Accessed from: *http://ebba2.info* (22/01/2024).

eBird Basic Dataset. Version: EBD_relJul-2023. Cornell Lab of Ornithology, Ithaca, New York. Jul 2023.

Elith, J., H. Graham*, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat Suitability and Distribution Models: With Applications in R* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781139028271

IUCN 2021: International Union for Conservation of Nature annual report. (2022). IUCN. https://portals.iucn.org/library/node/49944

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, *4*(1), 170122. https://doi.org/10.1038/sdata.2017.122

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2021). *Climatologies at high resolution for the earth´s land surface areas*. https://doi.org/10.16904/envidat.228.v2.1

Keller, V., Herrando, S., Voříšek, P., Franch, M., Kipson, M., Milanesi, P., Martí, D., Anton, M., Klvaňová, A., Kalyakin, M.V., Bauer, H.-G. & Foppen, R.P.B. (2020). *European Breeding Bird*

Funded by
the European Union

Atlas 2: Distribution, Abundance and Change. European Bird Census Council & Lynx Edicions, Barcelona.

König, C., Wüest, R. O., Graham, C. H., Karger, D. N., Sattler, T., Zimmermann, N. E., & Zurell, D. (2021). Scale dependency of joint species distribution models challenges interpretation of biotic interactions. *Journal of Biogeography*, *48*(7), 1541–1551. https://doi.org/10.1111/jbi.14106

Kurlavicius, P. (2004). *MONITORING OF BREEDING BIRDS IN LITHUANIA*.

McGill, B. J. (2010). Matters of Scale. *Science*, *328*(5978), 575–576. https://doi.org/10.1126/science.1188528

Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Pearson, R. G., Dawson, T. P., & Liu, C. (2004). Modelling species distributions in Britain: A hierarchical integration of climate and land-cover data. *Ecography*, *27*(3), 285–298. https://doi.org/10.1111/j.0906-7590.2004.03740.x

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., & Thuiller, W. (2021). On the Interpretations of Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, *36*(5), 391–401. https://doi.org/10.1016/j.tree.2021.01.002

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Thuiller, W., Guéguen, M., Renaud, J., Karger, D. N., & Zimmermann, N. E. (2019). Uncertainty in ensembles of global biodiversity scenarios. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-09519-w

Funded by
the European Union

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

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. https://doi.org/10.1002/ecm.1486

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, *10*(2), 198–211. https://doi.org/10.1111/2041-210X.13106

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261–1277. https://doi.org/10.1111/ecog.04960

Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2019). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, *47*(1), 101–113. https://doi.org/10.1111/jbi.13608

Funded by
the European Union