

Waldock et al. (2023)

– ODMAP Protocol –

2023-11-06

Section 1: Overview

Authorship

Authors: Conor Waldock, Bernhard Wegscheider, Dario Josi, Barbara Calegari, Jakob Brodersen, Luiz Jardim de Queiroz, Ole Seehausen

Contact: conor.waldock@unibe.ch, conorwaldock@gmail.com

Webpage: <https://github.com/wyss-swiss-fish/xAI-demonstration>

Model objective

Model objective: We aimed to model the spatial distribution of fish species in Switzerland, identify the key factors explaining species distributions, and apply model-agnostic locally explainable artificial intelligence approaches to explain the how environmental factors spatially structure species geographic ranges.

Focal Taxon

Alburnoides bipunctatus, *Barbus barbus*, *Cottus gobio*, *Gobio gobio*, *Lampetra planeri*, *Oncorhynchus mykiss*, *Perca fluviatilis*, *Squalius cephalus*, *Thymallus thymallus*

Location

Switzerland.

Scale of Analysis

Spatial extent: Aare, Limmat, Reuss and Rhine catchments that form the majority of Switzerland's river network north of the Alps.

Latitude: 6.053483 to 10.230689 (EPSG:4326, WGS84)

Longitude: 46.325695 to 48.023463 (EPSG:4326, WGS84)

Temporal extent: Presence-absence records spanning from 2010 to 2023.

Boundary: Political and watershed boundaries of Switzerland.

Biodiversity data

Observation type: Presence-absence records came from quantitative and semi-quantitative field surveys conducted by scientific researchers, cantonal and national monitoring agencies, ecological consultancies.

Response data type: Presence-absence (0s and 1s).

Predictors

Predictor types: Bioclimate, habitat quality, water quality, flow regime, land-use, connectivity (see Table 3 for full details).

Hypotheses

Species environment relationships were expected to be multidimensional with multiple important factors that show spatial variation across species ranges. We expected variation between species in responses to environmental factors. We expected relatively consistent negative effects of human impacts but with a few species that could benefit from human related factors, most likely non-native species. Generally, we expected positive effects of temperature on warmer affinity families (Cyprinids, Perciformes) with biogeographic histories of post-glacial alpine recolonization from warm refugia. We expected that agriculturally dominated lowland Switzerland may be particularly depleted in local biodiversity, and therefore have low predicted habitat suitability, due to the combined effects of land-use change, water pollution, and river morphological modification, despite having higher potential for biodiversity because rivers exhibit higher discharge, shallow slopes, slower flows, and warmer water temperatures. We expected at cold-range edges temperature will be a dominant structuring factor of species geographic distributions, however, we expected the capacity for species to exist at their natural range limits was altered by natural and anthropic barriers that reduced connectivity of these river stretches.

Assumptions

Relevant biological, ecological, and human-related drivers of species distributions, or proxy variables, were included.

Sampling was adequate and representative of the biological, ecological, and human-related drivers of species distributions.

Quantitative surveys had few important detection errors, these did not covary spatially, or covary with environmental conditions.

Species were at equilibrium with their environment (no dispersal limitation beyond those included as covariates).

Biological interactions had a limited influence on species distributions.

Algorithms

Modelling techniques: We used down-sampled random forests as fit in Valavi *et al.* (2021) and shown to be one of the best models of species' spatial distributions in a large benchmarked review of presence-only species distribution models (Valavi *et al.* 2022). These models are designed to overcome the class imbalance issue often associated with over-fitting in random forest models, for full descriptions see Valavi *et al.* (2021).

Model complexity: Random forests are complex algorithms but the down-sampling approach is designed to avoid overfitting that can occur in random forests due to class imbalance (many more 0s than 1s) and lead to poor out-of-sample performance. We used a low number of trees (1000) and standard 'mtry' parameter of the square root of the number of variables to further avoid overfitting.

Model averaging: None applied

Valavi, R., Elith, J., Lahoz-Monfort, J.J. & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. Ecography, 44, 1731–1742.

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.

Workflow

Model workflow: We extracted the environment at the coordinates of all presence-absence records. To this data matrix, we fitted down-sampled random forests. Given the environmental conditions of 2km sub-catchments across the Aare, Limmat, Reuss and Rhine river network (within the borders of Switzerland) we made predictions, given the fitted random forest. This provided a prediction of habitat suitability per sub-catchment per species. The performance of all models was cross-validated by spatially blocking our data into 5-folds. For the predicted habitat suitability in every 2km sub-catchment we applied SHAPLEY analysis, which assessed the local contribution of environmental variables to species habitat suitability prediction.

Software

Software: R version 4.1.0 (2021-05-18), randomForest 4.7-1.1

Code availability: <https://github.com/wyss-swiss-fish/xAI-demonstration>

Data availability: Data came from a variety of sources with different public availability. Progetto Fiumi data are available on request from Dr. Jakob Brodersen at EAWAG. Dr. Pascal Vonlanthen provided a compiled dataset of private fish monitoring data compiled for Switzerland. Kanton Berne data are available on request from the responsible authorities. The authors of this document performed electrofishing surveys across the Aare catchment and data are available on request from the authors.

Section 2: Data

Biodiversity data

Taxon names: Taxa see Table 1 below.

Taxonomic reference system: We updated all taxonomic information based on primary literature and genetic barcoding of specimens from work completed at EAWAG and University of Bern in recent years to better delimit species distributions (Alexander et al. 2021; Brodersen et al. 2023). Note that “*Cottus gobio*” and “*Perca fluviatilis*” have high phenotypic and genetic variation between populations in Switzerland indicating distinct post-glaciation colonization waves (Vonlanthen et al. 2007; Lucek et al. 2018) and likely presence of multiple unrecognized species (Alexander and Seehausen 2021).

Alexander, T., & Seehausen, O. (2021). Diversity, distribution and community composition of fish in perialpine lakes. "Projet Lac" synthesis report. Eawag: Das Wasserforschungsinstitut des ETH-Bereichs. <https://doi.org/10.55408/eawag:24051>

Brodersen, J., Hellmann, J., & Seehausen, O. (2023). Erhebung der Fischbiodiversität in Schweizer Fliessgewässern. Progetto Fiumi Schlussbericht. Eawag: Swiss Federal Institute of Aquatic Science and Technology. <https://doi.org/10.55408/eawag:30020>

Lucek, K., Keller, I., Nolte, A.W. & Seehausen, O. (2018). Distinct colonization waves underlie the diversification of the freshwater sculpin (*Cottus gobio*) in the Central European Alpine region. *Journal of Evolutionary Biology*, 31, 1254–1267.

Vonlanthen, P., Excoffier, L., Bittner, D., Persat, H., Neuenschwander, S. & Largiadèr, C.R. (2007). Genetic analysis of potential postglacial watershed crossings in Central Europe by the bullhead (*Cottus gobio* L.). *Molecular Ecology*, 16, 4572–4584.

Ecological level: Population level (i.e., assuming records come from independent populations and variation between populations in relative habitat suitability depends on environmental conditions).

Biodiversity data sources:

- *Project fieldwork data (CW, DJ, BW, BC, OS):* Data were collected in the context of the University of Bern, Federal Office for the Environmental, Kanton Bern and Wyss Academy for Nature LANAT-3 project titled “*Stopping the biodiversity loss of water bodies - despite climate change*”.
- *Progetto Fiumi:* The main objective of Progetto Fiumi (Erhebung der Fischbiodiversität in Schweizer Fliessgewässern) was to survey fish biodiversity and its distribution in Swiss rivers and surveys were undertaken between 2013 and 2017.
- *Module Stufen Konzept electrofishing data:* Compiled by Dr. Pascal Vonlanthen and Dr. Sébastien Lauper in the context of compiling data collected under the biomonitoring protocol ‘Module Stufen Konzept’ (<https://modul-stufen-konzept.ch>).

- *Kanton Berne electrofishing data*: The Canton of Bern has provided us with an extract from its fisheries database for the period 1994 to 2022.

See Tables 1 and 2 for an overview of species presence records and available data from different monitoring datasets.

Table 1. Summary of presence records across different monitoring schemes in our analysis. University of Bern 2022 refers to our summer sampling campaign undertaken by (BC, BW, CW, DJ, OS). Progetto Fiumi was a Switzerland wide river biodiversity monitoring project coordinated by Jakob Brodersen and Ole Seehausen at EAWAG (1). Kanton Bern provided monitoring data undertaken by.. Consultancies refers to data compiled by Dr. Pascal Vonlanthen and Dr. Sebastien Lauper compiling various surveys from private and federal surveys and monitoring initiatives conforming to standardised electrofishing surveys (2).

	University of Bern 2022	EAWAG Progetto Fiumi ¹	Kanton Bern	MSK ²	Total
<i>Alburnoides bipunctatus</i>	10	7	32	61	110
<i>Barbus barbus</i>	12	24	34	99	169
<i>Cottus gobio</i>	30	59	558	247	894
<i>Gobio gobio</i>	11	11	34	79	135
<i>Lampetra planeri</i>	2	2	23	59	86
<i>Oncorhynchus mykiss</i>	2	3	66	11	82
<i>Perca fluviatilis</i>	6	8	33	33	80
<i>Squalius cephalus</i>	22	24	100	170	316
<i>Thymallus thymallus</i>	4	3	27	27	61

1. Brodersen, J., Hellmann, J., & Seehausen, O. (2023). *Erhebung der Fischbiodiversität in Schweizer Fließgewässern. Progetto Fiumi Schlussbericht*. Eawag: Swiss Federal Institute of Aquatic Science and Technology. <https://doi.org/10.55408/eawag:30020>
2. <https://modul-stufen-konzept.ch/en/fish/>

Table 2. Summaries of records in across databases.

	Number of species	Number of locations	Number of populations	First year	Median year	Last year	95 th -percentile year
1. University of Bern 2022	9	38	99	2022	2022	2022	2022
2. EAWAG Progetto Fiumi	9	70	141	2013	2014	2017	2013
3. Kanton Bern	9	655	871	2010	2016	2022	2010
4. Module Stufen Konzept	9	215	570	2010	2018	2020	2011

Sampling design:

- *Project fieldwork data (DJ, BW, BC, CW, OS)*: Electrofishing campaign from August to October 2022. We stratified sampling of sites based on 9 classes combining three human impacts levels (low, medium, high) and three elevation strata (low, medium, high) based on equal sized strata. From these strata, we sampled 46 sites and attempted to balance site selection across strata. We fished 100 meter stretches of wadeable streams with electrofishing at 46 sites in Switzerland. All fish were identified in the field, as best as possible, to species level, measured, and weighed. The determination of species from difficult groups was later confirmed in the laboratory or on the basis of the photos. Approximately 10-15 individuals of all

species (license permitting) were retained to form long-term research collections at the Naturhistorisches Museum Bern.

- *Progetto Fiumi*: Sites were surveyed between 2013 and 2017. Site selection was intended to representatively survey the diversity of Swiss watercourses. Surveys were conducted in September and October each year. A stretch of approximately 100 meters was quantitatively fished when rivers were small. Waterbodies that were too wide, too deep, or too fast-flowing were fished qualitatively, taking into account all habitats as much as possible. Non-wadeable, large rivers were sampled by boat equipped with an anode rake using the strip fishing method. Captured fish were stocked separately by strip, or by passage in the case of quantitative surveys. All fish were identified in the field, as best as possible, to species level, measured, and weighed. The determination of species from difficult groups was later confirmed in the laboratory or on the basis of the photos.
- *Module Stufen Konzept electrofishing data*: These data were compiled if they conformed to the electro-fishing component of the 'Module Stufen Konzept' (modular stepwise procedure) providing methods for the analysis and assessment of surface waters in Switzerland (<https://modul-stufen-konzept.ch>). These methods are developed to ensure assessment of watercourse status is in accordance with water protection legislation and was developed in collaboration with the federal government, cantonal government, EAWAG and the VSA. All data provided in this compilation had at least two electrofishing 'runs' recording all fish species encountered with the aim to describe the present species community (personal communication Dr. Pascal Vonlanthen and Dr. Sebastien Lauper).
- *Kanton Bern electrofishing data*: This mainly contains data collected at all construction sites during this time-period as well as some data from long-term monitoring sites.

Sample size:

Alburnoides bipunctatus (110); *Barbus barbus* (169); *Cottus gobio* (894); *Gobio gobio* (135); *Lampetra planeri* (86); *Oncorhynchus mykiss* (82); *Perca fluviatilis* (80); *Squalius cephalus* (316); *Thymallus thymallus* (61)

Absence data:

Combining all surveys together, we obtain 3,216 sampling events that provided absence records if species were not recorded on a survey.

- *Project fieldwork data (CW, BW, DJ, BC, OS)*: As far as logistically possible, we applied standardized fishing protocol that had to be adapted depending on river width, flow velocities and depth. Given that we collected all observed individuals and identified these at a species level we assume that missing species are absent from sampling events, and therefore sites.

- *Progetto Fiumi*: For absence data, we filtered Progetto Fiumi data to only quantitative and semi-quantitative sampling events, rather than non-systematic qualitative sampling events. Given that all observed individuals were identified at a species level we assume that missing species are absent from sampling events, and therefore sites.
- *Module Stufen Konzept electrofishing data*: All data provided in this compilation had at least two electrofishing ‘runs’ recording all fish species encountered with the aim to describe the present species community and as such we assumed all missing species were absent from sampling events.
- *Kanton Bern electrofishing data*: As far as we are aware, all species recovered were reported in electrofishing events, however, sampling occurred for different purposes depending on the monitoring needs that may introduce differences in the detectability of different species.

Background data: We used presence-absence data, so no background data were generated.

Data partitioning

Training data: To evaluate our models, we split our presence-absence datasets into 5-fold spatially blocked cross validation sets using the blockCV package in R (version 2.1.4; (Valavi *et al.* 2019)). We set the size of blocks to 10 km which was identified as the spatial scale of spatial autocorrelation in our environmental variables during preliminary analyses in our predictor set. We optimized our blocking grid to provide the most equal split in presence-absence across folds selected blocks, and arranged these blocks randomly.

Validation data: We did not use external validation data, but instead validated our models on the held-out cross validation set described above.

Valavi, R., Elith, J., Lahoz-Monfort, J.J. & Guillerá-Arroita, G. (2019). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods in Ecology and Evolution, 10, 225–232.

Predictor variables

For spatial distributions of relevant predictor variables see Figure S14 in Appendix 2. These are further summarized in detail below in Table 3. Our final models included 11 potential variables for model fitting, see Section 3 on model fitting below for further details on variable selection.

Table 3. Relevant details on predictor variable and processing steps. Coordinate reference systems were different amongst data products and harmonized to epsg:3035 equal area projection of Europe (ETRS89-extended / LAEA Europe). Final predictors used in our models are highlighted in bold.

Variable	Unit	Source	Time period	Spatial scale raw data	Temporal aggregation	Spatial scale in model	Spatial aggregation	Transformation
Monthly minimum	degrees C	CHELSA V2.1 climatology	1981–2010	1km	Minimum across years	STARS entire upstream	NA	NA

temperature						catchment		
Monthly maximum temperature	degrees C	CHELSEA V2.1 climatology	1981-2010	1km	Maximum across years	STARS entire upstream catchment	NA	NA
Monthly mean temperature	degrees C	CHELSEA V2.1 climatology	1981-2010	1km	Mean across years	STARS entire upstream catchment	NA	NA
Maximum discharge	m3/s	MQ-GWN-CH	1981-2000	river reach	Mean across years	BAFU subcatchment	Maximum per subcatchment	Log10+1
Minimum slope	%	swissALTI3D	NA	river reach	NA	BAFU subcatchment	Minimum per subcatchment	Log10+1
Flow velocity	m/s	derived	NA	river reach	NA	BAFU subcatchment	Mean weighted by river length per subcatchment	log+0.01
River morphological modification	Unitless index	BAFU Ecomorphology-F	2013-2016	river reach	NA	BAFU subcatchment	Mean weighted by river length per subcatchment	NA
Cropland cover	%	Corine Land Cover (CLC) 2018, Version 2020_20u1	2017-2018	25ha	NA	STARS reach contributing area	Proportion cover	NA
Floodplain	presence-absence	BAFU	2017	Polygon of object	NA	Local	Mean coverage per sub-catchment	NA
Wetland	Presence-absence	BAFU	2017	Polygon of object	NA	Local	Mean coverage per subcatchment	NA
Livestock unit density	500kg livestock per km2	BFS GEOSTAT (2008), Landwirtschaftliche Betriebszählung / Census of farming, Bundesamt für Statistik, CH-2010 Neuchâtel and Corine Land Cover (CLC) 2018, Version 2020_20u1	2008	25ha (CLC); 2km catchment (BAFU)	NA	STARS reach contributing area	Mean	NA
Tree cover density	%	High Resolution Layer - Tree Cover Density (TCD) 2018	2018	10m	NA	Local	NA	NA
Imperviousness density (urbanization)	%	High Resolution Layer - Imperviousness Density (IMD) 2018	2017-2019	10m	NA	Local	NA	NA
Diffuse nitrogen inputs in waters	total loss (kg/ha*a)	BAFU	2004-2009	100m	Model reference year 2010	STARS reach contributing area	Mean	NA
Diffuse phosphorous inputs into water	total loss (kg/ha*a)	BAFU	2004-2009	100m	Model reference year 2010	STARS reach contributing area	Mean	NA
Insecticide application rate	treatments per % cropland	BFS GEOSTAT (2008), Landwirtschaftliche Betriebszählung / Census of farming, Bundesamt für Statistik, CH-2010 Neuchâtel	2008	Polygon of municipality	NA	STARS reach contributing area	Mean	NA
Wastewater concentration	m3/year	BAFU Gewässerabschnittsbasierte Einzugsgebietsgliederung der Schweiz GAB-EZGG-CH	2014	2km catchment	NA	STARS reach contributing area	Mean	log10
Asymmetric colonization index (connectivity)	relative connectivity	derived from EU hydro data	NA	river reach	NA	BAFU subcatchment	NA	Log10+0.001
Distance to lake	Meters	derived from EU hydro data and BAFU barrier dataset	NA	river reach	NA	BAFU subcatchment	NA	Log10+1

Data processing

We compiled data on the spatial distribution of 19 environmental variables representing bioclimate, flow regime, land use, habitat quality, water quality and river connectivity from publicly available sources. In addition to selected variables, we evaluated the spatial scale of effects on freshwater fish species distributions based on review, elicitation, and discussion amongst co-authors and processed data according to the expected scale of effect (Table 3). Depending on the variable and the dataset, we calculated variables at four potential spatial scales, i) the local values 100m buffered from river, ii) aggregated to “reach contributing area” derived from openSTARS, iii) aggregated to “catchment area”

derived from openSTARS and iv) if data were originally provided aligned to Switzerland's Federal Office for the Environment geodata sub-catchments ($\sim 2\text{km}^2$) they are aggregated to this product for consistency. We processed our data to account for the directionality and structure of river networks by using the openSTARS workflow, an open source implementation of STARS using R and GRASS GIS (Peterson & Hoef 2014; Kattwinkel *et al.* 2020). We first derived a topologically correct river network from the European Digital Elevation Model (v1.1) with a burn-in defined by the EU-Hydro River Network Database (v1.3) setting the accumulation threshold parameter to 1500 and burn-in depth of 50m. We used openSTARS to identify the adjacent land area of each segment that provides the lateral overland flow, known here the "reach contributing area". We also derived the full upstream "catchment area" of each river segment, representing a larger spatial scale than reach contributing areas. We calculated the mean, sum and/or proportion of area at these two spatial scales. In addition to the openSTARS workflow, we used spatial operations in R to buffer stream segments by 100m and calculated the mean, sum, or proportion of area of each relevant variable at a local scale. Finally, we aggregated data to the 2km sub-catchment scale if provided by the Swiss geodata portal (e.g., BAFU Ecomorphology-F dataset).

Transfer data

We did not transfer these models to new temporal domains.

Section 3: Model

Multicollinearity

Multicollinearity: We used several strategies to limit the effects of multicollinearity in our modelling. However, we also recognized the inherent uncertainty in correlative statistical approaches to predict spatial ecological patterns which cannot perfectly accommodate causal inference from observational data (Arif & MacNeil 2022). As such, 'removing multicollinearity' by removing variables can artificially decrease uncertainty (Morrissey 2018). We therefore favored simpler models that were interpretable over more predictive models that had poor links between biological mechanisms for how each environmental factor affects species' populations and thus habitat suitability predictions. As such, we aimed to include the most direct drivers possible, but note that some known to be important variables were not available at the broad spatial scale required for creating spatially continuous predictions from species distribution modelling, and so could not be included in our analyses (e.g., cover of in-stream vegetation, river-bed properties, fine-scale spatial and temporal flow dynamics). We also chose not to reduce the dimensionality of our variables because we want to interpret the relationships more directly between drivers and species occurrences.

We ranked an initially more comprehensive list of variables ($n=31$) by importance and identified the expected spatial scale of importance ($n=3$; local scale, reach contributing area, upstream catchment). We selected in our final set of variables the highest-ranking variables that also represented independent ecological gradients without redundancy in variables, leaving 18 variables for further consideration (e.g., we excluded coniferous and

deciduous tree cover in favor of overall tree cover, we excluded the land use class ‘pasture’ in favor of a more quantitative livestock unit density, we removed elevation due to covariation with land use, connectivity and temperature).

We removed 6 of these selected variables which were strongly confounded with elevation, slope, river discharge or distance to lakes (see main manuscript). We removed these variables only when confounding effects could not be removed by the two-step modelling approach described in the main manuscript.

In the modelling steps, after the variable pre-selection, we applied the BORUTA algorithm to automatically select relevant variables when using random forest models. This algorithm was specifically developed to find all relevant variables, instead of building a predictive black-box model (Kursa & Rudnicki 2010). The BORUTA algorithm iteratively adds “shadow features”, which are randomized versions of original features, to random forests and removes real features with a lower Z-score than the maximum Z score of shadow features (full details in Kursa and Rudnicki 2010). We retained all features identified as ‘important’ using BORUTA R package (version 7.0.0).

Kursa, M.B. & Rudnicki, W.R. (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36, 1–13.

Model settings

Model settings (fitting): We fitted down-sampled random forests using the randomForest package in R (version 4.7-1.1). The number of trees was set to 1000. The number of parameters in each tree (mtry) was set as the default for classification problems (square-root of the number of variables). Most importantly, to fit down-sampled random forests, we set the ‘sampsize’ parameter to ensure both classes in the response data (0 and 1) had a sample size equal to lowest number of records in either class (as in (Valavi *et al.* 2021)).

Valavi, R., Elith, J., Lahoz-Monfort, J.J. & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. Ecography, 44, 1731–1742.

Model estimates

Coefficients:

Random forests do not output coefficients, but we interpret variable importance and effect using Shapley values.

Variable importance:

We calculated an indicator of global variable importance by calculating the mean absolute Shapley value per variable (as per Wadoux & Molnar 2022).

Variable effects:

We determined variable effects by plotting Shapley values against the original environmental values (as per Wadoux & Molnar 2022).

Analysis and Correction of non-independence

None

Threshold selection

For binary predictions, we selected the threshold that maximized the TSS (only used in Figure S5)

Section 4: Assessment

Performance statistics

We derived a set of metrics from contingency matrix containing true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

- Overprediction rate = $FP / TP + FP$ = proportion of all presence predictions that are incorrect.
- Underprediction rate = $FN / TP + FN$ = proportion of all absence predictions that are incorrect.
- Sorensen index = $(2TP) / (FN + 2TP + FP)$ = ignoring true negatives, what is the overlap between true and estimated classes.
- Jaccard index = $TP / FN + TP + FP$ = ignoring true negatives, what is the ratio between true and estimated classes.
- Sensitivity = $TP / TP + FN$ = proportion of all presence values that are correctly predicted
- Specificity = $TN / TN + FP$ = proportion of all absence values that are correctly predicted
- True skill statistic = sensitivity + specificity – 1 = balance of sensitivity and specificity
- Matthew's correlation coefficient =
$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} =$$
 contingency matrix version of Pearson's product moment correlation between predictions and observations.

In addition, we estimated the threshold-independent area under the receiver operative characteristic curve.

Plausibility check

Response shapes: We reviewed all response curves within the project team to ensure plausibility and refitted the species modelling pipeline removing variables that had implausible relationships. We did not remove variables that were against expectations but had theoretical foundations (i.e., while we expected human stressors to have, on average, a

negative effect, we did not remove cases with positive effects of human stressors can occur through indirect predator release). Instead, we removed highly non-linear responses that have no theoretical foundations or plausible explanation.

Expert judgement: All range maps were reviewed by the project team against input data and expert knowledge.

Section 5: Prediction

Prediction output

Prediction unit: We interpret the predicted response value as an environmental suitability index, although it is built on presence-absence data. This is because of our down-sampling procedure balances the presences and absences within each tree of the random forest such that the model predictions do not represent probability of occurrence.

Uncertainty quantification

Scenario uncertainty: We did not quantify future scenarios.

Novel environments: We did not quantify future scenarios so no novel environments exist.