Supplementary Materials

Table S1. ODMAP Protocol (ver. 1.0) for food-probability models for the brown bear in the Central Forest State Nature Biosphere Reserve (West-European Russia). Details on Data, Model, Assessment, and Prediction. For the Overview section and references, please refer to the main text.

ODMAP element	Contents
	OVERVIEW
Authorship	• **Authors:** Ogurtsov, S.S.; Khapugin, A.A.; Zheltukhin, A.S.; Fedoseeva, E.B.; Antropov, A.V.; Delgado, M.d.M.; Penteriani, V. • **Contact email:** etundra@mail.ru • **Title:** Brown bear food-probability models in West-European Russia: on the way to the real resource selection function. • **DOI:** https://doi.org/10.3390/f13081247
Model objective	• **SDM objective:** ecological inference/explanation; mapping/interpolation. • **Target outputs:** Environmental variable importance and response curves/maps of the probability of presence and binary maps of potential presence for each species.
Taxon	Brown bear main food resources include *Apiaceae* forbs (*Angelica sylvestris*, *Aegopodium podagraria*, *Chaerophyllum aromaticum*), dwarf-shrubs (*Vaccinium myrtillus*, *Vaccinium microcarpum*, *Vaccinium oxycoccos*), shrubs (*Corylus avellana*), trees (*Populus tremula, Sorbus aucuparia*, *Malus domestica*), insects (anthills, xylobiont insects, social wasps), and mammals (*Alces alces*).
Location	Central Forest State Nature Biosphere Reserve (CFNR), Tver region, West-European Russia.
Scale of analysis	• **Spatial extent (Lon/Lat):** Longitude 32.61° E – 33.24° E, Latitude 56.42° N – 56.64° N. • **Spatial resolution:** 30 m. • **Temporal extent/time period:** March–November from 2008 to 2020. • **Type of extent boundary:** Administrative (boundary of the Protected Area).
Biodiversity data overview	• **Observation type:** Field survey. • **Response/Data type:** Presence only.
Type of predictors	Vegetation indices, terrain, distances to rivers, landcover, treecover
Conceptual model/Hypotheses	**Hypotheses about species–environment relationship:** We believe that the distribution of the focal species is associated to a greater extent with the abundance of phytomass and landcover types, and to a lesser extent with terrain variables (except elevation). We also believe that the distribution of *Vaccinium myrtillus* is more related to the protected boreal forests of the CFNR core area. The distributions of *Populus tremula*, *Sorbus aucuparia*, *Corylus avellana*, *Malus domestica*, and anthills are largely associated with the human-modified territory of the CFNR buffer zone. The most productive areas for many species are located on terrain elevations along moraine–kame ridges.
Assumptions	**We assumed that:** • Relevant ecological drivers (or proxies) of species distributions are included. Predictors are measured (or estimated) without errors. • Detectability does not change across habitat gradients. • Species are at equilibrium with their environment. • Sampling is adequate and representative (and any biases are accounted for/corrected). • All presence records are independent observations. • There were no dramatic changes in the environment during the study period.
SDM algorithms	• **Algorithms:** We fitted MaxEnt to the field data. MaxEnt was chosen due to its competitive performance on small sample sizes and ease of use, and its outputs were approximated to true probabilities using published equations.

	• **Model complexity:** MaxEnt models were built with linear, quadratic, product, and hinge features in different combinations for different species. A data-driven approach with a genetic algorithm through *SDMtune* functions was selected to choose the optimal combinations of hyperparameters (features and regularization multiplier) for each species.
Model workflow	For model workflow, see Figure 2 in the main text.
Software	• **Software:** Analyses were conducted using RStudio 1.1.447 software based on R 4.0.3 [40] with the packages *SDMtune* [79,82], *dismo* [83], *blockCV* [50], *ecospat* [91], *spThin* [45], *spatialEco* [47] and MaxEnt version 3.4.1 (https://biodiversityinformatics.amnh.org/open_source/maxent/; accessed on 10 March 2021). Field data collection was performed with mobile applications using ArcGIS QuickCapture and ArcGIS Survey123 (Esri Inc., Redlands, California, U.S.). Geodata processing was performed with the help of ArcMap 10.6.1 (Esri Inc., Redlands, California, U.S.) and SAGA GIS 7.7.1 [66]. • **Code:** code not shared, available on request. • **Data:** data not shared, available on request.
DATA	
Biodiversity data	• **Taxon names:** *Apiaceae* forbs (*Angelica sylvestris*, *Aegopodium podagraria*, *Chaerophyllum aromaticum*), dwarf-shrubs (*Vaccinium myrtillus, Vaccinium microcarpum, Vaccinium oxycoccos*), shrubs (*Corylus avellana*), trees (*Populus tremula, Sorbus aucuparia, Malus domestica*), insects (anthills, xylobiont insects, social wasps), and mammals (*Alces alces*). *Vaccinium microcarpum* and *Vaccinium oxycoccos* are hereinafter presented together as *Oxycoccus* spp. • **Details on taxonomic reference system:** Latin names of plants are standardized according to the database POWO [122]. Taxonomy of ants is given according to the AntWeb database [123]. Taxonomy of social wasps is given according to Carpenter and Kojima [124] and Daglio [125]. Latin names of mammals are given according to Wilson and Reeder [126]. • **Ecological level:** Species level. • **Data source:** Survey data collected in the field from March to November in 2008–2020. • **Sampling design:** Sampling was carried out on regular hiking routes through the study area four times a week (2466 km in reserve core area and 2167 km in buffer zone). The routes passed along clearings, roads, and paths. The collection of data on plant distribution was performed using the specially developed form in the ArcGIS QuickCapture (Esri Inc., Redlands, California, U.S.) mobile application. The collection of data on animal distribution was performed using the specially developed form in the ArcGIS Survey123 (Esri Inc., Redlands, California, U.S.) mobile application. Given minimum georeference accuracy was 4.6 m. Only those individuals that are suitable for brown bear consumption were recorded. • **Sample size:** *Angelica sylvestris* (196), *Chaerophyllum aromaticum* (113), *Aegopodium podagraria* (56), *Populus tremula* (56), *Vaccinium myrtillus* (325), *Oxycoccus* spp. (170), *Corylus avellana* (203), *Sorbus aucuparia* (151), *Malus domestica* (95), anthills (274), social wasps (193), xylobiont insects (151), and *Alces alces* (229). • **Regional mask:** We clipped all data to the boundary of the study area (Central Forest Nature Reserve). • **Scaling:** Records were spatially thinned (within 30 m). • **Data cleaning/filtering:** All records were rarefied according to average nearest neighbour index values. Sampling bias was eliminated by randomly removing records within the distance at which NNI > 1. All final record sets (except *Oxycoccus* spp.) showed dispersal distribution without clustering.

	• **Background data:** We generated 10,000 random background points within the study area based on minimum convex polygon (convex hull), which reflected survey intensity. • **Errors and biases:** Error rates were deemed low, as species presence locations were recorded with a GPS accuracy of 4.6 m. All records were collected by the same author (S.S.O.). Misidentification rates were deemed low, as all species were previously identified in the laboratory by botanist (for plants; A.A.K.) and entomologists (for insects; E.B.F., A.V.A.).
Data partitioning	While tuning the model for optimal hyperparameters, we used block cross-validation with spatial blocking strategy with a random pattern and 100 iterations for dividing the data. For the final model evaluation (testing), we also used truly independent datasets from the CFNR archive.
Predictor variables	• **Predictor variables:** Vegetation indices (EVI, GNDVI, NDMI, GCVI, ARVI, wetness), terrain variables (elevation, slope, northness and eastness, hillshade, TRI, solar radiation, CTI), distance to rivers, landcover (landcover types in %), forest canopy cover. • **Data sources:** Vegetation indices: average values of vegetation indices were derived from Landsat 8 OLI-TIRS nine no-cloud scenes from 2014 to 2020 (26.04.2014, 06.06.2014, 10.09.2014, 29.04.2015, 25.09.2017, 07.05.2018, 11.08.2018, 19.05.2019, 11.06.2019). Before calculating the vegetative indices, a radiometric correction was performed: first, the band values of multispectral images were converted from standard digital numbers (DN) to surface reflectance values by performing a Top of Atmosphere (TOA) correction, and then a correction for sun angle. All Landsat images were downloaded from https://earthexplorer.usgs.gov/ (accessed on 10 March 2021). Indices were calculated in ArcMap (Esri Inc., Redlands, California, U.S.). Terrain variables: All terrain variables were estimated from a digital elevation model (SRTM 1 Arc-Second Global; http://www.earthexplorer.usgs.gov; accessed on 16 November 2018). They were calculated in SAGA GIS. River distance was calculated from the 1:500,000 topographic map in ArcMap. Landcover map was made from semiautomated maximum likelihood classification of Landsat images in ArcMap. Noise removal was performed using the majority filter and focal statistics tools. Finally, manual post-classification processing and accuracy assessment were performed using test field data and available landcover data from the Global Land Cover service (https://lcviewer.vito.be/; accessed on 5 April 2021). Classification accuracy was 88% and Kappa was 0.85. Forest canopy cover was derived from the *treecover2010* product from the Global Land Analysis and Discovery (GLAD) service [68,127] (https://glad.umd.edu/; accessed on 26 April 2021). All environmental variables were prepared as ASCII raster maps with the help of ArcMap. For this, a polygon grid was created in ArcMap, covering the entire study area. The values of all environmental variables were set for each cell of this grid using the Zonal Statistic tool, and the proportion of each landcover type was calculated as a percentage. Then, separate rasters of all variables were created and translated into ASCII format. A set of 24 preliminary environmental parameters was created at a resolution of 30 m. • **Spatial extent:** 6277442.58999, 476938.137795, 515038.137795, 6251042.58999 (top, left, right, bottom). • **Spatial resolution:** 30 m. • **Projection:** WGS 1984, UTM zone 36N. • **Temporal extent:** Landsat images: 2014–2020. • **Data processing:** Radiometric correction for Landsat images.

	Before predictors were included in the models, they were additionally checked for outliers using a Cleveland plot [128]. • **Dimension reduction:** We used Spearman's rank correlation and variance inflation factor (VIF) analysis to avoid highly correlated variables. We only included the 15 noncollinear variables for each species to avoid overfitting for the models [69].
MODEL	
Variable pre-selection	The choice of initial environmental variables was made as a compromise between their availability and their ecological relevance as a direct or indirect proxy of species distributions. Only weakly correlated variables were included in each model.
Multicollinearity	Multicollinearity between predictors was investigated using Spearman's rank correlation coefficients and VIF. When variables were strongly related ($r > \lvert 0.7 \rvert$ and/or VIF > 10), we only retained one from each pair to minimize the possibility of overfitting.
Model settings	ModelFit: algorithm (maxent), featureSet (L, H, LQ, LQH, LQHP), featureRule (using a genetic algorithm; chosen based on spatial block 10-fold cross-validation on a list of linear, quadratic, product, and hinge features), regularizationMultiplierSet (1, 1.5, 2, 2.5, 3, 4, 4.5, 7.5), regularizationRule (chosen based on spatial block 10-fold cross-validation on a grid of regularization multipliers from 0 to 8 with increment of 0.5), convergenceThresholdSet (1.00E-05), samplingBiasRule (MCP), and iterations (1000).
Model estimates	Variable importance was calculated with the jackknife test removing one variable at a time. We used *with only TSS* for testing data as a measure of the importance of the variables.
Model selection / averaging / ensembles	Model selection was performed on the basis of AUC_{test} while tuning model hyperparameters. Then the *best* combination of features and RM was obtained to train the final model. Model averaging was performed on the basis of 10-fold block cross-validation.
Non-independence	To reduce the effects of sampling bias, all occurrence points were rarefied [44–46]. The clustering of points was assessed by calculating the average nearest neighbour index (NNI). Index values less than 1 indicate clustering, while values greater than 1 indicate dispersion [47]. Rarefied points satisfying the dispersed distribution were used as training points for building models. To account for spatial autocorrelation, we used block cross-validation [49] with spatial blocking strategy with a random pattern and 100 iterations [50], where the block size was determined from the median of the spatial autocorrelation range among all predictors (664 m for plants, 663 m for insects, and 1590 m for mammals).
Threshold selection	As a threshold for dividing continuous predictions into binary classes (presence/absence of a species), we used the threshold value maxSSS (maximum sum of sensitivity and specificity), which is considered to produce the best results for models based on presence-only data [24,92].
ASSESSMENT	
Performance statistics	We used multiple lines of independent evidence according to the "gold standard" from Araújo et al. [12]: spatial block cross-validation and held apart fully independent data. Predictive model performance on evaluation data was assessed using four different performance measures: area under the receiver operating characteristic curve (AUC), difference between training and testing AUC, true skill statistic (TSS), and continuous Boyce index (CBI). The TSS constitutes a threshold-dependent performance measure and was calculated using a TSS-maximisation threshold. Evaluation: trainingDataStats(AUC), testingDataStats (AUC), testingDataStats (AUCDiff), testingDataStats (trueSkillStatistic), testingDataStats (boyce).
Plausibility checks	In our pre-analyses, we used response curves and presence probability maps to understand model behaviour for different hyperparameter settings, and on the basis of these checks decided on intermediate model complexity.

PREDICTION	
Prediction output	**Prediction unit:** For further analyses, we used continuous averaged predictions of presence probability per species as well as predicted presence per species that were obtained by binarizing the predicted presence probabilities using the maxSSS threshold. For estimating prevalence of each food resource in the CFNR core area and its buffer zone we calculated area of its presence from a binary map for each type of territory and multiplied it by a correction factor (1.32 for the reserve core and 0.68 for the buffer zone). The correction factor was estimated from the proportional area of each territory type. To convert all of our binarized maps for all species into one map of food resource richness, we combined all of them and summed the number of food resources in each pixel. Prediction: output (cloglog), transferEnv1 (absolute probability), minVal (0.001), maxVal (0.991), thresholdSet (0.53, 0.40, 0.41, 0.28, 0.58, 0.43, 0.66, 0.49, 0.52, 0.45, 0.10, 0.59, 0.65, 0.59), thresholdRule (maxSSS).
Uncertainty quantification	• **Algorithmic uncertainty:** In addition, to evaluation the models based on independent datasets, we calculated the mean validation estimates and presence probability from 10-fold block cross-validation. We also calculate standard errors for all metrics and sd-intervals for the response curves following Araujo et al. [12]. This approach can reduce algorithmic-based uncertainty from SDMs.