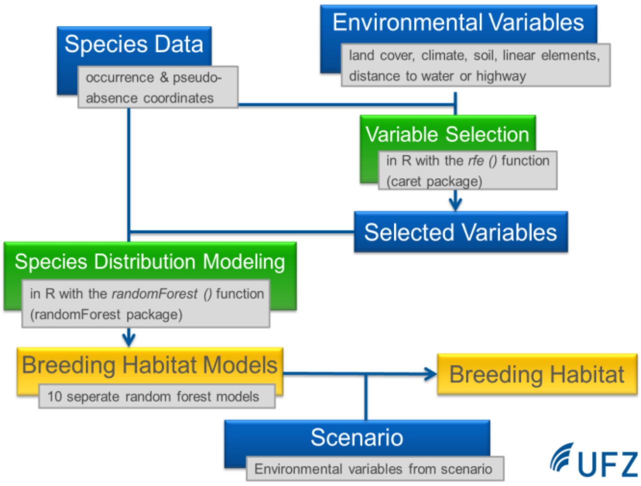
Species distribution models (random forest) based on environmental variables

1. General information	
Methodology name	**Species distribution models (random forest) based on environmental variables**
Short description	Environmental predictors (e.g. land cover, climate and soil properties, the occurrence of linear landscape elements and the distance parameters) are used to model the habitat of a species. A feature selection process prior to creating the model is described to identify the relevant environmental predictors for the species.
Created by	Dr. Anne Jungandreas, Helmholtz Centre for Environmental Research (Leipzig, Germany)
Flowchart	 Species Data occurrence & pseudo-absence coordinates Environmental Variables land cover, climate, soil, linear elements, distance to water or highway Variable Selection in R with the *rfe ()* function (caret package) Selected Variables Species Distribution Modeling in R with the *randomForest ()* function (randomForest package) Breeding Habitat Models 10 seperate random forest models Breeding Habitat Scenario Environmental variables from scenario UFZ

2. Data collection	
2a. Literature review	yes
Guidance for literature review	The habitat of a species is defined by certain environmental boundaries. A literature review can help with the selection of fitting environmental variables (land use, climate and soil conditions, etc.). Care should be taken that different spatial resolutions might result in a different importance of the variables.
2b. Quantitative data	yes
Dataset	Species occurrence data and (pseudo-)absence data with specific coordinates.
2c. Spatial data	yes
Dataset	Raster data on all environmental variables used in the models.
2d. Stakeholder input	no

3. Methodology	
Description	General steps in creating a species distribution model with random forest with preceding selection of environmental variables.
Stepwise description	
Step 1	**Prepare R** Install the packages ‘dismo’, ‘caret’ and ‘randomForest’ in R.
Step 2	**Collect the data** Collect occurrence and (pseudo-)absence data for the species. Collect data on environmental variables (e.g. temperature, amount of agricultural area, etc.) that might be useful for creating the species distribution model in a raster format. The variables should be related to environmental conditions known to have an impact on the species (i.e. do a literature review). Extract the environmental variables from the occurrence and (pseudo-)absence coordinates into tables. Give a value of 0 to (pseudo-)absences and of 1 to occurrences.
Step 3	**Variable selection** If you are not sure that all collected environmental variables are indeed important for defining the habitat of the species, a variable selection is recommended. Create a dataset for the variable selection (= feature selection) from the data. Care should be taken that the ratio between occurrence and (pseudo-) absence data does not exceed 1:10 (Barbet-Massin et al., 2012). If the balance between occurrences and (pseudo-)absences is not given, the resulting models will focus on the much bigger data set and the prediction quality will be poor despite a possible good model quality. Then use the function *rfe ()* (caret package) to do a feature selection. The function *rfeControl ()* can be used to detail the feature selection run. As a result, you will get a file with information on variable sets with a different number of variables and the pseudo-R2 and RMSEs of resulting random forest models (see the ***Example*** on more in-depth information).
Step 4	**Species distribution modelling with random forest** The method orients itself on the thorough and recommendable vignette about species distribution modelling by Hijmans & Elith (2017). Create a training dataset for the feature selection (the selection of variables) from the data. Hold back a part of the data for model testing (we usually hold back 1/6th of the data). We also use a different set of randomly chosen (pseudo-) absence data for feature selection and for modeling. Then use the function *randomForest ()* (randomForest package) to create a model (see the ***Example*** on more in-depth information). The *predict ()* function in R can be used to predict the occurrence of the species between 0 and 1. If you want to map the probable habitat of a species you need to calculate a threshold for presence/absence prediction. Therefore, create a testing dataset from the data held back. Then use *evaluate ()* and *threshold ()* (both ‘dismo’ package) to calculate the threshold of your model. (see the ***Example*** on a coding example).

4. Example	
Description	This is a step-by-step example on how species distribution models with random forest were created in the German case study of the TALE project with a preceding selection of environmental variables. The file *“SDM_with_randomForest.zip”* contains two R-scripts that contain code to run steps 3 and 4 and the folder *“source”* containing all example data. Additionally, the document *“Example_description.docx”* describes all steps and the corresponding R-code thoroughly.
Data access	The species data and part of the environmental data used in the German case study is restricted. Therefore, only example data sets without coordinates and processed data is provided for the present example.
Step 1	**Prepare R** Install the packages ‘dismo’, ‘caret’ and ‘randomForest’ in R.
Step 2	**Collect the data** Example data with extracted environmental information is provided in the folder “source” and is loaded at the appropriate positions in the R-scripts of step 3 and 4.
Step 3	**Variable selection** The selection of environmental variables is done with a recursive feature selection in R. You will find the code in the file “Variable_Selection.R”.
Step 4	**Species distribution modelling with random forest** The species distribution modelling is done with random forest in R. You will find the code in the file “SDM_modelling.R”.

5. References	
1.	Barbet-Massin, M, Jiguet, F, Albert, CH, Thuiller, W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* 3: 327-338
2.	Hijmans, RJ, Elith, J (2017) Species distribution modeling in R. URL: https://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf