Explainable AI for Species Distribution Models

Quantifying expected and shadow distributions using Shapley values

06 November 2023

Abstract

Appendix to the manuscript Waldock et al (2023) Shadow distributions reveal species' spatial sensitivity to anthropogenic threats and ecological constraints. Here we outline the application of Explainable AI tools to species distributions models and show the code for how we derived the shadow distributions in Waldock et al (2023) with more detailed descriptions.

Background

Using SDMs in fundamental and applied ecology Fundamental questions in basic and applied ecology depend on understanding why species live in certain locations, and not in others. The environmental conditions of a location are an important determinant supporting or preventing species occurrence. Species distribution models (SDMs) combine the distribution of environmental conditions and the distribution of species occurrences through a statistical model with the aim of calculating maps of species distributions. Note that such spatial models can also use other available biological response variables such as presence-absence, abundance or population growth rate (Figure 1).

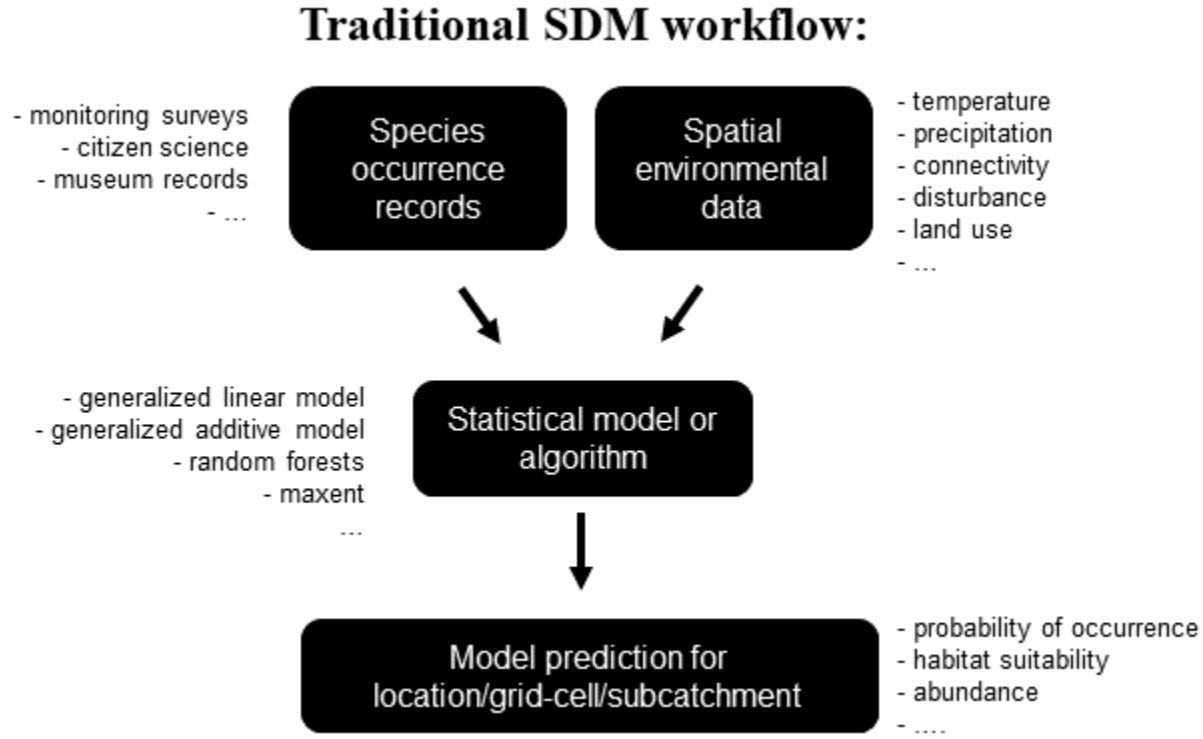

Figure 1. Broad overview of species distribution modelling frameworks

This workflow generates broad insights when assessing the overall model outputs and properties across all input observations in the model. This is referred to in the interpretable machine learning literature as investigating the *global model* scale, which commonly provides the following outputs:

- Mapping model predictions for unobserved areas to define a continuous map of the spatial variation in environmental suitability for a species.
- The overall importance of each variable for the models predictive accuracy e.g., permutation variable importance.
- The overall response of the species to each environmental factor.

Limitations of SDMs for fundamental and applied ecology Many ecological and conservation related questions demand a more detailed understanding of species distributions than the above scale of investigation can provide. For example, explaining why a population is expected to be present in a given location (with a high environmental suitability score) is very challenging. We simply know that a location has a high suitability, but it could be due to any of the environmental factors included in the model that have a positive effect on the species. Likewise, environmental managers often want to understand what would happen with a change in a given location. However, it is challenging to quantify why a species is expected to be present in a given location in the first instance, and therefore, it is hard to predict what an environmental manipulation might achieve and whether the manipulated factor is the main factor limiting a species locally.

We, therefore, lack knowledge on the importance of local ecological constraints and threat effects. Without this knowledge, it remains more difficult to effectively manage biodiversity. It would be useful to obtain localised insights from these broad scale models that highlight the influence of multiple co-occurring environmental factors on species distributions. This is the main challenge we attempt to provide one solution in Waldock et al. 2023.

To overcome this challenge and help explain why models make certain predictions, local explainable artificial intelligence approaches have been developed but have been rarely applied to ecological systems (but see (Ryo et al. 2021) and (Lucas 2020)). These approaches aim to provide explanations at the observation level of the model, elucidating why specific values for the variables in a model led to the prediction for that location. In our manuscript, we utilized Shapley values as they are model-agnostic and often yield more reliable estimates of local variable contributions compared with other methods (Scott M. Lundberg and Lee 2017a).

For a readable overview of local interpretations from explainable artificial intelligence and machine learning, please refer to the Local interpretable models chapter in this book by Christoph Molnar. For a detailed mathematical overview of the Shapley approach we implement see (Štrumbelj and Kononenko 2014a), and also (Scott M. Lundberg et al. 2018) for an alternative approach. Excellent tutorials for SHAP and kernelshap exist which provide overviews of the general approach and potential outputs. See also (Wadoux and Molnar 2022) for an application of Shapley values to ecosystem properties and (Scott M. Lundberg et al. 2018) for medical applications.

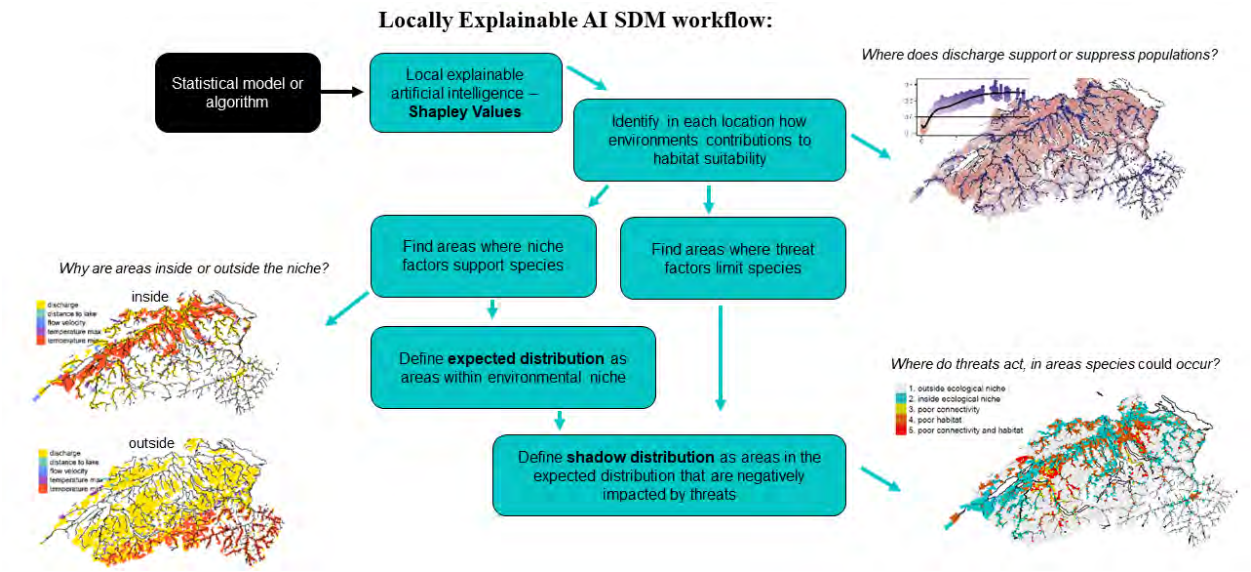

Figure 2. Overview of our application of explainable artificial intelligence to better understand drivers of species distributions.

We demonstrate in the code below how to implement the workflow from Waldock et al 2023. We briefly generate a species distribution model but focus on the analysis after the SDM has been created. We show the Shapley software implemented and summarize, for each observation level in our work (a river subcatchment), how we generated Shapley values that provide insights into the local contribution of each environmental variable to the overall model prediction. We then separated our variables into those assumed to characterize the natural niche of the species and those characterizing the threats expected to impact a species distribution. We then defined two new distributional concepts that are, to our knowledge, not quantifiable through traditional species distribution model workflows. We quantified the ***"expected distribution"***, defined as the areas within the natural niche of the species. We then identified the areas where threats contributed negatively to species distributions within this expected distribution. We refer to this as the "shadow distribution," as this property measures areas where species are living in the shadow of human influences.

Section 1. Traditional species distribution modelling approach

We first briefly outline the main inputs to species distribution modelling and generate spatially continuous predictions of environmental suitability in unsampled locations. We then later apply Shapley value analysis to the fitted species distribution model. As in our manuscript, we focus on the species Alburnoides bipunctatus shown below because it is a relatively widespread species in our catchments and is classified as 'Vulnerable' based on apparent population reduction and decline in the area and quality of habitat.

Load packages We first need to load the various packages used in this script. For full documentation of the package version used see the session info at the end of the document.

```
## LOAD PACKAGES

# local in pacman
if(!"pacman" %in% rownames(installed.packages())){install.packages("pacman")}
library(pacman)

# load in packages - version of packages used are displayed at the end of the workflow
p_load(tidyverse, terra, tmap, sf, randomForest, Boruta)
```

```
## LOAD OCCURRENCE AND ENVIRONMENTAL DATA

# occurrence dataset
pa <- st_read("scripts/workflow example/pa.shp", quiet  = T)

# environmental dataset
env_data <- rast("scripts/workflow example/env_data.tif")

# complete combined dataset
full_data <- readRDS(file = "scripts/workflow example/sp_example.rds")
```

1a. Species occurrence records The response variable in our modelling framework is the presence (1) or absence (0) of a species from a location. Our dataset has been compiled from electrofishing surveys where all species recovered from streams were identified. The Shapley value analysis, calculation of local variable contributions and the shadow distribution framework could also be applied to other types of data on ecological and biological responses to environmental gradients (such as presence-only, abundance, body size, trait properties, demographic rates etc).

```
## [1] "number of records = 3229"
```

```
## [1] "number of presence = 85"
```

```
## [1] "number of absence = 3144"
```

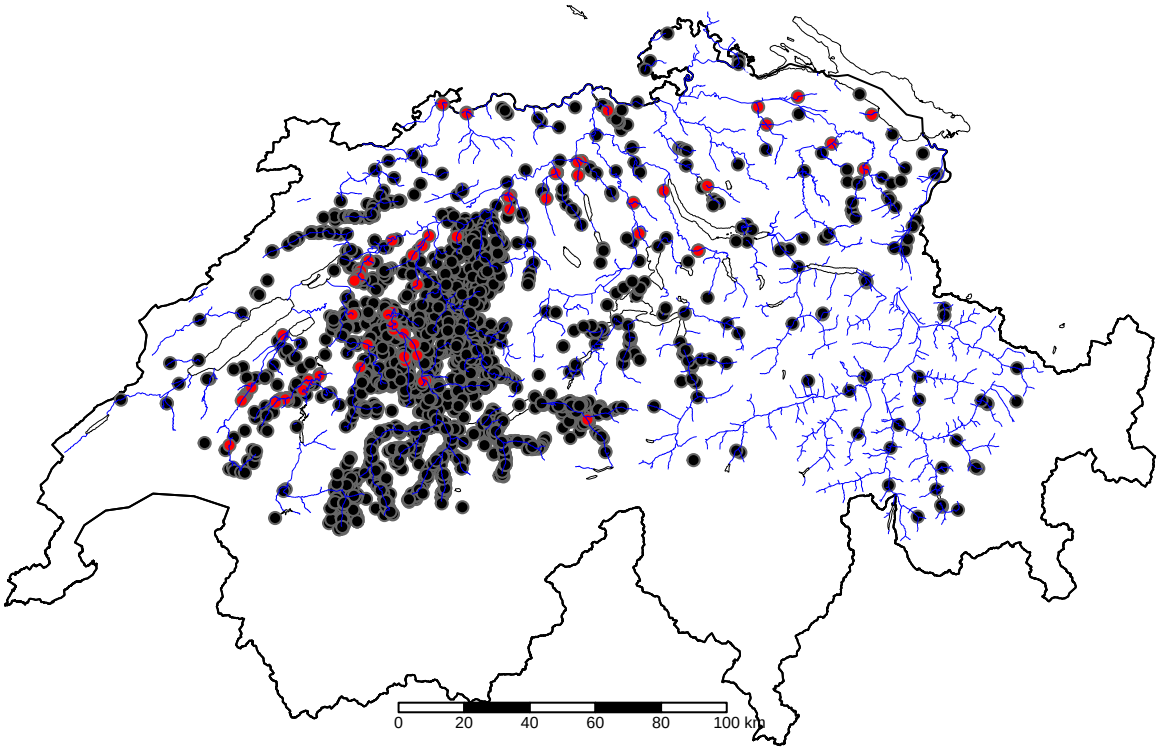

Figure 3. Spatial distribution of presence (red) and absence (black) points across Aare-Rhine riverscape.

1b. Spatial environmental data The second component of a species distribution model is the spatial environmental data. In our work, we compile freshwater specific variables for Switzerland that were available across the Aare catchment. However, there are many available data layers that have been used for SDM modelling purposes. This data has two purposes for SDM modelling: i) to provide explanatory variables for building relationships between occurrence and environmental gradients, ii) to provide spatially continuous predictions of occurrence probability or environmental suitability using the environmental values across full domain of interest (i.e., beyond the sampling locations). The environmental data are also important to the Shapley estimation as an input to estimate the local contribution of each environmental variable to the predicted suitability value.

Full name	Short name	Natural or threat
Monthly minimum temperature across 1981-2010	Minimum temperature	Natural
Monthly maximum temperature across years from 1981-2010	Maximum temperature	Natural
Maximum discharge in a subcatchment across years 1981-2000	Discharge	Natural
Minimum slope of river reach in subcatchment	Slope	Natural
Flow velocity of river reachs weighted-averaged within subcatchment	Flow velocity	Natural
Distance to lake	Distance to lake	Natural
River morphological modification (BAFU-Ecomophology F) weighted-average per subcatchment	Morphological modification	Threat
Floodplain proportion cover in subcatchment	Floodplains	Threat
Wetland proportion cover in subcatchment	Wetland	Threat
Imperviousness density COPERNICUS high resolution layer	Urbanisation	Threat
Asymmetric colonization index of river reach	Connectivity	Threat

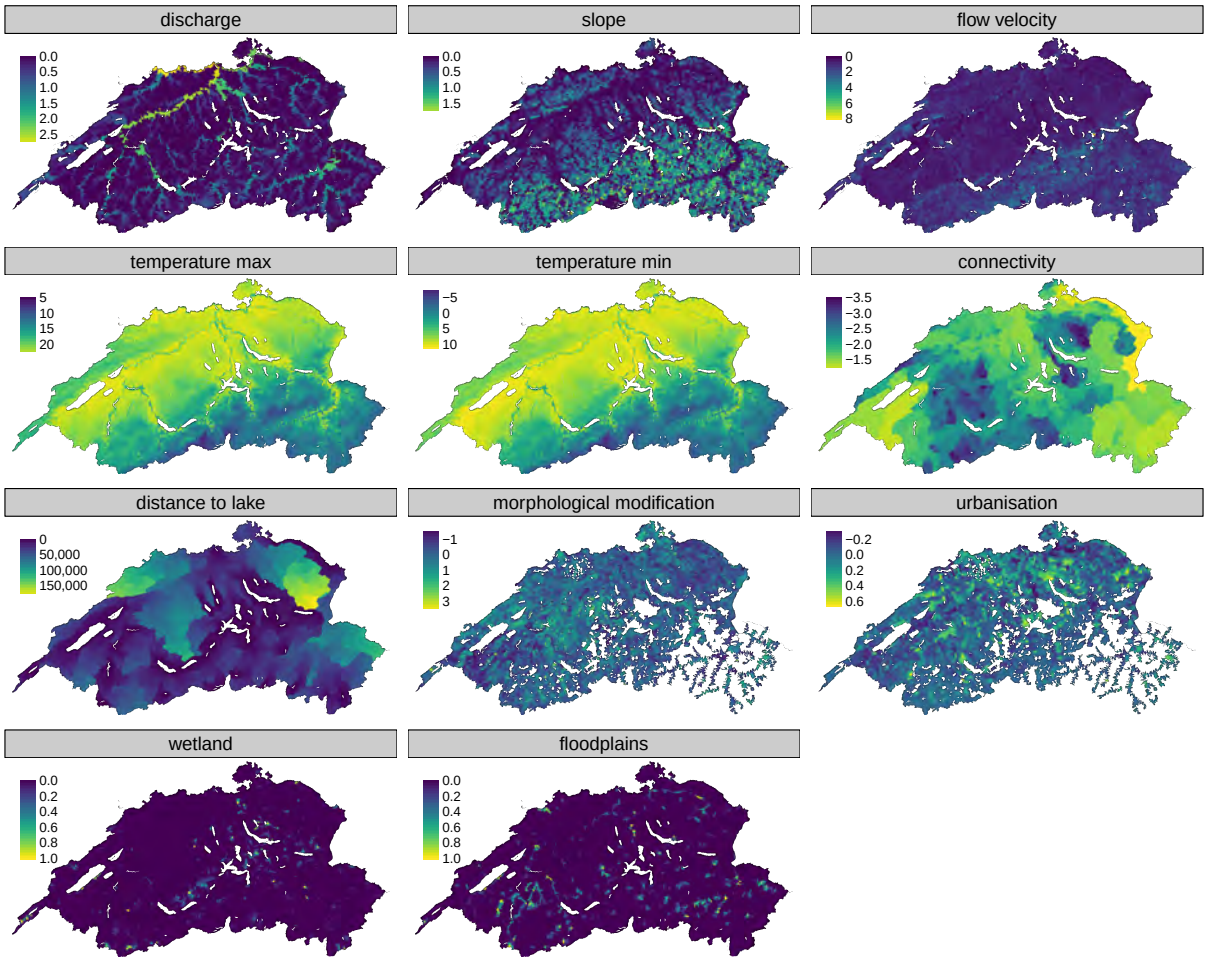

Figure 4. Map of the spatial environmental variables used to fit the species distribution model

1c. Species distribution model Here we used down-sampled random forests as in (Valavi et al. 2021) which provides good model performance when there are many more absences compared with presences. The explainable AI approach used here, Shapley values implemented in "fastshap," is agnostic to the exact underlying model. This is a major benefit as it could be applied to the commonly used ensemble modelling approaches e.g., (Thuiller et al. 2023). In the below script, we apply a custom wrapper for fitting down-sampled random forests using the approach in (Valavi et al. 2021) and using the "boruta" method for selecting variables (Kursa and Rudnicki 2010).

```
### RUN RANDOM FOREST MODEL

# Source functions that are wrappers to run random forests
# and variable selection available in this GitHub repository
funs <- lapply(list.files("scripts/workflow example/functions", full.names = T),
               function(x) source(x, echo = F))

# View data containing species occurrences and the relevant covariates
str(full_data, 2)
```

```
## 'data.frame':    3184 obs. of  14 variables:
##  $ occ                          : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
##  $ X                            : num  4141621 4121533 4159747 4123049 4119849 ...
##  $ Y                            : num  2718789 2670654 2717452 2635659 2642558 ...
##  $ ecoF_discharge_max_log10     : num  1.2977 0.0719 0.1931 0.0719 0.5289 ...
##  $ ecoF_slope_min_log10         : num  0.121 0.1404 0.0414 0.3964 0.0414 ...
##  $ ecoF_flow_velocity_mean      : num  1.606 0.778 1.046 0.494 0.589 ...
##  $ stars_t_mx_m_c               : num  21.2 23.1 22.9 19.3 20.2 ...
##  $ stars_t_mn_m_c               : num  8.25 9.9 9.59 6.27 7.16 ...
##  $ local_asym_cl_log10          : num  -1.78 -2.26 -1.69 -2.46 -2.46 ...
##  $ local_dis2lake               : num  83140 16145 52609 32454 23181 ...
##  $ ecoF_eco_mean_ele_residual   : num  0.99 -0.97 -0.342 -0.418 -0.618 ...
##  $ local_imd_log10_ele_residual: num  0.261 -0.102 0.049 -0.093 -0.199 ...
##  $ local_wet                    : num  0 0 0 0 0 ...
##  $ local_flood                  : num  0 0 0 0.303 0.485 ...
##  - attr(*, "na.action")= 'omit' Named int [1:45] 3 18 19 22 24 29 84 91 99 118 ...
##   ..- attr(*, "names")= chr [1:45] "3" "18" "19" "22" ...
```

```
# Generate species distribution model
var_selection_method = "boruta"
pa_rf_final <- rf_wrapper(full_data)
pa_rf_final
```

```
##
## Call:
##  randomForest(formula = occ ~ ., data = x[c("occ", rf_vars)],      ntree = 1000, sampsize = spsize, 
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 9.39%
## Confusion matrix:
##      0   1 class.error
## 0 2814 286  0.09225806
## 1   13  71  0.15476190
```

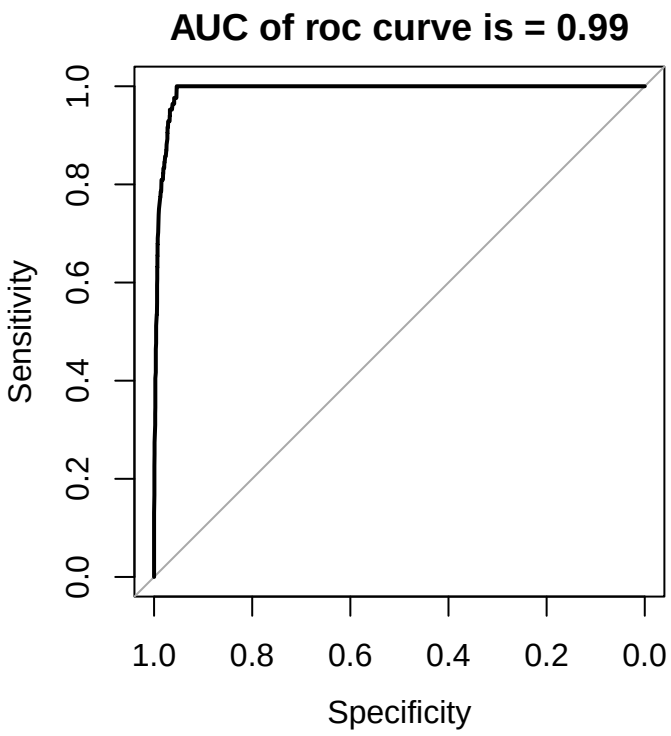

Figure 5. ROC curve of random forest SDM indicating within-sample performance

We do not aim to provide an in-depth overview of fitting and evaluating species distribution models and assume users are familiar with the biases, choices, and pitfalls involved (e.g.(Araújo et al. 2019; Sillero and Barbosa 2021; Guillera-Arroita et al. 2015)). In our manuscript, we use spatially-blocked cross validations and evaluate multiple metrics of model performance to assess model adequacy. However, our focus here is to demonstrate how to apply Shapley value analysis to SDMs in order to derive local contributions to environmental suitability scores and then quantify species shadow distributions.

1d. Environmental suitability predictions A key output of species distribution modelling is a spatially continuous map of the modelled response variable (e.g., occurrence, abundance). The name of this variable depends on the structure of the data and model, but is often called the "environmental suitability," "environmental suitability," or "probability of occurrence." Here, we used "environmental suitability"

```
### MAKE MODEL PREDICTIONS OF ENVIRONMENTAL SUITABILITY

# generate threshold using ecospat based on TSS
thresh <- ecospat::ecospat.max.tss(as.numeric(predict(pa_rf_final,
                                                      type = "prob")[,2]),
                                   as.numeric(pa_rf_final$y)-1)

# make prediction of environmental suitability from random forest model
habitat_suitability <- terra::predict(env_data, pa_rf_final,
                                      type = "prob")[[2]]
threshold_suitability <- as.numeric(habitat_suitability > thresh$max.threshold)

# extract the value of suitability per subcatchments (TEILEZGNR)
suit_sp <- terra::extract(habitat_suitability,
                          terra::vect(subcatchments_final),
                          fun = function(x) mean(x, na.rm = T),
                          touches = T)

# rename
suit_sp <- suit_sp %>%
  cbind(., subcatchments_final %>% select(TEILEZGNR)) %>%
  rename(., c("suitability" = "X1"))
```

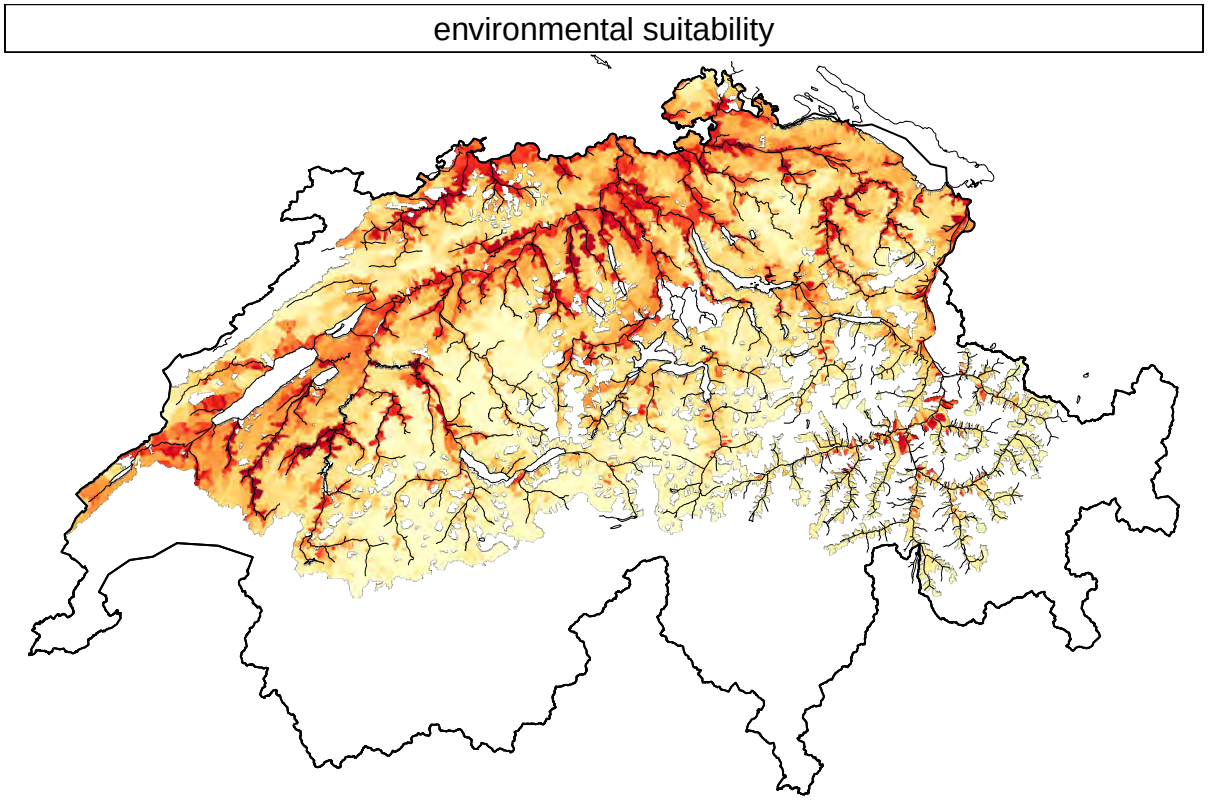

Figure 6. Environmental suitability prediction as the predicted values from our fitted species distribution model

One of the main limitations with the insight gained from a map of environmental suitability scores is that we do not know why the environmental suitability is high or low in particular areas. We only have a measure of how potentially suitable the habitat is in a specific location For many fundamental and applied questions it would be important to determine what are the specific environmental factors that contribute to a location being suitable or unsuitable. For example, under climate change predictions, it would be important to know that at range edges species are actually limited by fast flow velocities in rivers but not temperature. Or alternatively, before restoring a river, it would be important to know that low habitat complexity is an important factor leading to low environmental suitability, rather than cold water temperatures limiting the species locally.

Section 2. Shapley analysis of a Species Distribution Model

Here we calculate the Shapley values for all subcatchments based on the subcatchments environmental values and the random forest models. We aim to explain why the model makes a prediction of environmental suitability for a given subcatchment given the specific set of environmental conditions in that particular subcatchment.

2a. Calculate Shapley values of the random forest model Before running the Shapley analysis we need to set up some technicalities. We must load the fastshap package (Greenwell 2021) and define the prediction function used by the model. In addition, we must create a vector of all features/covariates names in the model.

```
### PREAMBLE TO SET UP SHAPLEY ANALYSIS

# load in the fast shap package used to calculate shapley values
p_load(fastshap)

# define the prediction function to use in fastshap
pfun <- function(object, newdata) {
 as.numeric(as.character(predict(object, newdata = newdata,
                                 type = "prob")[,2]))
}

# get the variables in a way that is model specific
vars <- colnames(attr(pa_rf_final$terms, "factors"))
```

We use the function `fastshap::explain()` in the fastshap package to run the Shapley analysis. This function takes the model object, the names of the features to explain, the data used to fit the model, new data to predict Shapley values, the prediction function to use, and the number of simulations to generate to create average Shapley values. In our main manuscript, we set the number of simulations to 10,000 and this number should be set as high as computationally possible to obtain the most accurate Shapley values. One simulation is the number of Monte Carlo repetitions used to generate the random coalition of variables used to generate the local contribution of any given variable as explained in the helpfile `?fastshap::explain` and in (Štrumbelj and Kononenko 2014b).

```
### RUN SHAPLEY VALUE ANALYSIS

# get the shapley values
shapley_pa <- fastshap::explain(

  # model object
  object = pa_rf_final,

  # names of features to explain
  feature_names = vars,

  # X data used to fit the model
  X = full_data[vars],

  # new data to predict on
  newdata = all_env_subcatchments[vars],

  # predictive function
```

```
  pred_wrapper = pfun,

  # number of replicates
  nsim = 100 # here should be set to as high as possible.
  # In the manuscript we used 10,000 taking a couple of days to run each time.

  )
```

The output of the Shapley analysis has the same column names as the covariates put in. To make it easier to process this data, we append the Shapley columns with "_SHAP" to differentiate from the environmental data values.

We create a combined spatial dataset with the subcatchments (TEILEZGNR), suitability scores from the random forests, the raw environmental data values, and the Shapley values.

```
### ORGANISE OUTPUT OF SHAPLEY ANALYSIS

# convert to a dataframe
shapley_pa <- shapley_pa %>% data.frame

# rename shapley data so it doesn't have the same names as the variables
names(shapley_pa) <- paste0(names(shapley_pa), "_SHAP")

# bind back in with the subcatchment environmental values
sp_shapley_pa <- cbind(all_env_subcatchments[c("TEILEZGNR", vars)],
                       shapley_pa)

# join in the spatial subcatchments in the Aare river catchment
# with the env. values and the shapley values.
shap_final <- left_join(left_join(subcatchments_final["TEILEZGNR"],
                                  suit_sp),
                        sp_shapley_pa)
```

Here we now have a dataframe that provides for each subcatchment the value of the environment, the modelled environmental suitability, and the effect of the subcatchment environment on the environmental suitability score (the Shapley value).

Section 3. Insights gained from Shapley values

Here we generate multiple summaries based on Shapley values including:

- 3a Variable contribution importance
- 3b Variable contribution direction
- 3c Species response curves
- 3d Spatial distribution of variable contributions
- 3e Relative contribution of variables in one location

3a. Average Shapley value across all subcatchments (variable contribution importance) The average absolute value of Shapley values gives an estimate of the overall variable importance in the model. For each subcatchment the Shapley value is turned to an absolute value so that the direction of the contribution (positive or negative) is ignored when calculating the overall importance. A variable has a low overall importance when most of the Shapley values have a low value. This is because this variable is not contributing significantly to a change in the environmental suitability score on average, and is therefore interpreted as globally unimportant. Note that, this global interpretation does not exclude that in some local subcatchments a variable that is unimportant on average can have a very strong local effect.

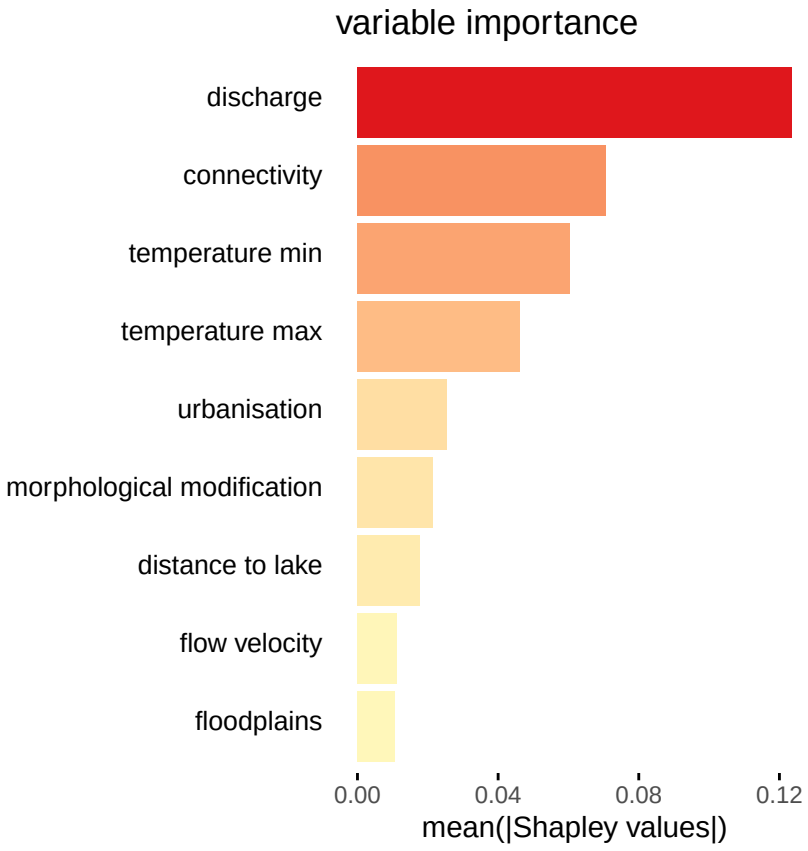

Figure 7. Shapley based variable importance scores

3b. Variation in Shapley values per environmental gradient (variable contribution direction) We can also show the distribution of the Shapley values per variable which shows if the overall effect of the variable is generally positive or negative. This approach does not show the relationship between the variables but instead the general impression of negative or positive contributions. For example, in the below plot we see that discharge contributes negatively in most subcatchments to environmental suitability. In contrast, temperature variables tend to have a more balanced distribution of effects across all subcatchments, some positive and some negative.

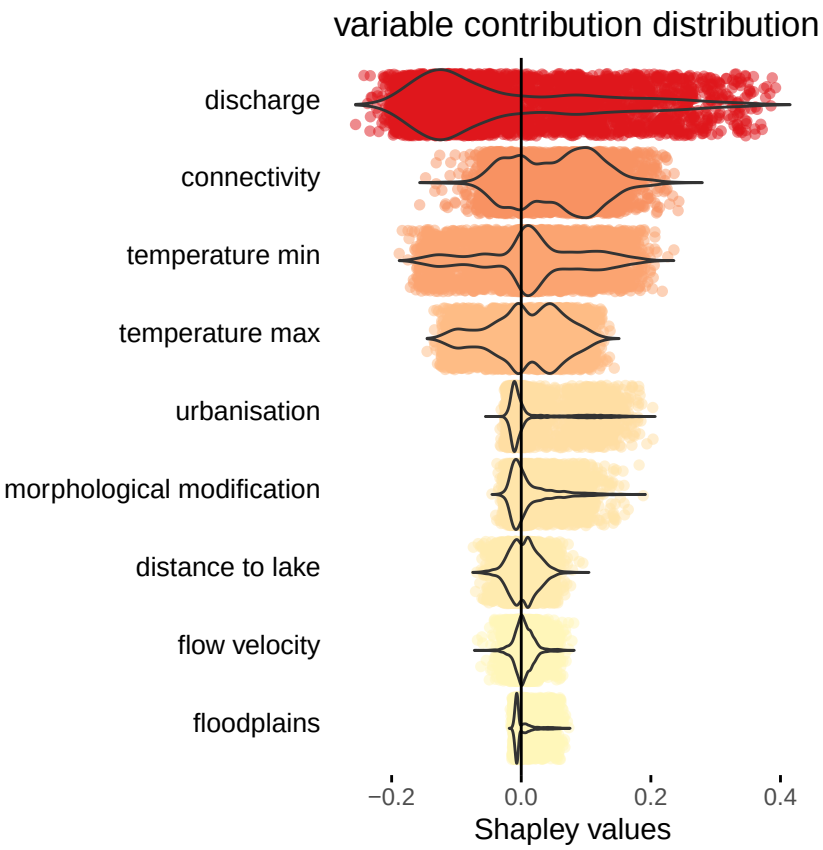

Figure 8. Shapley based distribution of variable contribution scores

3c. Species responses to environmental gradients (response curves) The Shapley value at a given environmental value indicates the contribution to the overall suitability, as summarised in 3a and 3b. We can also look at whether a variable has a positive or negative effect on environmental suitability using Shapley values (i.e., species response to environmental gradients).

When a variable has a positive overall effect on environmental suitability this indicates higher Shapley values at higher values of the environmental gradient (and vice versa). In this way, Shapley values plotted against environmental values indicate response curves much like other techniques for plotting these curves (e.g., predictions from models over all values of environmental gradients, Accumulated Local Effect plots).

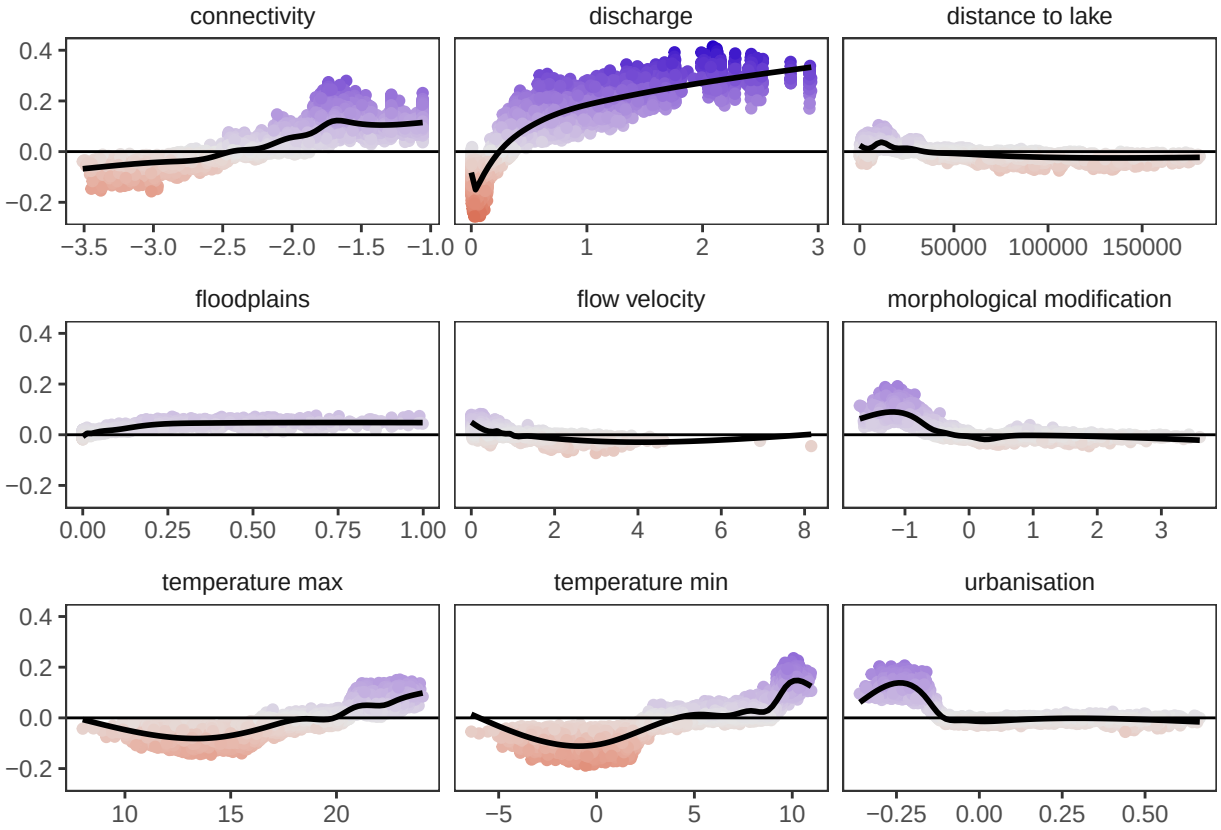

Figure 9. Shapley based response curves

3d. Spatial distribution of Shapley contributions to environmental suitability Given that we estimate the Shapley value per subcatchment, and we have spatial information on the subcatchments across the whole landscape, we can then predict the spatial distribution of the effect of a given variable on the environmental suitability. This can be interpreted as the *spatial distribution of species sensitivity* to each environmental variable.

This insight can be used to generate fundamental knowledge on the main ecological niche constraints on species distributions across a landscape. If used in more applied domains, we can see the main areas that are impacted by each threat which could help inform conservation decision making to find locations where species populations respond negatively to threats.

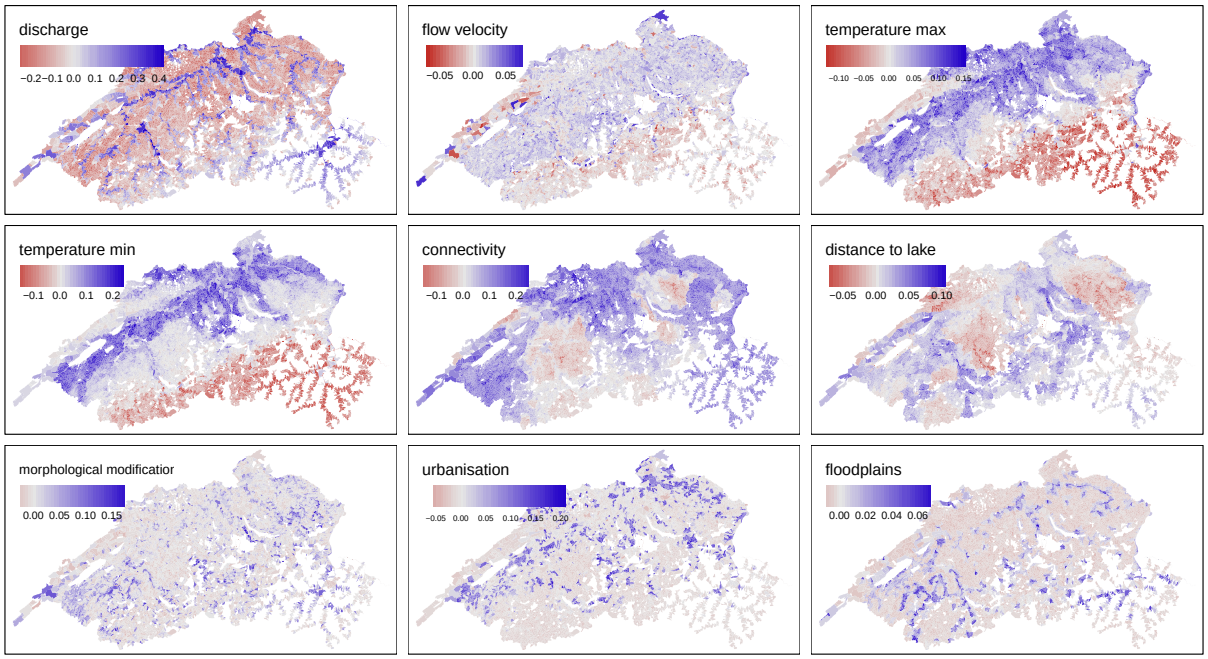

Figure 10. Map of Shapley values across Aare-Rhine subcatchments

3e. Relative contribution to environmental suitability in specific locations We demonstrate how shapley values can be summarised to identify the main environmental constraints on species distributions in a specific sub-catchment. We first must identify the subcatchments of interest, which here we choose the Sense river in Switzerland which has a relatively high degree of connectivity, natural and largely unmodified flow regime. We extracted the Shapley values for all variables for all subcatchments that fall in the Sense drainage and took the mean Shapley value across subcatchments.

First we identify the subcatchments that fall in the Sense river:

```
#### ESTIMATE SHAPLEY VALUE BASED ON SINGLE CATCHMENT

# take one subcatchment in the sense river
test_subcatchment <- 79104

# get the shapley values
shapley_sense <- fastshap::explain(

  # model object
  object = pa_rf_final,
  # names of features to explain
  feature_names = vars,
  # X data used to fit the model
  X = full_data[vars],
  # new data to predict on
  newdata = all_env_subcatchments %>%
    filter(TEILEZGNR == test_subcatchment) %>%
    select(vars),
  # predictive function
  pred_wrapper = pfun,
  # number of replicates
  nsim = 10000, # here should be set to as high as possible.
  # In the manuscript we used 10,000 taking a couple of days to run each time.
  )

shap_sense <- pivot_longer(data = data.frame(shapley_sense),
                           cols = vars)
```

```
#### SHOW SECOND METHOD FOR CALCULATING SHAPLEY VALUES (GIVES IDENTICAL RESULT)

# other approaches are available to calculate shapely value
# explanations, but here appear to give the same result.
library(kernelshap)
SHAP <- kernelshap(object = pa_rf_final,
           X = all_env_subcatchments %>%
             filter(TEILEZGNR == test_subcatchment) %>%
             select(vars),
           bg_X = full_data[vars],
           pred_fun = pfun,
           exact = T)
SHAP_df <- SHAP$S
SHAP_df <- data.frame(SHAP_df)
```

Next, we estimate the mean predicted environmental suitability which is our baseline on which the Shapley values modify environmental suitability in the Sense river:

```
#### CALCULATE MEAN MODEL PREDICTION ACROSS ALL SUBCATCHMENTS

# create baseline as the mean prediction of the model output
baseline_prediction <- mean(as.numeric(predict(pa_rf_final))-1)
baseline_prediction
```

```
## [1] 0.1121231
```

Next, we estimate the mean predicted environmental suitability of the Sense catchment. Once the shapley values are summed, this is what the estimated environmental suitability converges.

```
#### CALCULATE LOCAL ENVIRONMENTAL SUITABILITY

sense_prediction <- mean(shap_final %>%
                           filter(TEILEZGNR == test_subcatchment) %>%
                           pull(suitability), na.rm = T)
sense_prediction
```

```
## [1] 0.6388616
```

We see clearly that the average prediction for the Sense is much higher than the average suitability across all other locations in the Aare and Rhine catchments. A critical question is, why is this habitat better for this species?

We can investigate this question by looking at the model predicted deviation from the mean suitability. This visualises the contribution of each variable to the environmental suitability in the Sense, given this rivers particular environmental conditions, which helps explain why the suitability score is predicted from the model.

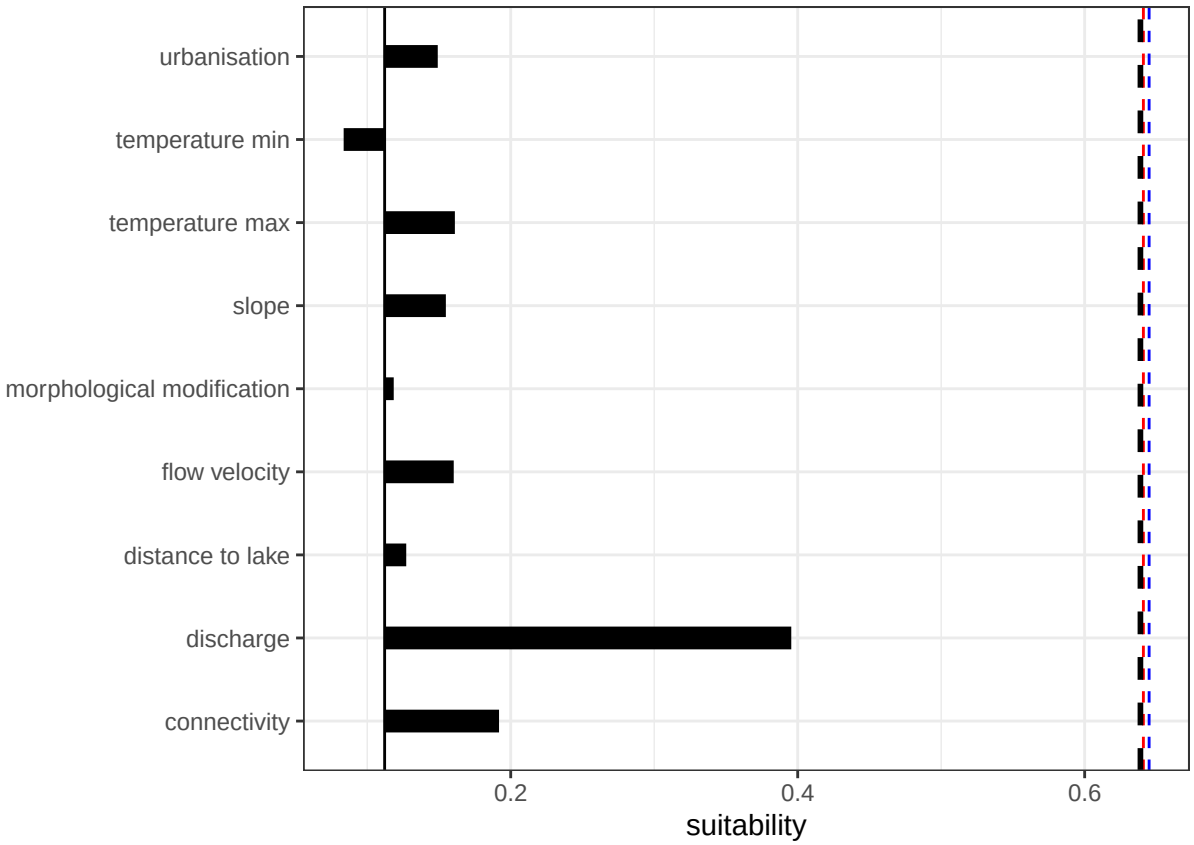

Figure 11. Shapley contributions to model local prediction (dashed lines) expressed as deviation from baseline model prediction. Red and blue dashed lines indicate the summed shapley values + baseline prediction for fastshap (blue) and SHAP (red) methods, giving almost identical results.

The plot above indicates how each variable contributes to the deviation from the overall baseline prediction (0.11) to obtain the predicted suitability for this location (0.64). By estimating local explanations, we can understand why we obtained this prediction. The Shapley values highlight which variables at this location contributed to the difference between the baseline average prediction and the local suitability prediction. We observe that discharge and connectivity made positive contributions to the local prediction, along with urbanization, flow velocity, and slope, all contributing positively as well. Additionally, we notice a few negative contributions from habitat-related threats, indicating that these factors are unlikely to be limiting species occurrence in these locations. In our manuscript, we contrast two river systems with different ecological and anthropic conditions and their impact on species occurrence locally.

Section 4: Shadow distributions

We next demonstrate the calculation of shadow distributions based on Waldock et al. (2023). We introduce the concept of a shadow distribution as the areas where a species would be expected to occur but where threats negatively affected the species. We called this the "shadow" distribution to reflect that species are in the shadow of human influences.

Two types of shadow distributions We define two types of shadow distributions: the *"binary shadow distribution"* and the *"quantitative shadow distribution"*, described in more detail below. The key components to quantify each shadow distribution are:

i) a set of variables that define the ecological niche of the species, and

ii) a set of variables that define threats to a species.

In any given system, the factors defining the niche and the threats to species must be informed by well-grounded ecological theory and expert knowledge within the specific system.

Defining expected distributions from abiotic niche factors The ecological niche is defined by the relevant abiotic or biotic environmental factors that are expected to naturally determine individual fitness and population performance. The geographic space falling inside the ecological niche is defined as the *expected distribution*. The actual or realized distribution may deviate from this expected distribution due to human impacts within the species' distribution. The threat factors quantify human-related changes to the environment that are expected to reduce individual fitness and population performance, although it's important to note that for some species, threat effects may be positive.

In more specific terms shadow distribution is region within the expected distribution (defined by the natural niche) where human impacts negatively affect species. This property is expected to deviate strongly from traditional predictions from species distribution models (i.e., environmental suitability) if threat effects are important determinants of habitat suitability. This deviation occurs because the environmental suitability score combines the effects of all variables on a model prediction, including threats that reduce suitability in areas that would otherwise be suitable based on ecological niche factors alone. In contrast, we utilize Shapley values to calculate the contribution of each variable to the prediction. This separation of variable effects enables us to quantify the negative contribution of threats within areas where natural niche factors positively contribute to model predictions.

We will now outline how we calculate the expected distribution, as well as the species' binary shadow distribution and quantitative shadow distribution separately.

4a. Expected distribution to define baselines Before calculating the shadow distribution, we must quantify the expected distribution, and do so using binary and quantitative representations. The binary expected distribution refers to whether a location is inside or outside abiotic niche of species. The quantitative expected distribution is the environmental suitability of the areas inside the niche.

First we must define the natural niche factors. These are factors that we expect to define the natural conditions that constrain a species distribution, and therefore help define the realized ecological niche of the species. In many respects, this definition is somewhat subjective and should reflect your expertise on the focal species in addition to how well the variables represent the key non-human related ecological processes constraining a species distribution. We also opted to include habitat variables, such as the presence of "nationally important floodplains," as a threat in our exercise. This decision was taken because most subcatchments are not inside this category so the absence of floodplains can be perceived as having a negative impact on biodiversity.

```
#### DEFINE THE NATURAL NICHE FACTORS (THAT ARE USED IN THE MODELS)

# define natural niche factors
natural_niche_factors = c("ecoF_discharge_max_log10_SHAP",  # discharge
                          "stars_t_mn_m_c_SHAP", # minimum temperature
                          "stars_t_mx_m_c_SHAP", # maximum temperature
                          "ecoF_flow_velocity_mean_SHAP", # flow velocity
                          "local_dis2lake_SHAP",  # distance to lake
                          "ecoF_slope_min_log10_SHAP") # slope

# subset the natural niche shapley values relevant to the specific species model
natural_niche_factors <- natural_niche_factors[natural_niche_factors %in%
                                                 names(shap_final)]
```

Here we next define areas as inside or outside of the natural niche of the species. We define subcatchmets as falling inside the natural niche of the species if the ***sum of the natural niche variable Shapley values is > 0.***

```
#### DEFINE BINARY EXPECTED DISTRIBUTION

# is the sum of the natural niche variable Shapely values positive?
shap_final$natural_niche <-
  rowSums(st_drop_geometry(shap_final[,natural_niche_factors])) > 0

# what is the sum of the natural niche variable Shapley values?
shap_final$natural_niche_value <-
  rowSums(st_drop_geometry(shap_final[,natural_niche_factors]), na.rm = T)
```

We ask whether what proportion of subcatchments is the sum of natural niche factor Shapley values positive (i.e. > 0). This highlights a simple summary that around `round(sum(shap_final$natural_niche==0, na.rm = T) / sum(!is.na(shap_final$natural_niche), na.rm = T)*100)` of subcatchments are outside of the ecological niche of the species.

```
#### MAKE TABLE OF INSIDE OR OUTSIDE OF NICHE

signif(table(shap_final$natural_niche) / nrow(na.omit(shap_final)), 3)
```

```
##
## FALSE  TRUE
##  0.62  0.38
```

We next look at a histogram of the summed of the natural niche Shapley value contributions. This is a continuous representation whether a subcatchment is inside or outside of the ecological niche. We see by how much subcatchments have negative or positive contributions of all natural niche variables to the overall environmental suitability prediction.

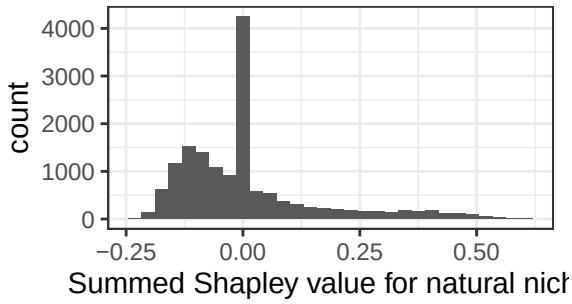

Figure 12. Distribution of summed Shapley value contributons to natural niche factors

The above summaries highlight, broadly, how natural niche factors contribute positively to environmental suitability predictions. Because each of the data points underlying these distributions is a geographic unit, we can therefore geographically map either binary areas or continuous areas where we may expected species to have suitable environments based on abiotic niche factors - the expected distribution. We can contrast this with the outputs of a traditional species distribution model which only shows the relative environmental suitability.

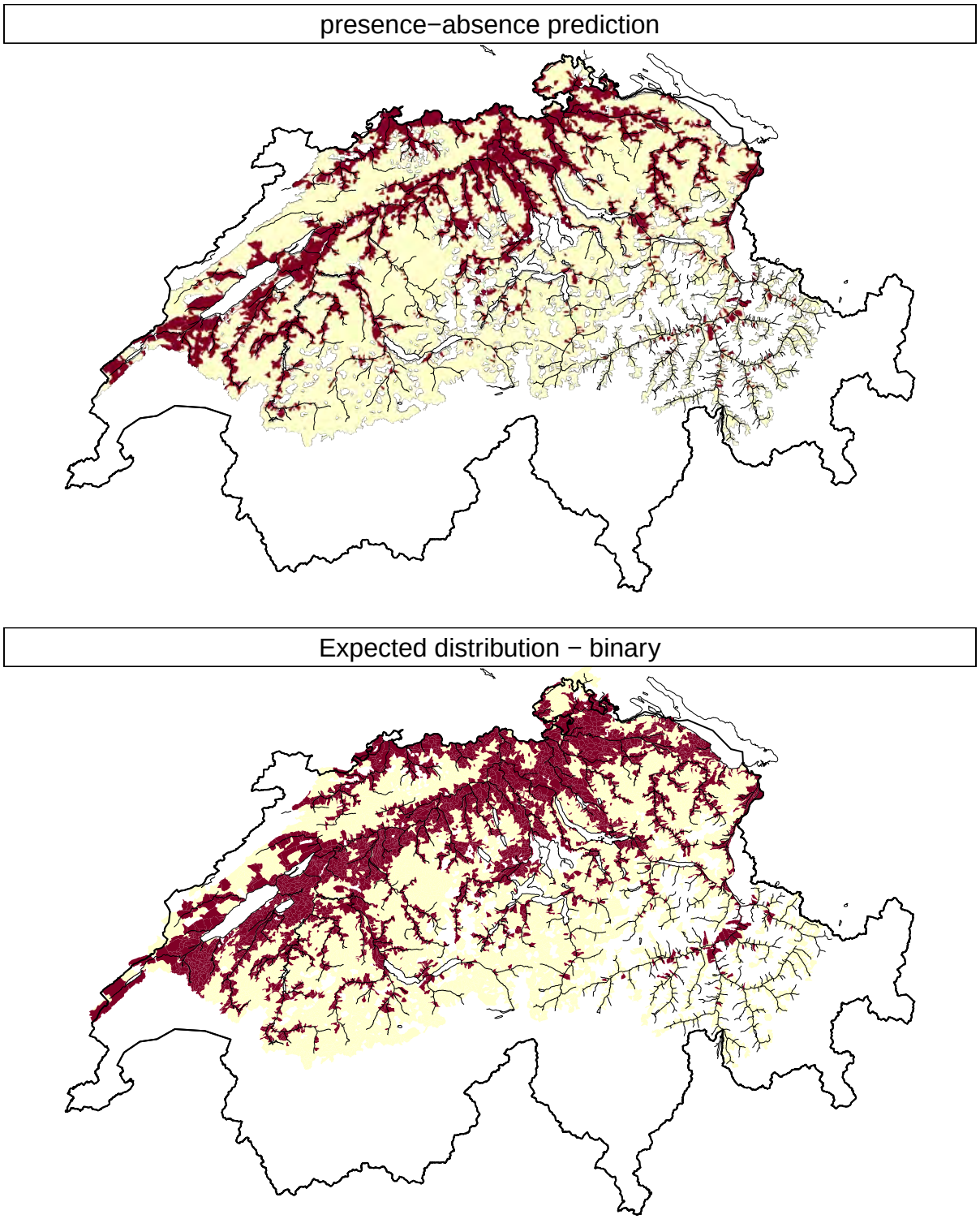

Figure 13. Comparison of predicted distribution of presence-absence from traditional species distribution model and the expected distribution quantified by Shapley values for natural niche factors

In the above plot, we see the contrast between the environmental suitability scores and species predicted presence (red) or absence (yellow) by thresholding these scores. These two plots take into account all variables

effects at once on the environmental suitability and presence-absence prediction. In contrast, we define the binary expected distribution only by the natural niche variables.

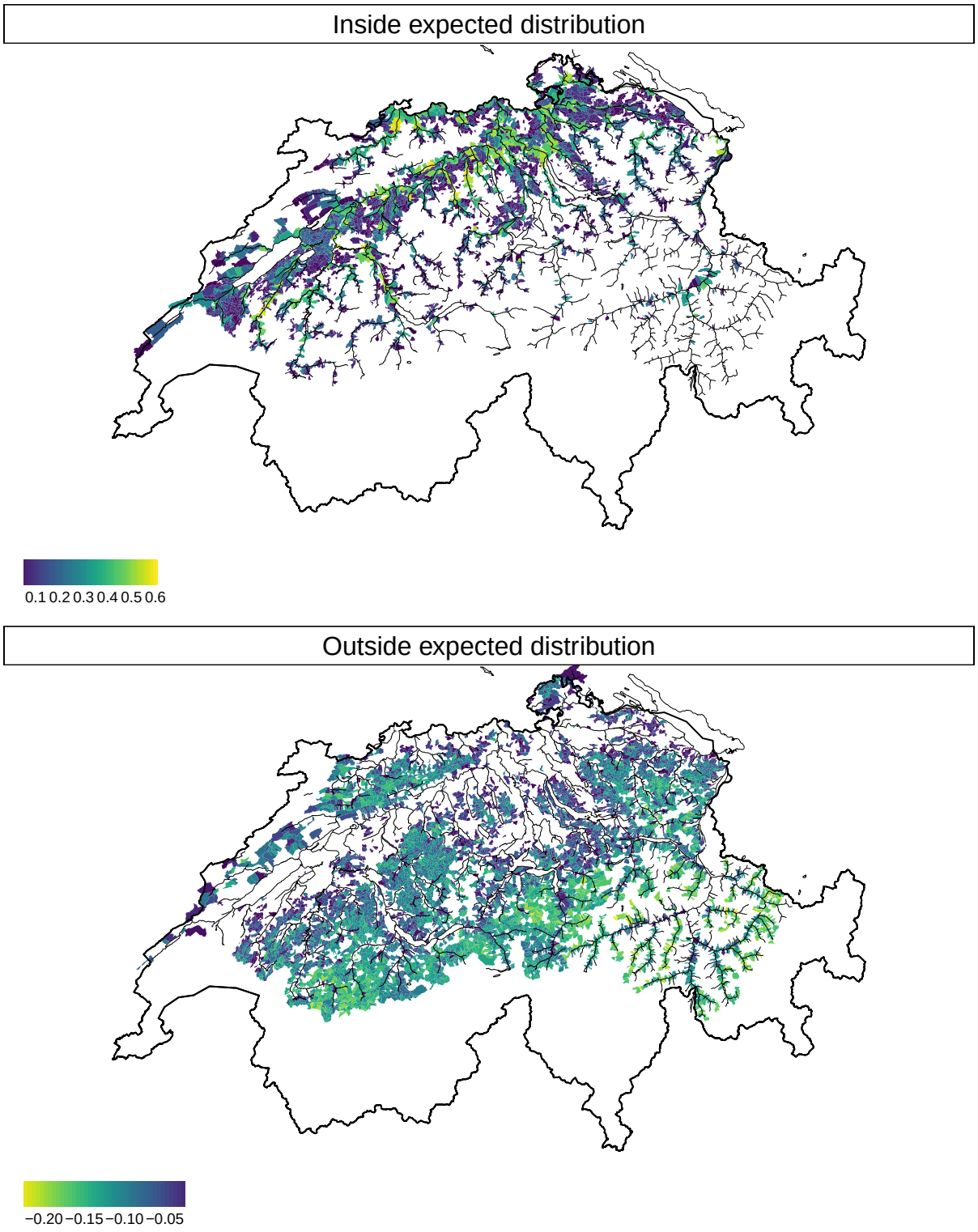

Figure 14. Comparison of areas inside and outside of binary expected distribution expressed as the habitat suitability scores inside and outside of the expected distribution.

The distributions in the above plot highlights the novelty of producing expected distributions, in addition to realized environmental suitability predictions, because we can now determine where a species should occur, but does not (i.e., the shadow distribution). Note that, fitting a species distribution model with only the niche related factors also does not solve this challenge. This is because the variables related to human factors still influence the species distribution. Therefore, simply ignoring these variables can confound the fitted model, the recovered response curves for the niche factors, and the resulting environmental suitability predictions. Instead, we fit the model including all the variables and then partition their individual effects at a local scale using the Shapley values to recover the regions where niche factors positively contribute to species distributions.

Now that we have defined the expected distribution, we can define the shadow distribution by expressing the effect of threats within the expected distribution.

4b. Binary shadow distribution We define the binary shadow distribution as the presence or absence of negative threat factor Shapley values within the expected distribution. As such, for a given sub-catchment, we first determine if the distribution first falls within expected distribution. If within the expected distribution, we determine within the catchment there is a negative Shapley value for a threat factor, indicating this threat negatively contributes to the environmental suitability prediction, despite a positive contribution of natural niche factors.

We determine two threat categories of "habitat-loss" threats and "connectivity" threats. The following variables were habitat-loss related threat factors in our framework: (low) floodplain cover, (low) wetlands cover, (high) river morphological modification index, and (high) urbanisation. We considered these together by average the net effect of habitat threats. We considered the connectivity alone and therefore take the raw Shapley value.

Define threat factors:

```
# define the variables that fall into each threat category
habitat_threat <- c("local_wet_SHAP", # wetland proportion cover
                    "local_flood_SHAP",  # floodplain proportion cover
                    "local_imd_log10_ele_residual_SHAP", # urbanisation proportion cover
                    "ecoF_eco_mean_ele_residual_SHAP") # river anthopic modification index

connectivity_threat <- c("local_asym_cl_log10_SHAP")
```

```
# add column to identify whether contribution of habitat
# variable to environmental suitability is positive or negative
shap_final$neg_habitat <-
  rowMeans(st_drop_geometry(shap_final[which(names(shap_final) %in% habitat_threat)])) < 0

# add column to identify whether contribution of connectivity
# variable to environmental suitability is positive or negative
shap_final$neg_con <-
  rowMeans(st_drop_geometry(shap_final[connectivity_threat])) < 0
```

Proportion of catchments inside niche with negative habitat effect (==T):

```
table(shap_final$neg_habitat) / sum(!is.na(shap_final$neg_habitat))
```

```
##
##     FALSE      TRUE
## 0.4757013 0.5242987
```

Proportion of catchments inside niche with negative connectivity effect (==T)

```
table(shap_final$neg_con) / sum(!is.na(shap_final$neg_con))
```

```
##
##     FALSE      TRUE
## 0.7393181 0.2606819
```

Binary shadow distribution: define threat categories We then used these contribution scores to demonstrate the areas of a species distributions that fall into the following categories:

1. outside of the expected distribution
2. inside the expected distribution - with no threats
3. inside the expected distribution - negative mean contribution of habitat variables
4. inside the expected distribution - negative mean contribution of connectivity variable
5. inside the expected distribution - negative mean contribution of habitat and connectivity threats

We would define the categories 3-5 as falling inside the shadow distribution of the species. We chose the above categories to help understand the major threats that lead to a location falling into the shadow distribution. For any given system, species, or threat landscape, these categories can be adapted.

```
# We used the following logical statements to develop this categorization - which of course should be a
  shap_final$niche_categories <- as.factor(

    ifelse(shap_final$natural_niche == T &
             shap_final$neg_con == F &
             shap_final$neg_habitat == F,

           "2. inside expected distribution + no threat",

           ifelse(shap_final$natural_niche == T &
                    shap_final$neg_con == T &
                    shap_final$neg_habitat == F,

                  "3. shadow - poor connectivity (C)",

                  ifelse(shap_final$natural_niche == T &
                           shap_final$neg_con == F &
                           shap_final$neg_habitat == T,

                         "4. shadow - poor habitat (H)",

                         ifelse(shap_final$natural_niche == T &
                                  shap_final$neg_con == T &
                                  shap_final$neg_habitat == T,

                                "5. shadow - poor C + H",

                                ifelse(shap_final$natural_niche == F,
                                       "1. outside ecological niche",
                                       NA))))))
```

We can investigate the proportion of the landscape that falls into each of the above categories:

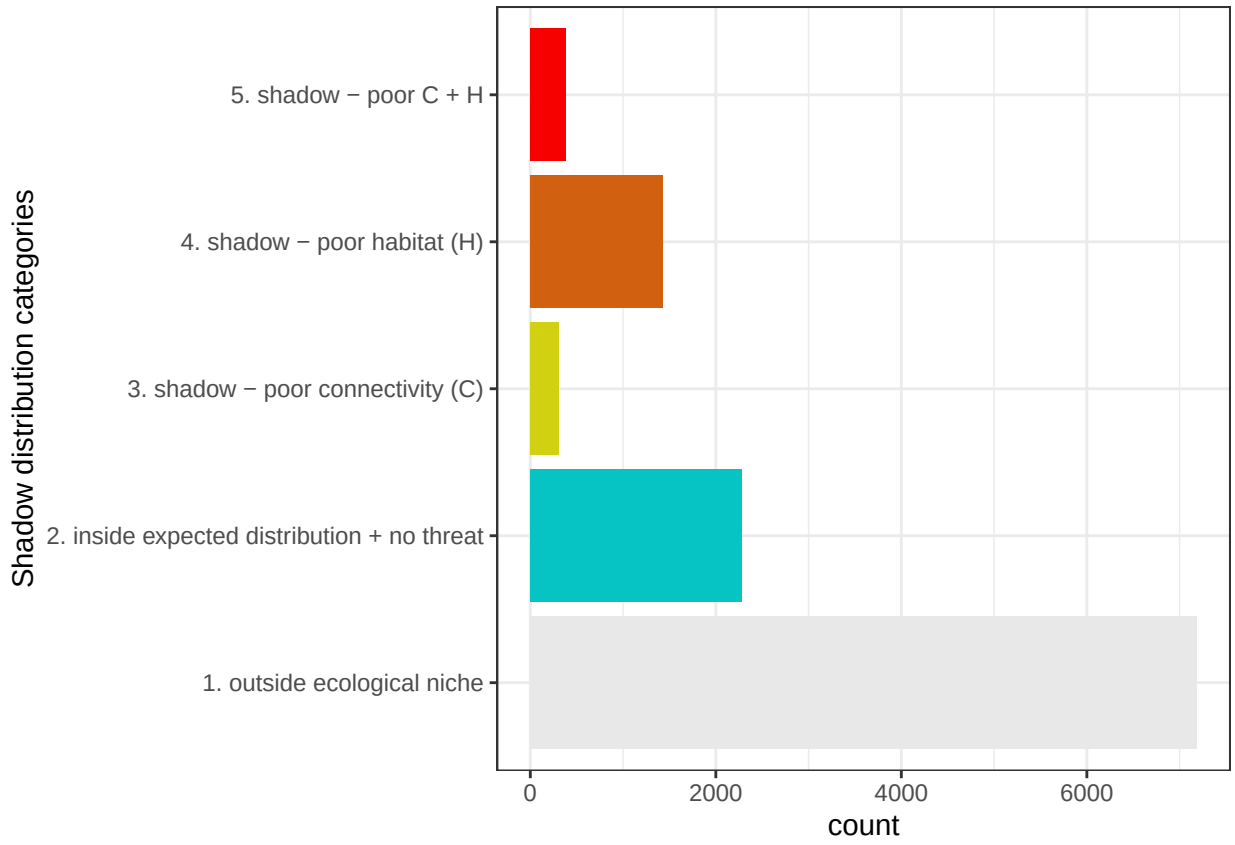

Figure 15. Relative proportion of subcatchments falling into different categories of the expected and shadow distribution

In the above plot we see relative proportions of each distribution category. Categories 1-2 represent inside or outside of the expected distribution, whereas categories 3-5 indicate the shadow distribution inside the expected distribution. We can also spatially map these categories to see the overall spatial distribution of different threats and their combinations, and also regions where threats are expected to have a limited effect on species distributions.

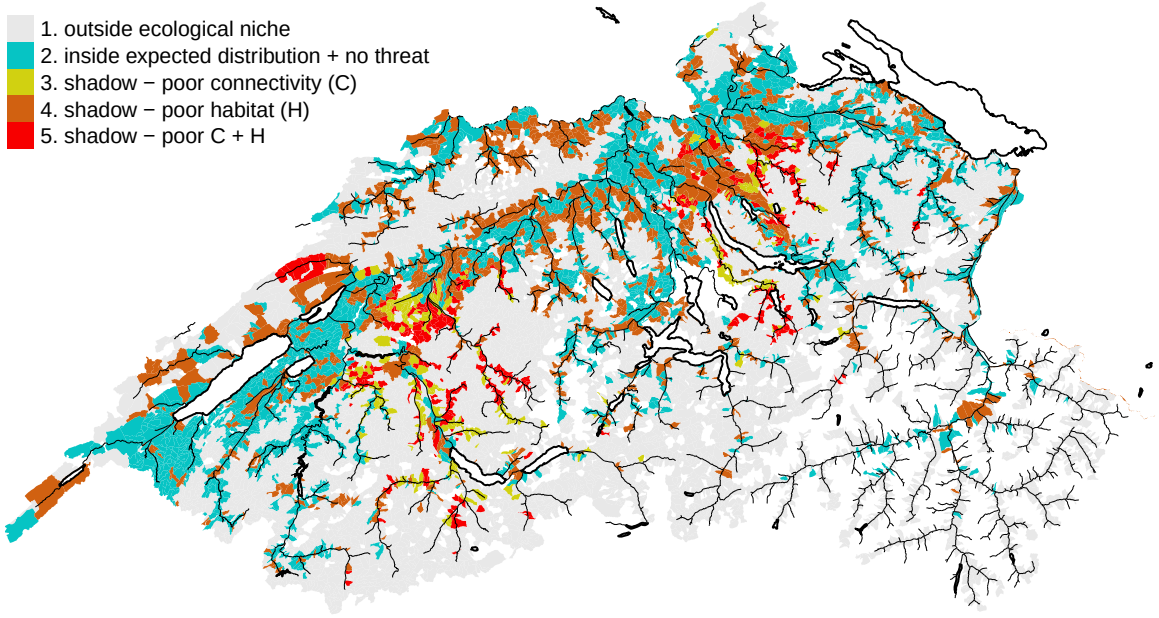

Figure 16. Map of expected and shadow distribution types across subcatchments

We interpret the above distribution as representing the shadow distribution of the species in this qualitative way. The areas coloured red, green or brown indicate the shadow distribution of the species. The areas in blue indicate the expected distribution that is not in the shadow distribution, and the areas in gray are outside of the expected distribution. In this way, we can visualise and summarise the spatial distribution of negative threat effects on species. We expect such workflows could provide important feature inputs to systematic conservation planning exercises such as features to zonation, marxan or prioritizr, especially where threats are to be alleviated as highlighted in (Salgado-Rojas, Hermoso, and Álvarez-Miranda, n.d.).

4c. Shadow distribution - quantitative While the summary of areas inside or outside of the expected distributions and shadow distribution is helpful as a broad overview of areas potentially influence by threats, it is also important to demonstrate the magnitude of reduction in environmental suitability due to threat factors. This can be achieved through summarising the **quantitative shadow distribution** of a species.

The quantitative shadow distribution estimates the loss of environmental suitability due to human threats in areas falling within the expected distribution. We estimate the quantitative shadow distribution by calculating the difference between environmental suitability scores of the expected distribution and the observed distribution. We express this as a proportional loss of environmental suitability in the observed distribution.

The observed distribution could simply be the model prediction (i.e., raw output) of the species distribution model, giving an indicator of environmental suitability. However, as a minor technicality, to maintain consistency in approaches we calculate the model prediction from the sum of all the Shapley values for all the variables in the model + a baseline prediction - which in theory gives the model prediction for an observation (Scott M. Lundberg and Lee 2017b). Deviations can occur due to the bootstrapped nature of the Shapley value but we observed these to be very minor (r^2 between both approaches ~0.99).

The following steps are necessary to calculate the quantitative shadow distribution:

1. *Definition of a baseline value*
2. *Prediction of environmental suitability*
3. *Partition Shapley value contributions to threats vs. non-threats.*
4. *Decide on strategy to estimate quantitative expected distribution*
5. *Identify qualitative expected distribution*
6. *Calculate the difference between observed distribution and expected distribution - giving the quantitative shadow distribution*

Working through the code to calculate quantitative shadow distribution

1. *Definition of a baseline value.* We must define a baseline value as the average of the predicted values. This is the reference value that a Shapley value is added to (+ or -) to determine the contribution to the local prediction of a given variable.

```
### ----
# step 1: Definition of baseline value
baseline_value <- mean(shap_final$suitability, na.rm = T)
```

2. *Prediction of environmental suitability.* Here we sum the Shapley values across all variables which calculates the prediction from the model (environmental suitability) for a given observation (sub-catchment).

```
### ----
# step 2: Prediction of environmental suitability
# we calculate environmental suitability from sum of Shapley values
shap_final$shap_all_sum <- rowSums(
  st_drop_geometry(
    shap_final[,names(shap_final) %in%
                 c(natural_niche_factors,
                   connectivity_threat,
                   habitat_threat)]
    ),
  na.rm = T)

# The environmental suitability is the shapley sum + the baseline value.
shap_final$shap_suit_baseline <- shap_final$shap_all_sum + baseline_value
# Here note that values < 0 or > 1 can occur in the Shapley
# version of environmental suitability due to the bootstrap
# nature of the Shapley values. As you approach infinity
# shapley value runs the simulated environmental suitability
# approaches the predicted environmental suitability of the
# model. We found no important changes in the shapley values
# between 1000 runs and 10,000 runs that we used in our final analysis.
```

3. *Partition Shapley value contributions to threats vs. non-threats.* We sum the Shapley values for threats and non-threats separately, to obtain the independent contributions to environmental suitability scores for different types of environmental variables.

```
### ----
# Step 3. Partition shapley value contributions to threats vs. non-threats

# Define threat factors
threat_factors <- c(connectivity_threat,
                    habitat_threat)

# Create matrix of only threat columns
threat_shaps <- st_drop_geometry(shap_final[,names(shap_final) %in%
                                              c(threat_factors)])
threat_shaps[threat_shaps<0] <- NA
# remove areas where threat level benefit species.
# This occurs in unthreatened locations where threats area already alleviated.
str(threat_shaps, 1)
```

```
## 'data.frame':    15130 obs. of  4 variables:
##  $ local_asym_cl_log10_SHAP         : num  0.1439 0.0544 NA 0.1018 0.1099 ...
##  $ ecoF_eco_mean_ele_residual_SHAP  : num  0.0179 0.04846 NA 0.01739 0.00023 ...
##  $ local_imd_log10_ele_residual_SHAP: num  0.00283 0.08812 NA 0.14517 NA ...
##  $ local_flood_SHAP                 : num  NA 0.0449 NA NA NA ...
```

```
# We calculated natural niche before
shap_final$nn_sum_mask <- ifelse(shap_final$natural_niche < 0,
                                 NA,
                                 shap_final$natural_niche)
```

4. *Decide on strategy to estimate quantitative expected distribution.* Depending on the users needs, there are different ways to compare the quantitative expected distribution to the observed distribution. These approaches vary in their assumption about the alleviation of threats to move from the observed environmental suitability (with threat effects) to the expected environmental suitability (with threat effects removed). We expect the most accurate representation of the expected suitability is to calculate expected suitability when threats are removed completely, i.e., simulating a baseline or reference state of the subcatchment. To simulate this alleviation of threats, we followed three strategies that converted Shapley values for threats to positive values:

    - Converting negative threat Shapley values to a maximum positive Shapley value for that threat (a best case-scenario). For caution, and to avoid spuriously large positive Shapley values in unusual sites, we used the 95th quantile of Shapley values per threat as our correction as our measure of the maximum.
    - Converting negative Shapley values to 0 which indicates if threats no longer have a negative contribution to environmental suitability but also do not support environmental suitability (very conservative scenario, assuming little impact of threat allevation).
    - Converting negative Shapley values to the mean positive Shapley values (conservative baseline scenario, assuming threat alleviation is inefficient).

These scenarios help quantify uncertainty the change in environmental suitability when "alleviating" threats under different assumptions.

The definition of the expected suitability inside the expected distribution is the key step in the definition of the shadow distribution. This step sums the natural niche Shapley values and the baseline prediction of the model with the **corrected** or **alleviated** threat values. In this way, we simulate the removal of the threat from the model prediction, which, if threats act negatively, then increases the predicted environmental suitability.

5. *Identify binary expected distribution.* We next remove all areas from consideration that are outside of the expected distribution (defined by positive summed natural niche Shapley values), and therefore inside the ecological niche, of the species. This step is important to ensure that only the areas where we expect a species to naturally occur are included in the calculation of the species shadow distribution.

```
### ----
# Step 4 Estimate quantitative expected distribution
# Step 5 Identify qualitative niches (through natural_niche column)

# Calculate for the threat shaps
corrected_threat_shaps <- rowSums(apply(threat_shaps, 2, function(x){
      x[is.na(x)] <- quantile(x, 0.95, na.rm = T)
      return(x)}))
```

```
# Add the alleviated effects of the threats back to the observed distributions
shap_final$expected_distribution <- shap_final$natural_niche_value +
  corrected_threat_shaps +
  baseline_value

# Mask the expected suitability by the qualitative definition
# as inside or outside of the expected distribution / natural niche of the species
shap_final$expected_distribution <- ifelse(shap_final$natural_niche_value < 0,
                                           NA,
                                           shap_final$expected_distribution)

# remove small number of shapley values which due to stochastic calculation generate
# values > 1
shap_final$expected_distribution[shap_final$expected_distribution > 1] <- 1
```

6. *Calculate the difference between observed distribution and expected distribution - giving the quantitative shadow distribution.* This estimates the environmental suitability loss due to human threats that fall within the natural niche of the species.

```
### ---
# Step 6 Calculate the quantitative shadow distribution

# We must first define the observed distribution, which we
# do as the predicted environmental suitability inside the
# natural niche of the species. Above we already . Remember,
# that this prediction has both the effects of the natural
# niche factors and the threat factors within the estimate
# of the environmental suitability.
shap_final$observed_distribution <- ifelse(is.na(shap_final$nn_sum_mask),
                                           NA,
                                           shap_final$shap_suit_baseline)

# Here, we expect the most important summary of the shadow
# distribution is the percentage loss of environmental suitability
# as a result of threats.
shap_final$SD_OratioE <- shap_final$observed_distribution /
  shap_final$expected_distribution

# However, we can also calculate the raw different between
# expected and observed distributions.
shap_final$SD_OminusE <- shap_final$observed_distribution -
  shap_final$expected_distribution

# And, the difference as a percentage of the expected distribution.
shap_final$SD_OpercentOfE <- (shap_final$SD_OminusE) /
  shap_final$expected_distribution
```

Next we can spatially map the shadow distribution of the species.

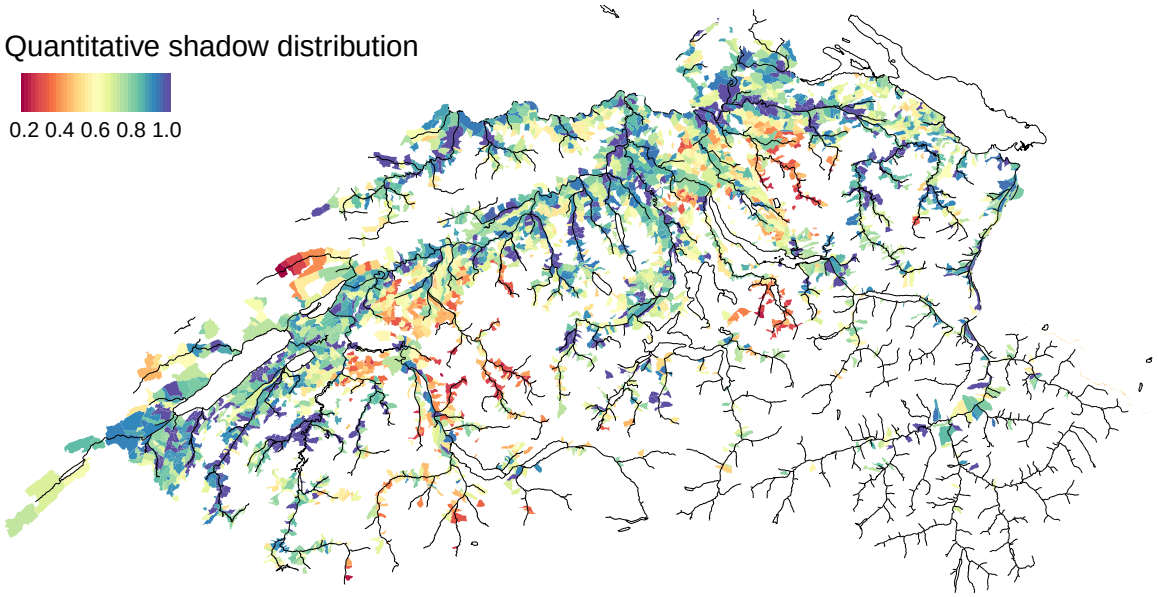

Figure 17. Species quantitative shadow distribution expressed as the ratio of the observed habitat suitability to the expected habitat suitability

```
print(sessionInfo())
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] kernelshap_0.3.8     fastshap_0.0.7       pROC_1.18.0
##  [4] Boruta_8.0.0         randomForest_4.7-1.1 sf_1.0-9
##  [7] tmap_3.3-3           terra_1.7-3          lubridate_1.9.2
## [10] forcats_1.0.0        stringr_1.5.0        dplyr_1.0.10
## [13] purrr_1.0.1          readr_2.1.4          tidyr_1.3.0
## [16] tibble_3.1.8         ggplot2_3.4.2        tidyverse_2.0.0
## [19] pacman_0.5.1
##
## loaded via a namespace (and not attached):
##   [1] leafem_0.2.0           colorspace_2.0-3       deldir_1.0-6
##   [4] class_7.3-19           mclust_6.0.0           leaflet_2.1.1
##   [7] htmlTable_2.4.1        base64enc_0.1-3        dichromat_2.0-0.1
##  [10] rstudioapi_0.14        proxy_0.4-27           farver_2.1.1
##  [13] earth_5.3.1            mvtnorm_1.1-3          fansi_1.0.3
##  [16] ranger_0.14.1          codetools_0.2-18       splines_4.1.2
##  [19] cachem_1.0.7           knitr_1.43             ade4_1.7-20
##  [22] Formula_1.2-4          jsonlite_1.8.4         mda_0.5-3
##  [25] tmaptools_3.1-1        cluster_2.1.2          png_0.1-8
##  [28] compiler_4.1.2         backports_1.4.1        assertthat_0.2.1
##  [31] Matrix_1.5-3           fastmap_1.1.0          cli_3.5.0
##  [34] htmltools_0.5.4        tools_4.1.2            gtable_0.3.3
##  [37] glue_1.6.2             reshape2_1.4.4         Rcpp_1.0.9
##  [40] PresenceAbsence_1.1.11 jquerylib_0.1.4        raster_3.6-14
##  [43] vctrs_0.6.1            ape_5.6-2              nlme_3.1-153
##  [46] iterators_1.0.14       leafsync_0.1.0         crosstalk_1.2.0
##  [49] lwgeom_0.2-10          xfun_0.39              maxnet_0.1.4
##  [52] timechange_0.1.1       lifecycle_1.0.3        gtools_3.9.4
##  [55] ecospat_3.4            XML_3.99-0.13          MASS_7.3-58.1
##  [58] scales_1.2.1           hms_1.1.3              parallel_4.1.2
##  [61] RColorBrewer_1.1-3     yaml_2.3.7             gridExtra_2.3
##  [64] TeachingDemos_2.12     sass_0.4.5             rpart_4.1-15
##  [67] latticeExtra_0.6-30    reshape_0.8.9          stringi_1.7.8
##  [70] foreach_1.5.2          plotrix_3.8-2          checkmate_2.1.0
##  [73] permute_0.9-7          e1071_1.7-12           poibin_1.5
```

```
##  [76] rlang_1.1.1           pkgconfig_2.0.3       pracma_2.4.2
##  [79] evaluate_0.21         lattice_0.20-45       labeling_0.4.2
##  [82] ks_1.14.0             htmlwidgets_1.6.1     tidyselect_1.2.0
##  [85] gbm_2.1.8.1           biomod2_4.2-1         plyr_1.8.8
##  [88] magrittr_2.0.3        R6_2.5.1              Hmisc_4.7-2
##  [91] generics_0.1.3        DBI_1.1.3             foreign_0.8-81
##  [94] mgcv_1.9-0            pillar_1.9.0          withr_2.5.0
##  [97] units_0.8-1           stars_0.6-0           survival_3.2-13
## [100] abind_1.4-5           sp_1.5-1              nnet_7.3-16
## [103] interp_1.1-3          KernSmooth_2.23-20    utf8_1.2.2
## [106] tzdb_0.3.0            rmarkdown_2.23        nabor_0.5.0
## [109] jpeg_0.1-10           grid_4.1.2            data.table_1.14.6
## [112] vegan_2.6-4           plotmo_3.6.2          digest_0.6.31
## [115] classInt_0.4-8        munsell_0.5.0         viridisLite_0.4.2
## [118] bslib_0.5.0
```

Citations

Araújo, Miguel B., Robert P. Anderson, A. Márcia Barbosa, Colin M. Beale, Carsten F. Dormann, Regan Early, Raquel A. Garcia, et al. 2019. "Standards for Distribution Models in Biodiversity Assessments." *Science Advances* 5 (1): eaat4858. https://doi.org/10.1126/sciadv.aat4858.

Greenwell, Brandon. 2021. *Fastshap: Fast Approximate Shapley Values.* https://CRAN.R-project.org/package=fastshap.

Guillera-Arroita, Gurutzeta, José J. Lahoz-Monfort, Jane Elith, Ascelin Gordon, Heini Kujala, Pia E. Lentini, Michael A. McCarthy, Reid Tingley, and Brendan A. Wintle. 2015. "Is My Species Distribution Model Fit for Purpose? Matching Data and Models to Applications." *Global Ecology and Biogeography* 24 (3): 276–92. https://doi.org/10.1111/geb.12268.

Kursa, Miron B., and Witold R. Rudnicki. 2010. "Feature Selection with the Boruta Package." *Journal of Statistical Software* 36 (September): 1–13. https://doi.org/10.18637/jss.v036.i11.

Lucas, Tim C. D. 2020. "A Translucent Box: Interpretable Machine Learning in Ecology." *Ecological Monographs* 90 (4): e01422. https://doi.org/10.1002/ecm.1422.

Lundberg, Scott M, and Su-In Lee. 2017b. "A Unified Approach to Interpreting Model Predictions." In. Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

———. 2017a. "A Unified Approach to Interpreting Model Predictions." In. Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Lundberg, Scott M., Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, et al. 2018. "Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia During Surgery." *Nature Biomedical Engineering* 2 (10): 749–60. https://doi.org/10.1038/s41551-018-0304-0.

Ryo, Masahiro, Boyan Angelov, Stefano Mammola, Jamie M. Kass, Blas M. Benito, and Florian Hartig. 2021. "Explainable Artificial Intelligence Enhances the Ecological Interpretability of Black-Box Species Distribution Models." *Ecography* 44 (2): 199–205. https://doi.org/10.1111/ecog.05360.

Salgado-Rojas, José, Virgilio Hermoso, and Eduardo Álvarez-Miranda. n.d. "Prioriactions: Multi-Action Management Planning in R." *Methods in Ecology and Evolution* n/a (n/a). https://doi.org/10.1111/2041-210X.14220.

Sillero, Neftalí, and A. Márcia Barbosa. 2021. "Common Mistakes in Ecological Niche Models." *International Journal of Geographical Information Science* 35 (2): 213–26. https://doi.org/10.1080/13658816.2020.1798968.

Štrumbelj, Erik, and Igor Kononenko. 2014b. "Explaining Prediction Models and Individual Predictions with Feature Contributions." *Knowledge and Information Systems* 41 (3): 647–65. https://doi.org/10.1007/s10115-013-0679-x.

———. 2014a. "Explaining Prediction Models and Individual Predictions with Feature Contributions." *Knowledge and Information Systems* 41 (3): 647–65. https://doi.org/10.1007/s10115-013-0679-x.

Thuiller, Wilfried, Damien Georges, Maya Gueguen, Robin Engler, Frank Breiner, Bruno Lafourcade, and Remi Patin. 2023. *Biomod2: Ensemble Platform for Species Distribution Modeling.*

Valavi, Roozbeh, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. 2021. "Modelling Species Presence-Only Data with Random Forests." *Ecography* 44 (12): 1731–42. https://doi.org/10.1111/ecog.05615.

Wadoux, Alexandre M. J-C., and Christoph Molnar. 2022. "Beyond Prediction: Methods for Interpreting Complex Models of Soil Variation." *Geoderma* 422 (September): 115953. https://doi.org/10.1016/j.geoderma.2022.115953.