Introduction to species distribution modelling (SDM) in R

Damaris Zurell

https://damariszurell.github.io (https://damariszurell.github.io)
Ecology & Macroecology Group
Inst. of Biochemistry and Biology
University of Potsdam

Last updated November 2020

Software notes:

I prepared all course materials in R version 4.0.2 (R Core Team 2020). However, the codes should als run in older versions of R >3.6.0. You need to install the following packages and their dependencies:

- ***data.table*** (Dowle and Srinivasan 2019)
- ***raster*** (Hijmans 2019)
- ***randomForest*** (Liaw and Wiener 2002)
- ***lattice*** (Sarkar 2008)
- ***RColorBrewer*** (Neuwirth 2014)
- ***PresenceAbsence*** (Freeman and Moisen 2008)

1 Background

Here, I provide a short, half-day introduction to species distribution modelling in R. The course gives a brief overview of the concept of species distribution modelling, and introduces the main modelling steps. Codes and data largely follow the materials from Zurell and Engler (2019) although we will use a different case study.

Species distribution models (SDMs) are a popular tool in quantitative ecology (Franklin 2010; Peterson et al. 2011; Guisan, Thuiller, and Zimmermann 2017) and constitute the most widely used modelling framework in global change impact assessments for projecting potential future range shifts of species (IPBES 2016). There are several reasons that make them so popular: they are

comparably easy to use because many software packages (e.g. Thuiller et al. 2009; Phillips, Anderson, and Schapire 2006) and guidelines (e.g. Elith, Leathwick, and Hastie 2008; Elith et al. 2011; Merow, Smith, and Silander Jr 2013; Guisan, Thuiller, and Zimmermann 2017) are available, and they have comparably low data requirements.

As input, SDMs require georeferenced biodiversity observations (e.g. individual locations, species' presence, species' counts, species richness; the response or dependent variable) and geographic layers of environmental information (e.g. climate, land cover, soil attributes; the predictor or independent variables). Such information are now widely available in digital format. For example, online repositories provide data on species distributions (e.g. GBIF (www.gbif.org) and OBIS (https://obis.org/)), on individual animal locations (e.g. Movebank (www.movebank.org)), on climate (e.g. WorldClim (http://www.worldclim.org/) and CHELSA (http://chelsa-climate.org/)) as well as land cover and other remote sensing products (e.g. Copernicus (https://land.copernicus.eu/global/products)). We can then relate the biodiversity observations at specific sites to the prevailing environmental conditions at those sites. Different statistical and machine-learning algorithms are available for this. Once we have estimated this biodiversity-environment relationship, we can make predictions in space and in time by projecting the model onto available environmental layers (Figure 1.1).

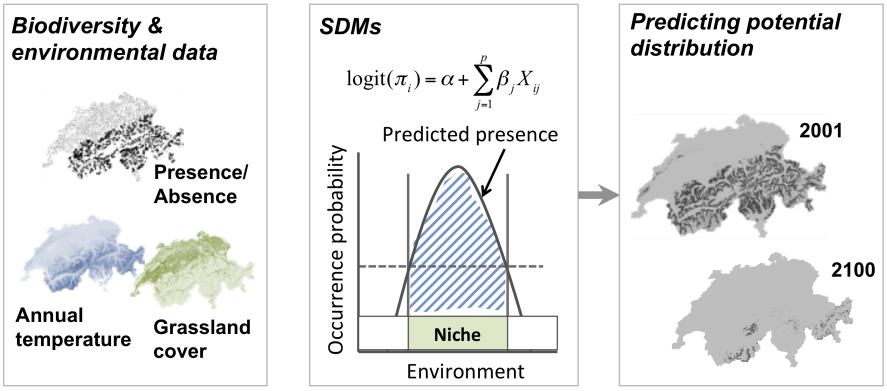


Figure 1.1: *Schematic representation of the species distribution modelling concept. First, biodiversity and environmental information are sampled in geographic space. Second, a statistical model (here, generalised linear model) is used to estimate the species-environment relationship. Third, the species–environment relationship can be mapped onto geographic layers of environmental information to delineate the potential distribution of the species. Mapping to the sampling area and period is usually referred to as interpolation, while transferring to a different time period or geographic area is referred to as extrapolation.*

1.1 Theory and Concept

Central to understanding species distributions is the niche theory that goes back to ideas formulated by Joseph Grinell and G. Evelyn Hutchinson (see Soberon 2007). Hutchinson distinguished the fundamental and the realised niche of a species (Hutchinson 1957). Thereby, the fundamental niche comprises all abiotic environmental conditions where a species can survive indefinitely, meaning where it has a positive population growth. We can envision this as an n-dimensional hypervolume where the axes are different environmental factors. Hutchinson defined the realised niche of the species as those parts of the fundamental niche where a species can survive despite the presence of competitors. Thus, Hutchinson assumed that the realised niche was smaller than the fundamental niche due to negative interspecific interactions. Today we know that facilitation (positive interspecific interactions) can also have a large effect on species distributions (Bruno, Stachowicz, and Bertness 2003). Dispersal ability or movement capacity of a species is a third important factor determining

whether a species is present in all suitable habitats (Soberon 2007). For example, sometimes species may be found in unsuitable sink habitats where they do not have positive population growth but which they can easily reach from source habitats (Pulliam 1988). Also, populations may go locally extinct simply in response to stochasticity and the dispersal ability will determine how fast the now empty patch can be recolonised (Hanski 1998). So-called BAM (biotic-abiotic-movement) diagrams emphasise the complex interplay between these three factors (Figure 1.2)(Soberon 2007; Peterson et al. 2011).

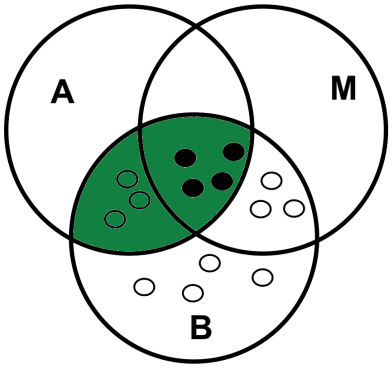


Figure 1.2: *The BAM diagram representing the factors that limit species distributions (Soberon 2007). A species can only survive in geographic areas where both the abiotic environmental conditions (A) and the prevailing biotic interactions (B) allow positive population growth. The intersection of A and B thus represents the potential distribution of the species, or the realised niche (the green elipse). The movement capacity (M) of a species will determine which geographic area is accessible within the time period of interest. The intersection of A, B and M represents the geographic area that is actually occupied by the species and where we can find source populations (filled circles). Open circles represent sink population with negative population growth.*

Although SDMs are clearly positioned within the niche theory, it is an ongoing debate whether the estimated species-environment relationship (Figure 1.1) approximates the fundamental niche (area A in Figure 1.2), the realised niche (green area in Figure 1.2) or the occupied niche (intersection of A, B and M in Figure 1.2). This debate is manifested in the multitude of names that are available for distribution modelling, for example climate envelopes, habitat model, resource selection functions, species distribution model (Elith and Leathwick 2009; Zurell and Engler 2019). It is important to understand that different factors will determine what part of the niche is reflected by the fitted species-environment relationship. Specifically, ecological processes are highly scale dependent, biodiversity data and environmental data can be gathered in very different ways and because of that may carry different biases, and also the different algorithms available for fitting SDMs (Elith et al. 2006) and the chosen model complexity (Merow et al. 2014) will largely affect the species-environment relationship. Thus, a careful reflection of the underlying assumptions and modelling choices in SDM studies is vital for understanding and applying these models.

2 SDM step by step

We can distinguish five main modelling steps for SDMs: (i) conceptualisation, (ii) data preparation, (iii) model fitting, (iv) model assessment, and (v) prediction (Figure 2.1). Even if the SDM siblings such as climate envelopes, habitat models, and resource selection functions, put slightly different emphasis on different aspects of the niche and are typically used at largely different spatial scales, the modelling steps are essentially the same. Below, I will illustrate these modelling steps with a **case study on the Ring Ouzel (*Turdus torquatus*) in Switzerland**. Generally, different data will require different treatment in SDM studies. Also, the model objective will affect modelling decisions. We can

distinguish three main objectives for SDMs: (a) inference and explanation, (b) mapping and interpiolation, and (c) forecast and transfer. I try to point out critical decision points but cannot be excaustive here. Rather, I would recommend getting more familiar with critical assumptions and modelling decisions by studying the many excellent review articles (Guisan and Zimmermann 2000; Guisan and Thuiller 2005; Elith and Leathwick 2009) and textbooks on SDMs (Peterson et al. 2011; Franklin 2010; Guisan, Thuiller, and Zimmermann 2017).

I would also like to emphasise that model building is an iterative process and there is much to learn on the way. In consequence, you may want to revisit and improve certain modelling steps, for example improve the spatial sampling design. Because of that I like to regard model building as a cycle rather than a workflow with a pre-defined termination point (Figure 2.1).

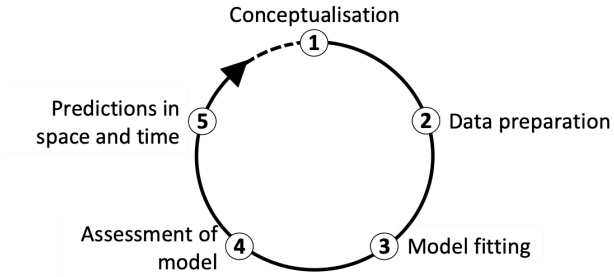


Figure 2.1: *The main modelling cycle in species distribution modelling.*

2.1 Conceptualisation

In the conceptualisation phase, we formulate our main research objectives and decide on the model and study setup based on any previous knowledge on the species and study system. An important point here is whether we can use available data or have to gather own biodiversity (and environmental) data, which would require deciding on an appropriate sampling design. Then, we carefully check the main underlying assumptions of SDMs, for example whether the species is in pseudo-equilibrium with environment and whether the data could be biased in any way (cf. chapter 5 in Guisan, Thuiller, and Zimmermann 2017). The choice of adequate environmental predictors, of modelling algorithms and of desired model complexity should be guided by the research objective and by hypotheses regarding the species-environment relationship. We can divide environmental variables into three types of predictors: resource variables, direct variables and indirect variables (Austin 1980; Guisan and Zimmermann 2000).

2.1.1 Ring Ouzel: Conceptualisation

We aim at assessing potential climate change effects on the Ring Ouzel (*Turdus torquatus*) in Switzerland (Figure 2.2), a typical mountain bird in the Alps and Jura mountains. The Swiss Ornithological Institute (https://www.vogelwarte.ch/en) provides comprehensive information for all birds breeding in Switzerland. Thus, we first have a look at their website to find out more about Ring Ouzel ecology (https://www.vogelwarte.ch/en/birds/birds-of-switzerland/ring-ouzel).

Figure 2.2: *A male ring ouzel. Copyright Ruedi Aeschlimann. Downloaded from www.vogelwarte.ch.*

According to the Swiss Ornithological Institute, the Ring Ouzel occurs mainly in the boreal and Alpine zones, it inhabits edge habitats mainly forests dominated by spruce, fir or larch, and adjacent meadows and pastures. It nests in trees. It is sensitive to climate warming, land use intensification on meadows and shrub encroachment in alpine pastures. Since the 1990s, density has decreased in Jura mountains and Pre-Alps while populations above 2000 m elevation have been comparably stable. In neighbouring countries, populations have also been assumed to be stable.

The Swiss Ornithological Institute has produced two distribution atlasses at 1 km resolution over the last two decades. These data, in particular the breeding bird atlas for 1993-1996 (Schmid et al. 1998) have been used previously. For example, Zurell et al. (2020) have made available the 1993-1996 distribution data of several breeding bird species along with environmental predictor variables, which we can use for our study. The study by Zurell et al. (2020) considered bioclimatic, topographic and remote-sensing based vegetation variables as relevant environmental predictors. For the Ring Ouzel, they identified the following five variables as the most important predictors: bio5 = maximum temperature of the warmest month, bio2 = mean diurnal range, bio14 = precipitation of driest month, std = standard deviation of vegetation height, and rad = annual total radiation.

Here, we will concentrate on the climate predictors in order to compare potential distribution under current and future climates. Also, the three climate variables had been identified as the most important predictors for Ring Ouzel by Zurell et al. (2020). In order to understand differences in model and extrapolation behaviour in different statistical algorithms, we will fit simple generalised linear models (GLMs) and random forests (RFs) (for a list of typically available algorithms see e.g. Elith et al. 2006). As output we want to generate maps of current and future occurrence probability as well as binary maps of potential occurrence.

Questions:

- Do you think the Ring Ouzel is/was at equilibrium during the time period when the atlas data were gathered?
- Do you think the five environmental predictors are appropriate? Should we use all of them for the climate impact assessment? What could be pros and cons?
- What relationship would you expect between Ring Ouzel occurrence and the five environmental predictors?
- Which of the proposed environmental variables constitute resource, direct and indirect predictors?
- Could the atlas data be biased?

2.2 Data preparation

In this step, the actual biodiversity and environmental data are gathered and processed. This concerns all data that are required for model fitting but also data that are used for making transfers. Special attention should be put on any scaling mismatches, meaning cases where the spatial (or temporal) grain or extent doe not match between biodiversity and environmental data or within environmental data. In these cases, we need to make decisions about adequate upscaling and downscaling strategies. Another important issue is the form of absence information available for the biodiversity data. Most SDM applications deal with some form of presence information for the species. These can be direct observations, GPS locations of data loggers, or museum records among others. All SDM algorithms require some form of absence or background information that they use as contrast to the presence data. Yet, absence data are rarely available. In such cases, adequate background data or pseudo-absence data needs to be selected. Again, the best strategy will depend on the research question, the data and the SDM algorithm (Guisan, Thuiller, and Zimmermann 2017). Finally, for later model assessment we may wish to partition the data into training data and validation data (Hastie, Tibshirani, and Friedman 2009).

2.2.1 Ring Ouzel: Data preparation

So, let's finally get started in R. As mentioned previously, the biodiversity and environmental data are available for download from the supplementary material of Zurell et al. (2020), or more specifically from a Dryad repository.

Download the data (https://doi.org/10.5061/dryad.k88v330) and unzip it into your data folder. Then, read in the data:

```
avi_dat <- read.table('data/Data_SwissBreedingBirds.csv', header=T, sep=',')

summary(avi_dat)
```

The data frame contains 2535 records with presence-absence information for 56 bird species, and 52 environmental predictor variables. The Ring Ouzel has a prevalence of ca. 0.25. Part of the data (70%) were used for single species distribution modelling (Zurell et al. 2020) and were already partitioned into spatial blocks in preparation for spatial block cross-validation (Roberts et al. 2017). The remaining 30% of the data were used for testing community level predictions, which is irrelevant for our purpose. Because the data were published without spatial coordinate information, we will not be able to re-partition the data into spatial blocks but could either employ a random split-sample approach or random cross-validation, or we could use the original spatial blocks (and the smaller data set) for validating our models.

Let's reduce the data frame to the relevant columns:

```
avi_cols <- c('Turdus_torquatus', 'bio_5', 'bio_2', 'bio_14', 'std', 'rad',
'blockCV_tile')

avi_df <- data.frame(avi_dat)[avi_cols]

summary(avi_df)
```

```
##  Turdus_torquatus     bio_5           bio_2            bio_14
##  Min.   :0.0000   Min.   :12.23   Min.   : 5.592   Min.   : 24.00
##  1st Qu.:0.0000   1st Qu.:18.81   1st Qu.: 7.789   1st Qu.: 64.00
##  Median :0.0000   Median :22.39   Median : 8.293   Median : 75.00
##  Mean   :0.2548   Mean   :21.42   Mean   : 8.243   Mean   : 80.97
##  3rd Qu.:1.0000   3rd Qu.:24.29   3rd Qu.: 8.689   3rd Qu.: 95.50
##  Max.   :1.0000   Max.   :28.27   Max.   :11.471   Max.   :177.00
##
##       std              rad         blockCV_tile
##  Min.   : 0.300   Min.   : 7942   Min.   :1.000
##  1st Qu.: 6.265   1st Qu.:20624   1st Qu.:2.000
##  Median : 7.930   Median :21530   Median :4.000
##  Mean   : 7.682   Mean   :21556   Mean   :3.289
##  3rd Qu.: 9.310   3rd Qu.:22704   3rd Qu.:4.000
##  Max.   :16.010   Max.   :27881   Max.   :5.000
##                                   NA's   :761
```

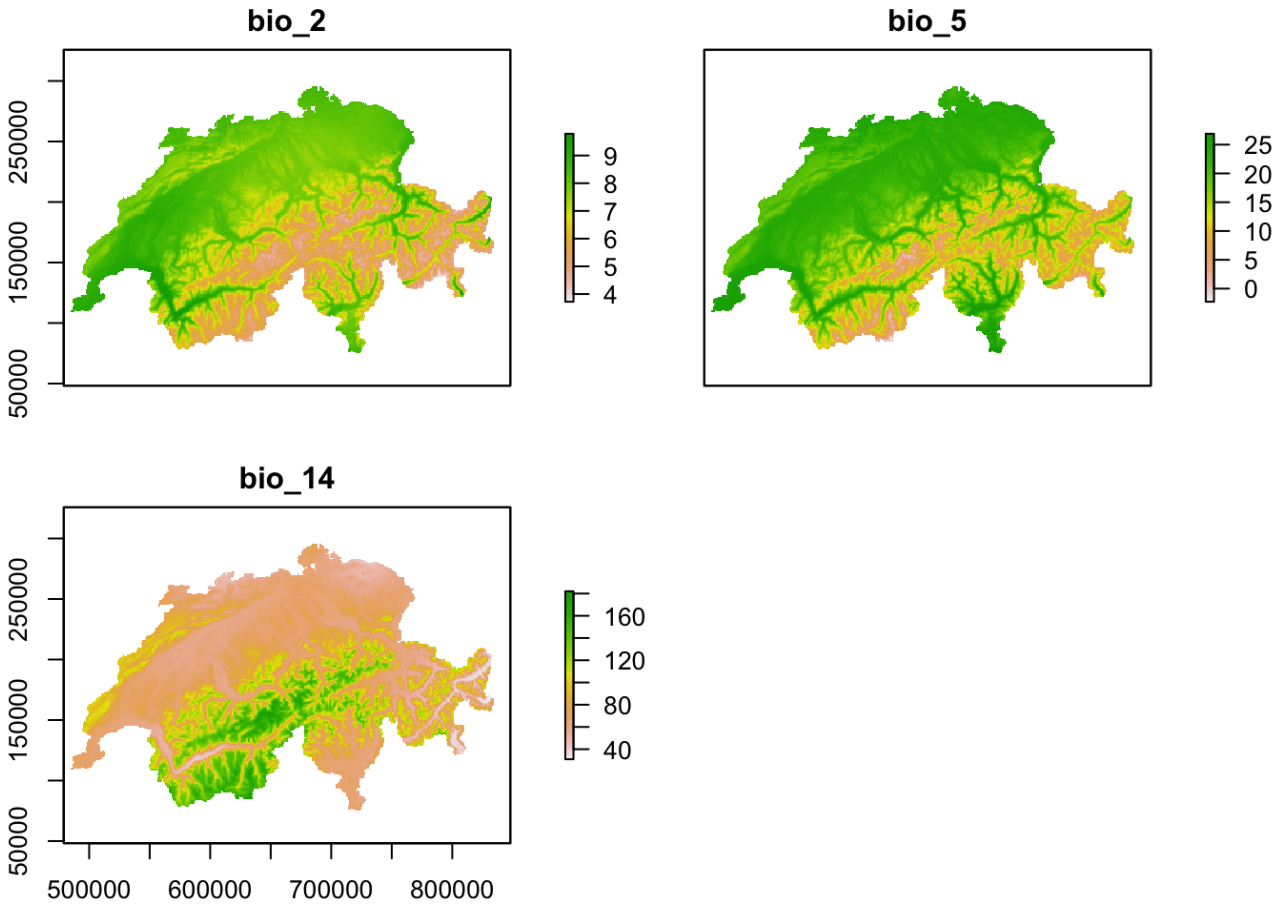
In order to map the potential distribution to the current environment and project it to future climate, we will also prepare the geographic layers for this. As mentioned earlier, the bioclimatic variables are available from WorldClim (http://www.worldclim.org/) and CHELSA (http://chelsa-climate.org/). For simplicity, we will here use the WorldClim data as we can download them directly from within R. We will use a background mask of Switzerland to clip the data.

```
library(raster)

# Please note that you have to set download=T if you haven't downloaded the d
ata before:
bio_curr <- getData('worldclim', var='bio', res=0.5, lon=5.5, lat=45.5, path=
'data')[[c(2,5,14)]]

# Please note that you have to set download=T if you haven't downloaded the d
ata before:
bio_fut <- getData('CMIP5', var='bio', res=0.5, lon=5.5, lat=45.5, rcp=45, mo
del='NO', year=50, path='data', download=F)[[c(2,5,14)]]
```

We will use a background mask of Switzerland to clip the data. This mask is in Swiss coordinates, which is the target coordinate system, and we thus need to reproject the worldclim layers. To speed things up, we will first crop the climate layers.

```
# A spatial mask of Switzerland in Swiss coordinates
bg <- raster('/vsicurl/https://damariszurell.github.io/SDM-Intro/CH_mask.tif'
)

# The spatial extent of Switzerland in Lon/Lat coordinates is roughly:
ch_ext <- c(5, 11, 45, 48)

# Crop the climate layers to the extent of Switzerland
bio_curr <- crop(bio_curr, ch_ext)

# Re-project to Swiss coordinate system and clip to Swiss political bounday
bio_curr <- projectRaster(bio_curr, bg)
bio_curr <- resample(bio_curr, bg)
bio_curr <- mask(bio_curr, bg)
names(bio_curr) <- c('bio_2', 'bio_5', 'bio_14')

# For storage reasons the temperature values in worldclim are multiplied by 1
0. For easier interpretability, we change it back to °C.
bio_curr[[1]] <- bio_curr[[1]]/10
bio_curr[[2]] <- bio_curr[[2]]/10

# Repeat above steps for future climate layers
bio_fut <- crop(bio_fut, ch_ext)
bio_fut <- projectRaster(bio_fut, bg)
bio_fut <- resample(bio_fut, bg)
bio_fut <- mask(bio_fut, bg)
names(bio_fut) <- c('bio_2', 'bio_5', 'bio_14')
bio_fut[[1]] <- bio_fut[[1]]/10
bio_fut[[2]] <- bio_fut[[2]]/10
```

```
plot(bio_curr)
```



Questions:

- What other predictor variables might be important for the Ring Ouzel that we haven't considered here?
- What kind of species data have you worked with so far or do you know?
- How would you design your own sampling?
- What kind of spatial and temporal scale mismatches could occur between biodiversity and environmental data?
- How would you sample background data for GPS tracking data or for citizen science data?
- Why do we standardise the environmental data?
- How would you assess multicollinearity in the environmental data?

2.3 Model fitting

Model fitting is the heart of any SDM application. Many different algorithms are available (Elith et al. 2006), and often several algorithms are combined into ensemble models or several candidate models with different candidate predictor sets are averaged (Hastie, Tibshirani, and Friedman 2009). The decisions on these matters should have been made during the conceptualisation phase. Important aspects to consider during the model fitting step are: how to deal with multicollinearity in the environmental data? How many variables should be included in the model (without overfitting) and how should we select these? Which model settings should be used? When multiple model algorithms or candidate models are fitted, how to select the final model or average the models? Do we need to test or correct for non-independence in the data (spatial or temporal autocorrelation,

nested data)? If the goal is to derive binary predictions, which threshold should be used? More detailed descriptions on these aspects can be found in Franklin (2010) and in Guisan, Thuiller, and Zimmermann (2017).

Exploration of model behaviour is strictly part of the model assessment step, e.g. checking the plausibility of the fitted species-environment relationship by visual inspection of response curves, and by assessing model coefficients and variable importance. However, for simplicity, we simultaneously look at model fitting and visualise model behaviour here.

2.3.1 Ring Ouzel: Model fitting

We decided to fit a generalised linear model (GLM) as a typical parametric regression method, and random forest (RF) as a machine-learning method. Specifically, we will fit quadratic functions in GLMs to test for unimodal relationships and will fit rather complex RFs. Then, we explore model coefficients in GLMs and variable importance in RFs, and plot response curves and response surfaces.

We leave out a number of aspects: The Pearson correlation coefficient between the three climate variables is $|r| < 0.7$ so we do not worry about multicollinearity issues (Dormann et al. 2013). As we do not have the spatial coordinates, we cannot test for spatial autocorrelation in the residuals. Temporal autocorrelation and nestedness of data are irrelevant for our dataset.

2.3.1.1 Generalised linear model (GLM)

Let's start with fitting a simple GLM. We can fit linear, quadratic or higher polynomial terms (check `poly()`) and interactions between predictors:
- the term `I()` indicates that a variable should be transformed before being used as predictor in the formula
- `poly(x,n)` creates a polynomial of degree n: $x + x^2 + ... + x^n$
- `x1:x2` creates a two-way interaction term between variables x1 and x2, the linear terms of x1 and x2 would have to be specified separately
- `x1*x2` creates a two-way interaction term between variables x1 and x2 plus their linear terms
- `x1*x2*x3` creates the linear terms of the three variables, all possible two-way interactions between these variables and the three-way interaction

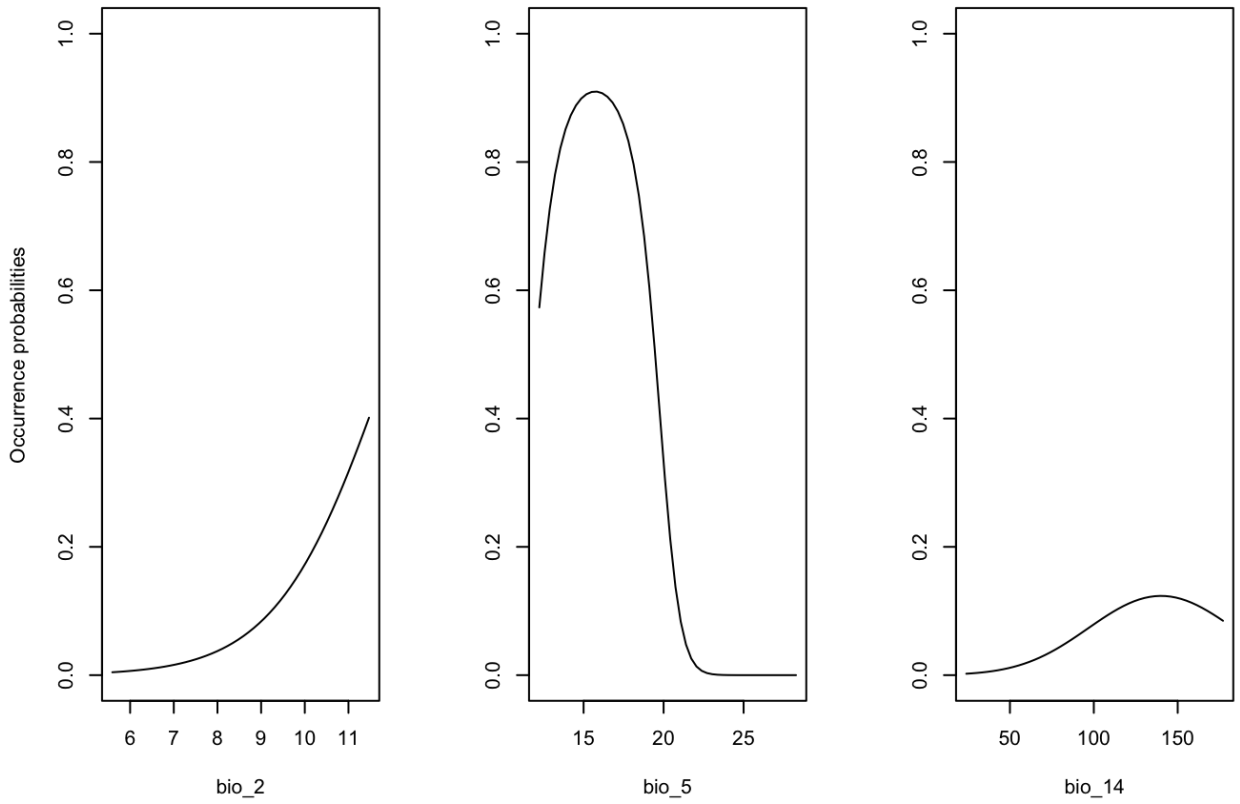
```
# Fit GLM
m_glm <- glm( Turdus_torquatus ~ bio_2 + I(bio_2^2) + bio_5 + I(bio_5^2) + bi
o_14 + I(bio_14^2), family='binomial', data=avi_df)

summary(m_glm)
```

```
##
## Call:
## glm(formula = Turdus_torquatus ~ bio_2 + I(bio_2^2) + bio_5 +
##     I(bio_5^2) + bio_14 + I(bio_14^2), family = "binomial", data = avi_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4440  -0.1061  -0.0090   0.3553   3.5462
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.139e+01  7.232e+00  -7.106 1.20e-12 ***
## bio_2        1.064e+00  1.648e+00   0.645  0.51862
## I(bio_2^2)  -1.255e-02  1.021e-01  -0.123  0.90219
## bio_5        5.191e+00  5.304e-01   9.786  < 2e-16 ***
## I(bio_5^2)  -1.650e-01  1.496e-02 -11.032  < 2e-16 ***
## bio_14       8.625e-02  1.830e-02   4.714 2.42e-06 ***
## I(bio_14^2) -3.079e-04  9.349e-05  -3.293  0.00099 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2877.6  on 2534  degrees of freedom
## Residual deviance: 1101.6  on 2528  degrees of freedom
## AIC: 1115.6
##
## Number of Fisher Scoring iterations: 8
```

Now, we plot partial response curves, meaning that for each predictor we plot the fitted species-environment relationship along the entire gradient while keeping the other predictors at their mean value. I have included a function in a little R package `mecofun` that our working group uses for teaching purposes.

```
# Install the mecofun package
library(devtools)
devtools::install_git("https://gitup.uni-potsdam.de/macroecology/mecofun.git"
)
```

```
# Load the mecofun package
library(mecofun)

# Names of our variables:
pred <- c('bio_2', 'bio_5', 'bio_14')

# We want three panels next to each other:
par(mfrow=c(1,3))

# Plot the partial responses
partial_response(m_glm, predictors = avi_df[,pred])
```

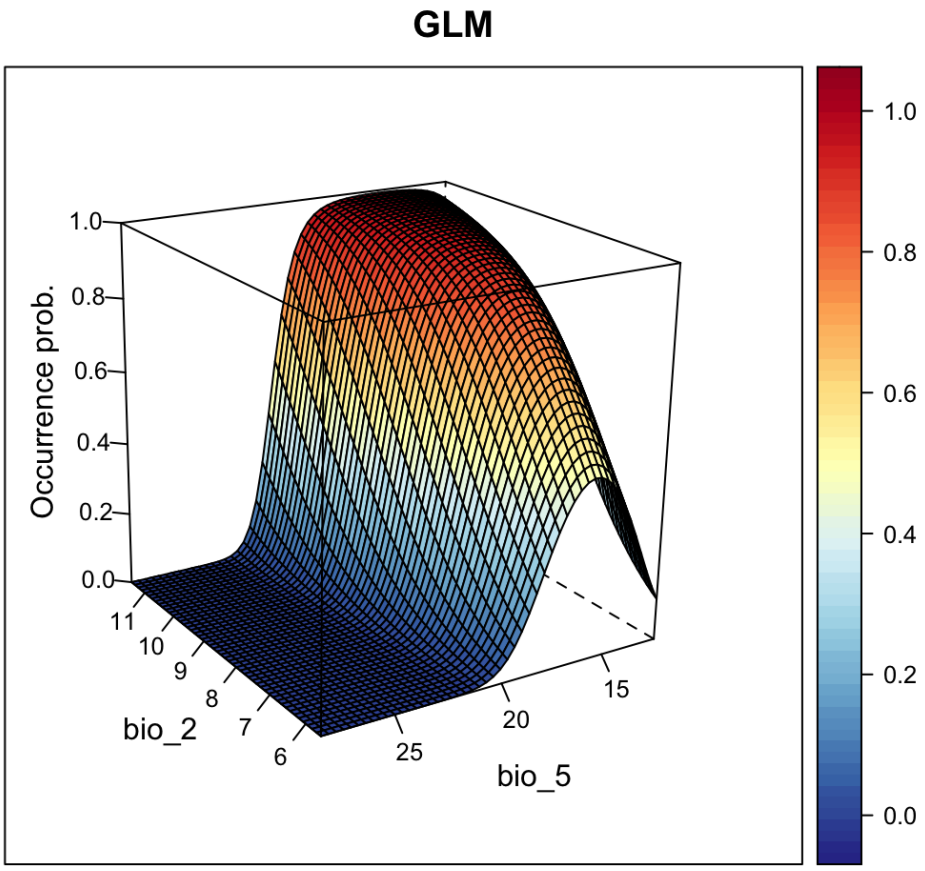


Often, partial response plots are difficult to interpret as they do not represent the full species-environment relationships. Alternatively, we can explore response surfaces, or inflated response curves that are a 2D-abstraction of response surfaces (Zurell, Elith, and Schroeder 2012).

```
library(RColorBrewer)
library(lattice)

# We prepare the response surface by making a dummy data set where two predic
tor variables range from their minimum to maximum value, and the remaining pr
edictor is kept constant at its mean:
xyz <- data.frame(expand.grid(seq(min(avi_df[,pred[1]]),max(avi_df[,pred[1
]]),length=50), seq(min(avi_df[,pred[2]]),max(avi_df[,pred[2]]),length=50)),
 mean(avi_df[,pred[3]]))
names(xyz) <- pred

# Make predictions
xyz$z <- predict(m_glm, xyz, type='response')
summary(xyz)
```

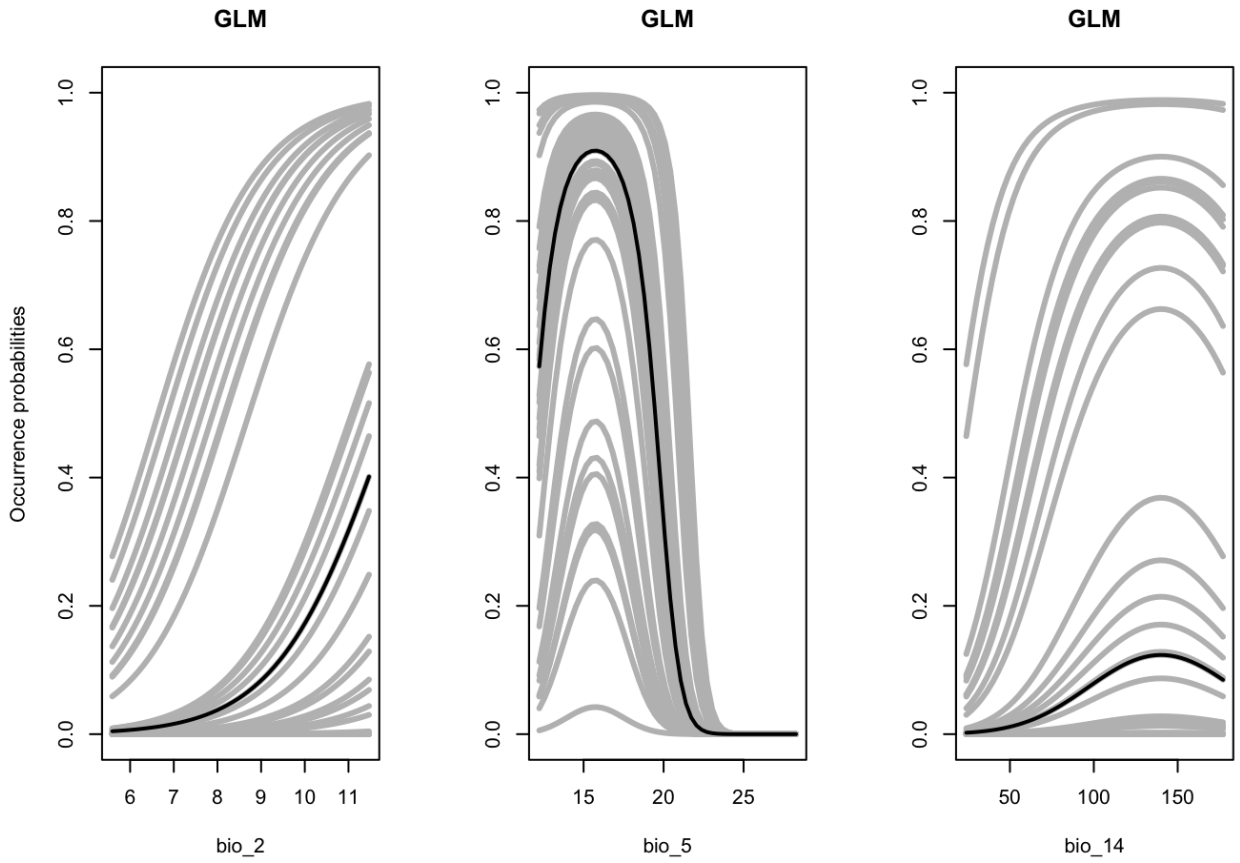
```
##      bio_2            bio_5           bio_14            z
##  Min.   : 5.592   Min.   :12.23   Min.   :80.97   Min.   :0.0000000
##  1st Qu.: 7.032   1st Qu.:16.16   1st Qu.:80.97   1st Qu.:0.0000582
##  Median : 8.531   Median :20.25   Median :80.97   Median :0.2366576
##  Mean   : 8.531   Mean   :20.25   Mean   :80.97   Mean   :0.3883815
##  3rd Qu.:10.031   3rd Qu.:24.35   3rd Qu.:80.97   3rd Qu.:0.8304635
##  Max.   :11.471   Max.   :28.27   Max.   :80.97   Max.   :0.9929319
```

```
# Make a colour scale
cls <- colorRampPalette(rev(brewer.pal(11, 'RdYlBu')))(100)

# plot 3D-surface
wireframe(z ~ bio_2 + bio_5, data = xyz, zlab = list("Occurrence prob.", rot=
90),
    drape = TRUE, col.regions = cls, scales = list(arrows = FALSE), zlim = c(
0, 1),
    main='GLM', xlab='bio_2', ylab='bio_5', screen=list(z = 120, x = -70, y =
3))
```



Codes for plotting inflated response curves (Zurell, Elith, and Schroeder 2012) are available in the `mecofun` package.

```
# Plot inflated response curves:
par(mfrow=c(1,3))
inflated_response(m_glm, predictors = avi_df[,pred], method = "stat6", lwd =
3, main='GLM')
```



2.3.1.2 Random forest (RF)

Random forests use a bagging procedure for averaging the outputs of many different CARTs (classification and regression trees)(Liaw and Wiener 2002). Bagging stands for “bootstrap aggregation”. Basically, we fit many CARTs to bootstrapped samples of the training data and then either average the results in case of regression trees or make a simple vote in case of classification trees (committee averaging)(Hastie, Tibshirani, and Friedman 2009; Guisan, Thuiller, and Zimmermann 2017). An important feature of random forests are the out-of-bag samples, which means that the prediction/fit for a specific data point is only derived from averaging trees that did not include this data point during tree growing. Thus, the output of Random Forests is essentially cross-validated. Random forests estimate variable importance by a permutation procedure, which measures for each variable the drop in mean accuracy when this variable is permutated.

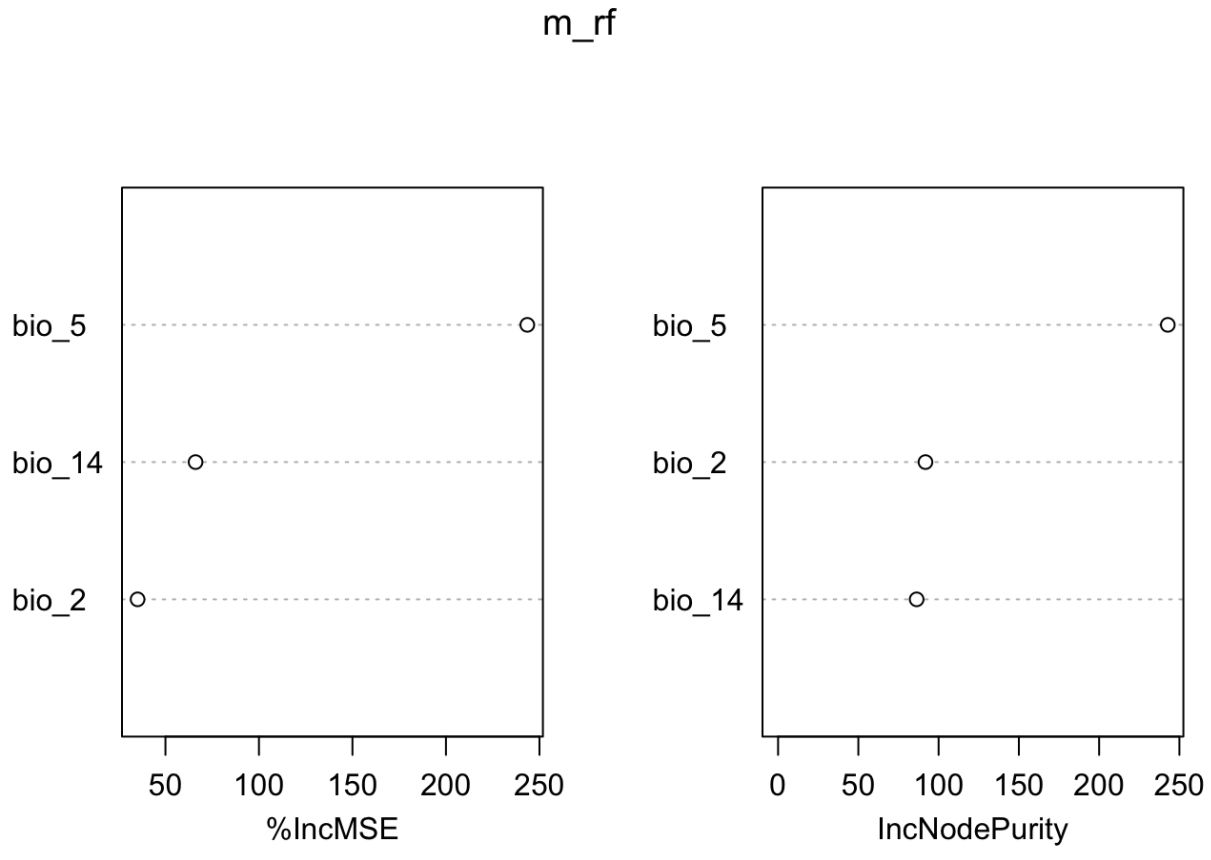
```
library(randomForest)

# Fit RF
(m_rf <- randomForest( x=avi_df[,2:4], y=avi_df[,1], ntree=1000, nodesize=10,
importance =T))
```

```
##
## Call:
##  randomForest(x = avi_df[, 2:4], y = avi_df[, 1], ntree = 1000,      nodes
ize = 10, importance = T)
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 0.07098877
##                     % Var explained: 62.62
```

```
# Variable importance:
importance(m_rf,type=1)
```

```
##          %IncMSE
## bio_5  243.50499
## bio_2   35.10351
## bio_14  66.10444
```

```
varImpPlot(m_rf)
```

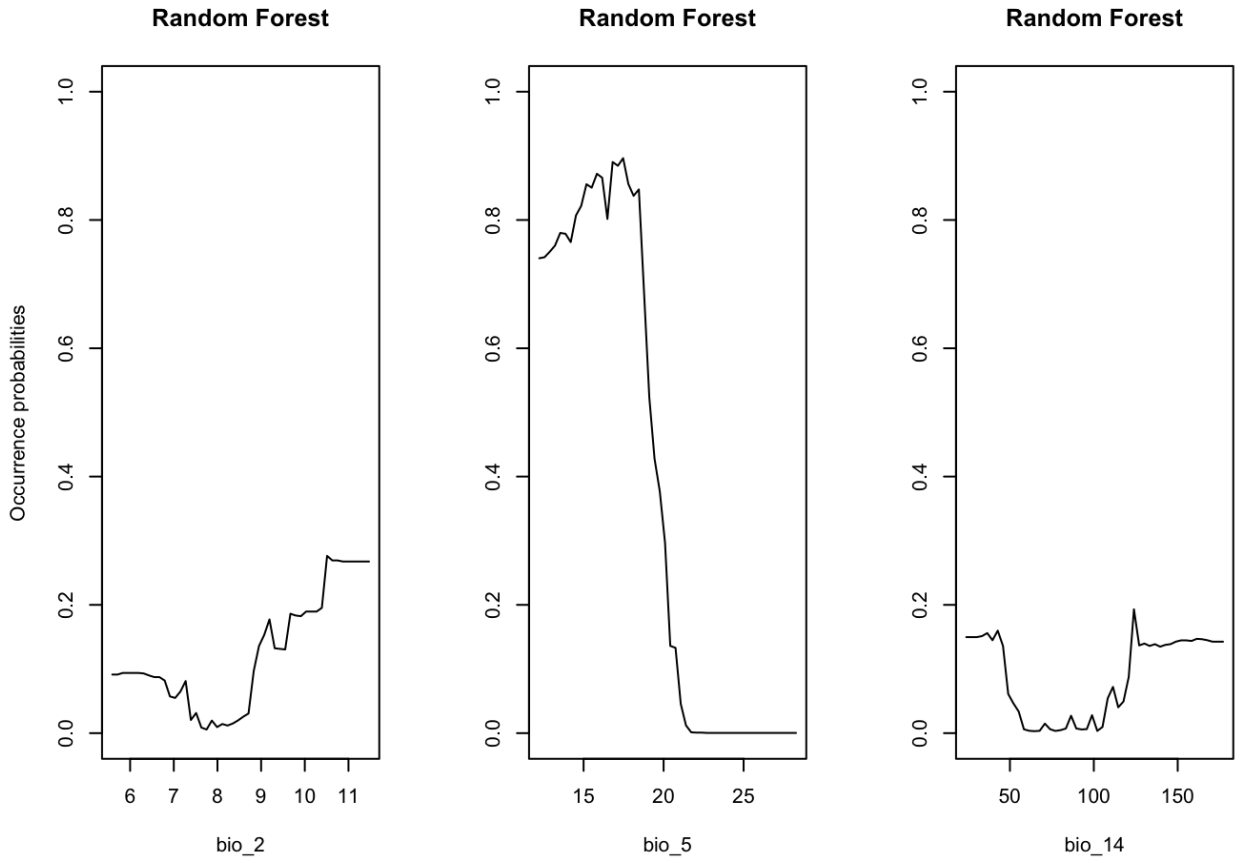



```
# Look at single trees:
head(getTree(m_rf,1,T))
```

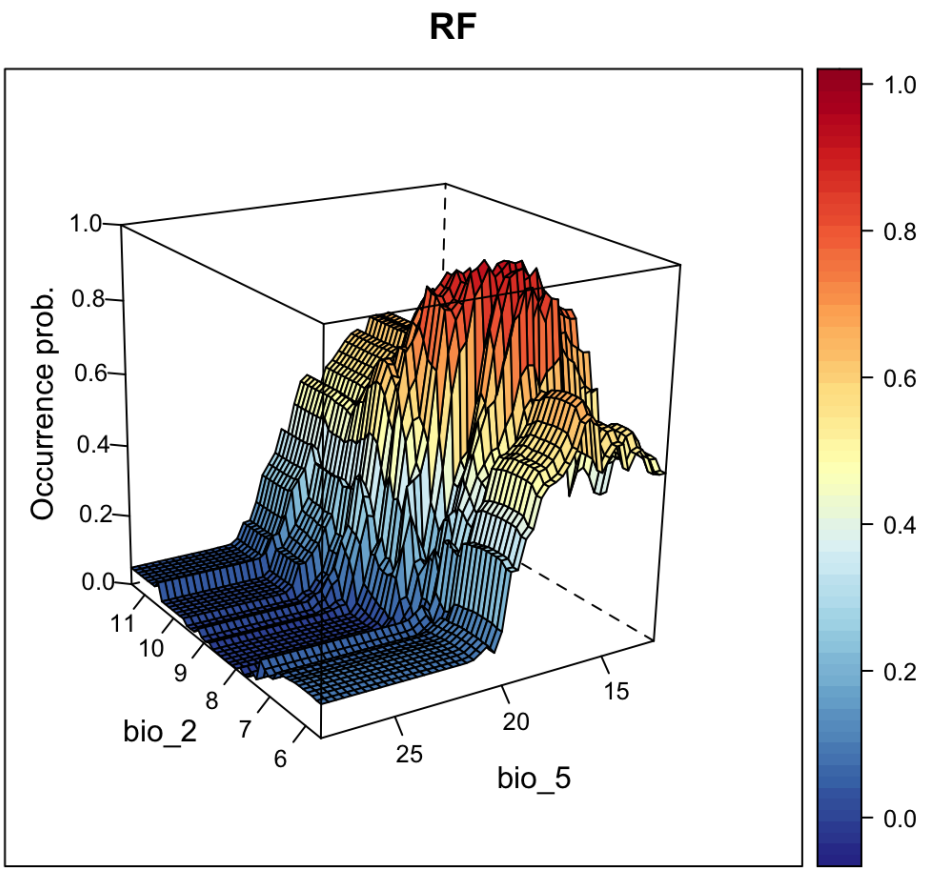
```
##   left daughter right daughter split var split point status prediction
## 1             2              3     bio_2    7.718725     -3  0.2331361
## 2             4              5    bio_14  111.500000     -3  0.6038062
## 3             6              7     bio_2    8.170200     -3  0.1236587
## 4             8              9    bio_14   76.500000     -3  0.4053156
## 5            10             11     bio_2    5.877262     -3  0.8194946
## 6            12             13    bio_14  107.500000     -3  0.2368421
```

Again, we first inspect partial response plots. Then, to get a better understanding of model behaviour over the entire environmental space, we explore 3D response surfaces, and inflated response curves that are a 2D-abstraction of response surfaces (Zurell, Elith, and Schroeder 2012).

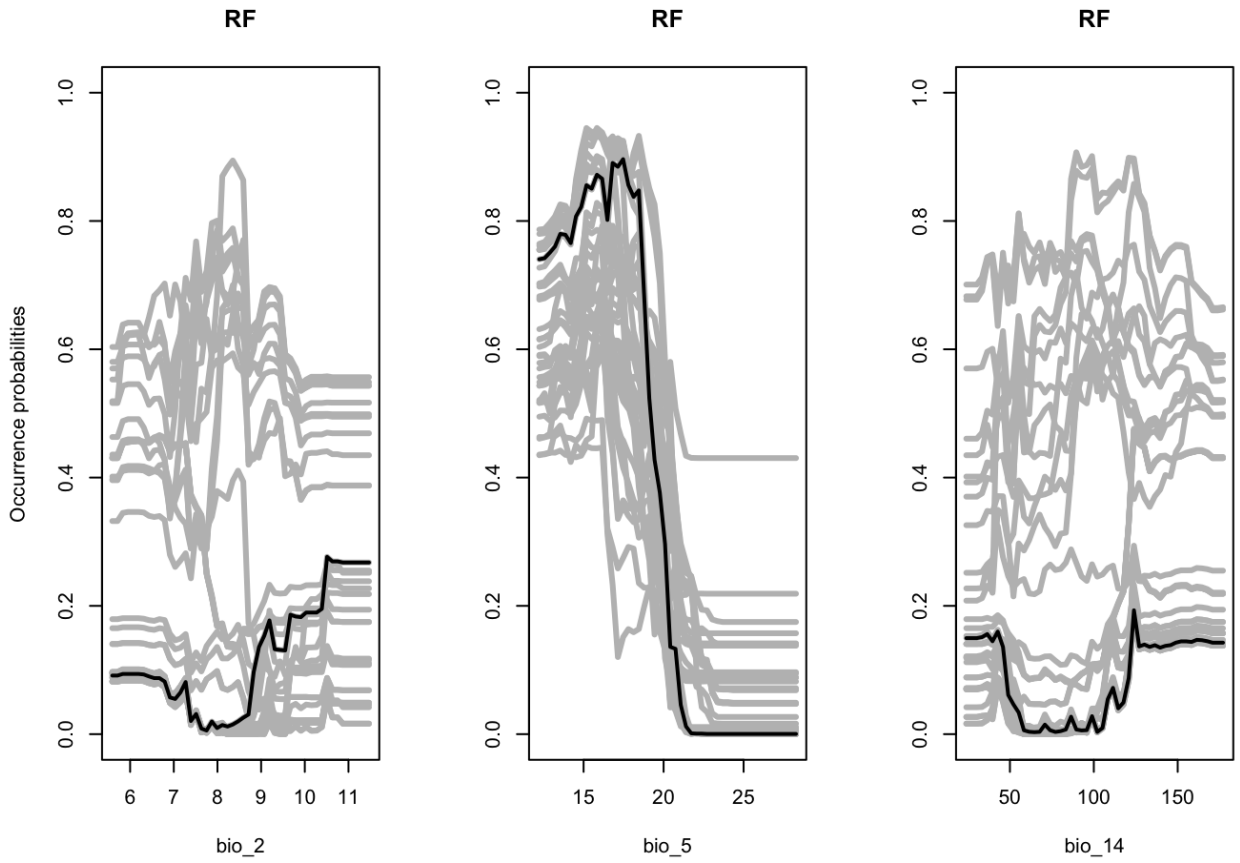
```
# Now, we plot response curves in the same way as we did for GLMs above:
par(mfrow=c(1,3))
partial_response(m_rf, predictors = avi_df[,pred], main='Random Forest')
```




```
# Plot the response surface:
xyz$z <- predict(m_rf, xyz)   # Note that we created the xyz data.frame in th
e GLM example above
wireframe(z ~ bio_2 + bio_5, data = xyz, zlab = list("Occurrence prob.", rot=
90),
    drape = TRUE, col.regions = cls, scales = list(arrows = FALSE), zlim = c(
0, 1),
    main='RF', xlab='bio_2', ylab='bio_5', screen=list(z = 120, x = -70, y =
3))
```




```
# Plot inflated response curves:
par(mfrow=c(1,3))
inflated_response(m_rf, predictors = avi_df[,pred], method = "stat6", lwd = 3
, main='RF')
```



Questions:

- What is the main difference in the extrapolation behaviour of GLMs and RFs?
- How could we achieve higher complexity of the species-environment relationship in GLMs?
- How could we achieve smoother responses, meaning less complexity of the species-environment relationship in RFs?
- What other SDM algorithms do you know?
- What are potential reasons for spatial autocorrelation in residuals?
- What kind of biodiversity data may show signals of temporal autocorrelation?
- What kind of biodiversity data could be nested?

2.4 Model assessment

In the model assessment step, we analyse the fitted model in depth. Strictly, checking the plausibility of the fitted species-environment relationship by visual inspection of response curves, and by assessing model coefficients and variable importance would also be part of the model assessment. However, to better understand what the different model algorithms were doing, we already explored this step during model fitting. Another crucial aspect of model assessment, which we will look at in more detail here, is assessing the predictive performance for a set of validation or test data (Hastie, Tibshirani, and Friedman 2009).

2.4.1 Ring Ouzel: Model assessment

We assess model predictive performance using (spatial block) cross-validation. Also, we select thresholds for binarising the predicted occurrence probabilities based on the cross-validated predictions.

First, we make cross-validated predictions to the spatial block tiles that were available from the data set in Zurell et al. (2020) (Figure 2.3).

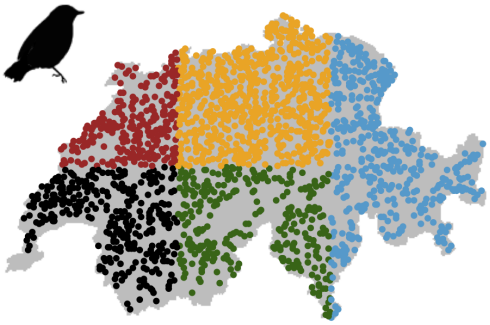
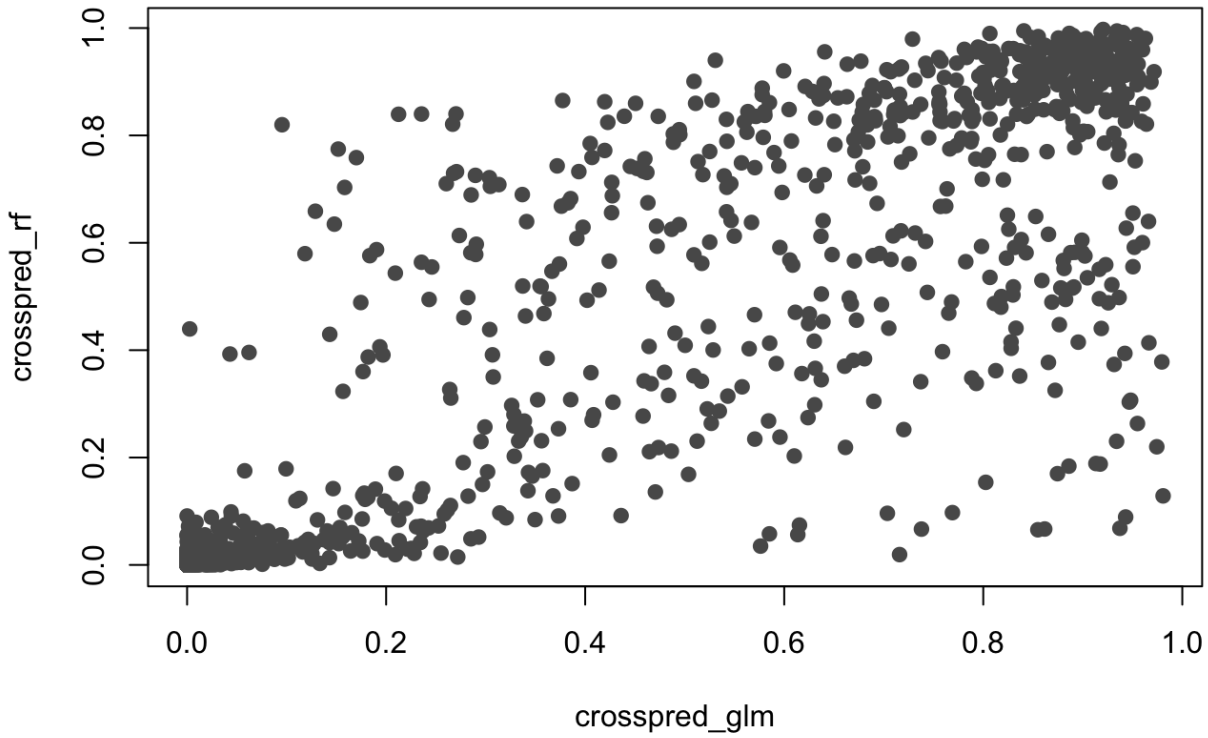


Figure 2.3: *Spatial blocks used in 5-fold spatial block cross-validation (from Zurell et al. 2020).*

```
# Make cross-validated predictions for GLM:
crosspred_glm <- mecofun::crossvalSDM(m_glm, traindat= avi_df[!is.na(avi_df$b
lockCV_tile),], colname_pred=pred, colname_species = "Turdus_torquatus", kfol
d= avi_df[!is.na(avi_df$blockCV_tile),'blockCV_tile'])

# Make cross-validated predictions for RF:
crosspred_rf <- mecofun::crossvalSDM(m_rf, traindat= avi_df[!is.na(avi_df$blo
ckCV_tile),], colname_pred=pred, colname_species = "Turdus_torquatus", kfold=
avi_df[!is.na(avi_df$blockCV_tile),'blockCV_tile'])

# Look at correlation between GLM and RF predictions:
plot(crosspred_glm, crosspred_rf, pch=19, col='grey35')
```



Next, we assess cross-validated model performance. We inspect different measures: AUC, the area under the receiver operating characteristic (ROC) curve (Hosmer and Lemeshow 2013); TSS, the true skill statistic (Allouche, Tsoar, and Kadmon 2006); sensitivity, the true positive rate; and specificity, the true negative rate. Simultaneously, we estimate the optimal threshold for making binary predictions. For this, we use a threshold that maximises TSS (= maximises the sum of sensitivity and specificity) (Liu et al. 2005).

```
(eval_glm <- mecofun::evalSDM(observation = avi_df[!is.na(avi_df$blockCV_til
e),1], predictions = crosspred_glm))
```

```
##         AUC       TSS     Kappa      Sens      Spec       PCC       D2 thr
esh
## 1 0.9458014 0.8146958 0.7091967 0.9803493 0.8343465 0.8720406 0.566055
0.17
```

```
(eval_rf <- mecofun::evalSDM(observation = avi_df[!is.na(avi_df$blockCV_til
e),1], predictions = crosspred_rf))
```

```
##         AUC       TSS     Kappa      Sens      Spec       PCC        D2 th
resh
## 1 0.9972359 0.9347666 0.9216467 0.9628821 0.9718845 0.9695603 0.8078386
0.51
```

We can also combine predictions from the two SDM algorithms and make an ensemble prediction, for example by taking the median.

```
# Derive median predictions:
crosspred_ens <- apply(data.frame(crosspred_glm, crosspred_rf),1,median)

# Evaluate ensemble predictions
(eval_ens <- mecofun::evalSDM(observation = avi_df[!is.na(avi_df$blockCV_til
e),1], predictions = crosspred_ens))
```

```
##         AUC       TSS     Kappa      Sens      Spec       PCC        D2 thr
esh
## 1 0.9838617 0.8784294 0.804557 0.9825328 0.8958967 0.9182638 0.7224253  0.
355
```

Questions:

- What is the main advantage of spatial block cross-validation over random cross-validation or random split-samples?
- What defines truely independent test data?
- Why is thresholding/binarising the probability output debated? What could be the problems?
- Which of the selected performance measures are more or less sensible to use in presence-only modelling?

2.5 Predictions

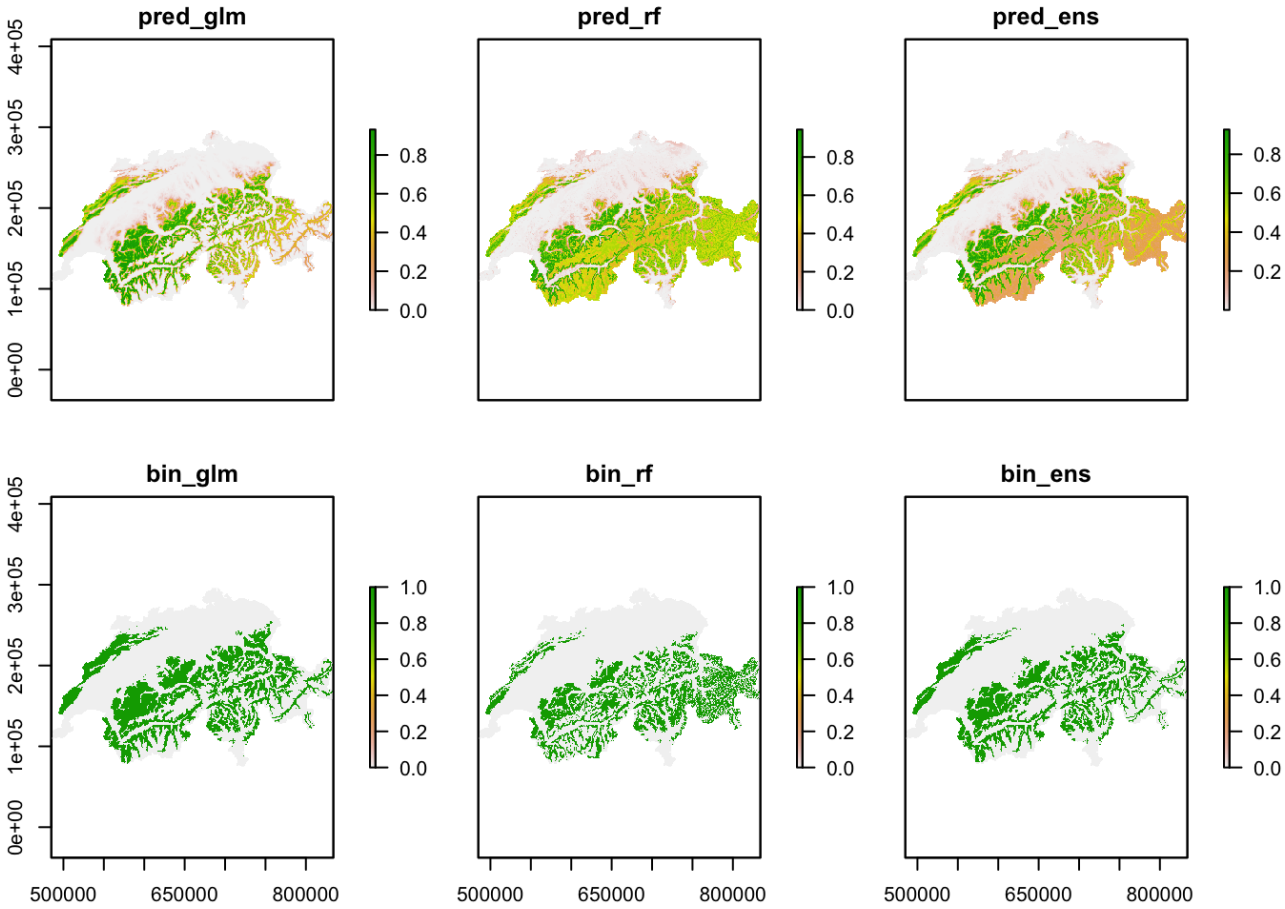
Now that we carefully fitted the SDMs, inspected model and extrapolation behaviour, and assessed predictive performance, it is finally time to make predictions in space and time. Importance points to consider here are quantification of uncertainty due to input data, algorithms, model complexity and boundary conditions (e.g. climate scenarios)(Araújo et al. 2019; Thuiller et al. 2019). When transferring the model to a different geographic area or time period, it is also recommended to quantify uncertainty due to novel environments (Zurell, Elith, and Schroeder 2012).

2.5.1 Ring Ouzel: Predictions

We aim to map potential current distribution of Ring Ouzel in Switzerland and compare this to potential future distribution under climate change. For simplicity, we only selected one climate model and one representative concentration pathway (RCP). Normally, it is recommended to compare the impact of several climate models and potentially assuming different RCPs.

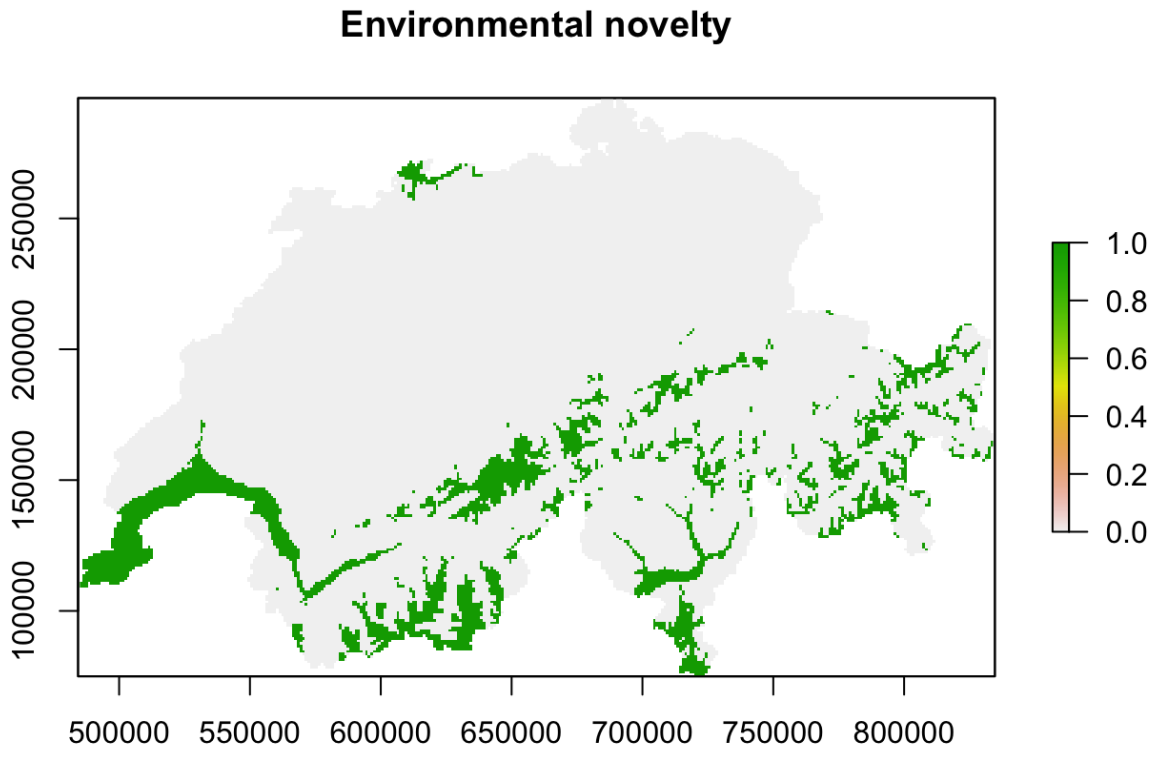
As we already prepared the spatial layers of current and future climate, mapping the potential distribution is straight forward.

```
# Make predictions to current climate:
bio_curr_df <- data.frame(rasterToPoints(bio_curr))
bio_curr_df$pred_glm <- mecofun::predictSDM(m_glm, bio_curr_df)
bio_curr_df$pred_rf <- mecofun::predictSDM(m_rf, bio_curr_df)
bio_curr_df$pred_ens <- apply(bio_curr_df[,-c(1:5)],1,median)

# Make binary predictions:
bio_curr_df$bin_glm <- ifelse(bio_curr_df$pred_glm > eval_glm$thresh, 1, 0)
bio_curr_df$bin_rf <- ifelse(bio_curr_df$pred_rf > eval_rf$thresh, 1, 0)
bio_curr_df$bin_ens <- ifelse(bio_curr_df$pred_ens > eval_ens$thresh, 1, 0)

# Make raster stack of predictions:
r_pred_curr <- rasterFromXYZ(bio_curr_df[,-c(3:5)])
plot(r_pred_curr)
```

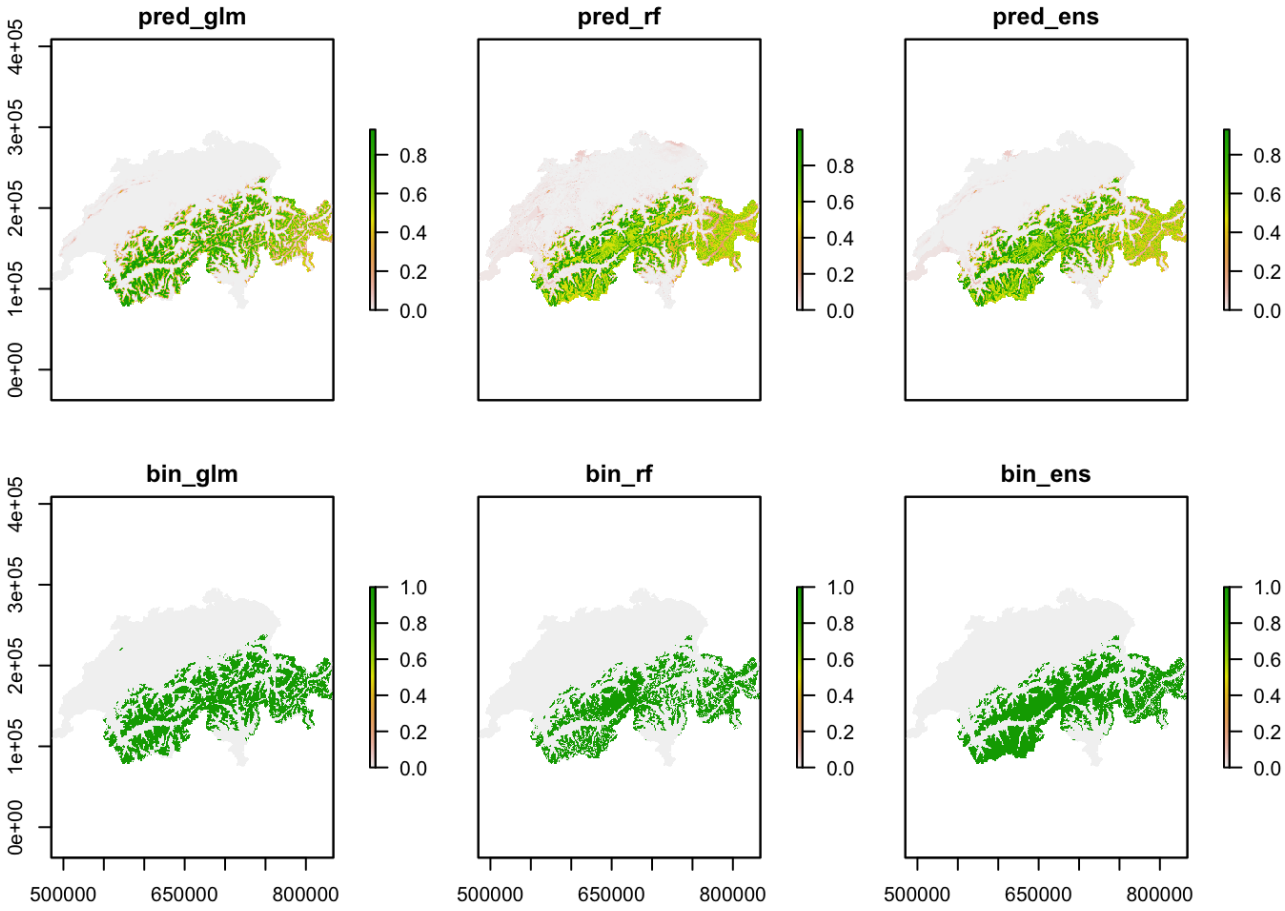


For the future climate layers, we proceed in the same way only that we additionally quantify novel environments. For this, we use codes from Zurell, Elith, and Schroeder (2012) that we already sourced earlier. The so-called *environmental overlap mask* will indicate where novel environmental conditions occur, meaning where the environmental conditions in the future climate layers lie beyond the range of environmental variables in the sample data, and will also indicate where novel combinations of environmental variables occur (Zurell, Elith, and Schroeder 2012).

```
# Assess novel environments in future climate layer:
bio_fut_df <- data.frame(rasterToPoints(bio_fut))
# Values of 1 in the eo.mask will indicate novel environmental conditions
bio_fut_df$eo_mask <- mecofun::eo_mask(avi_df[,pred], bio_fut_df[,pred])
plot(rasterFromXYZ(bio_fut_df[,-c(3:5)]), main='Environmental novelty')
```

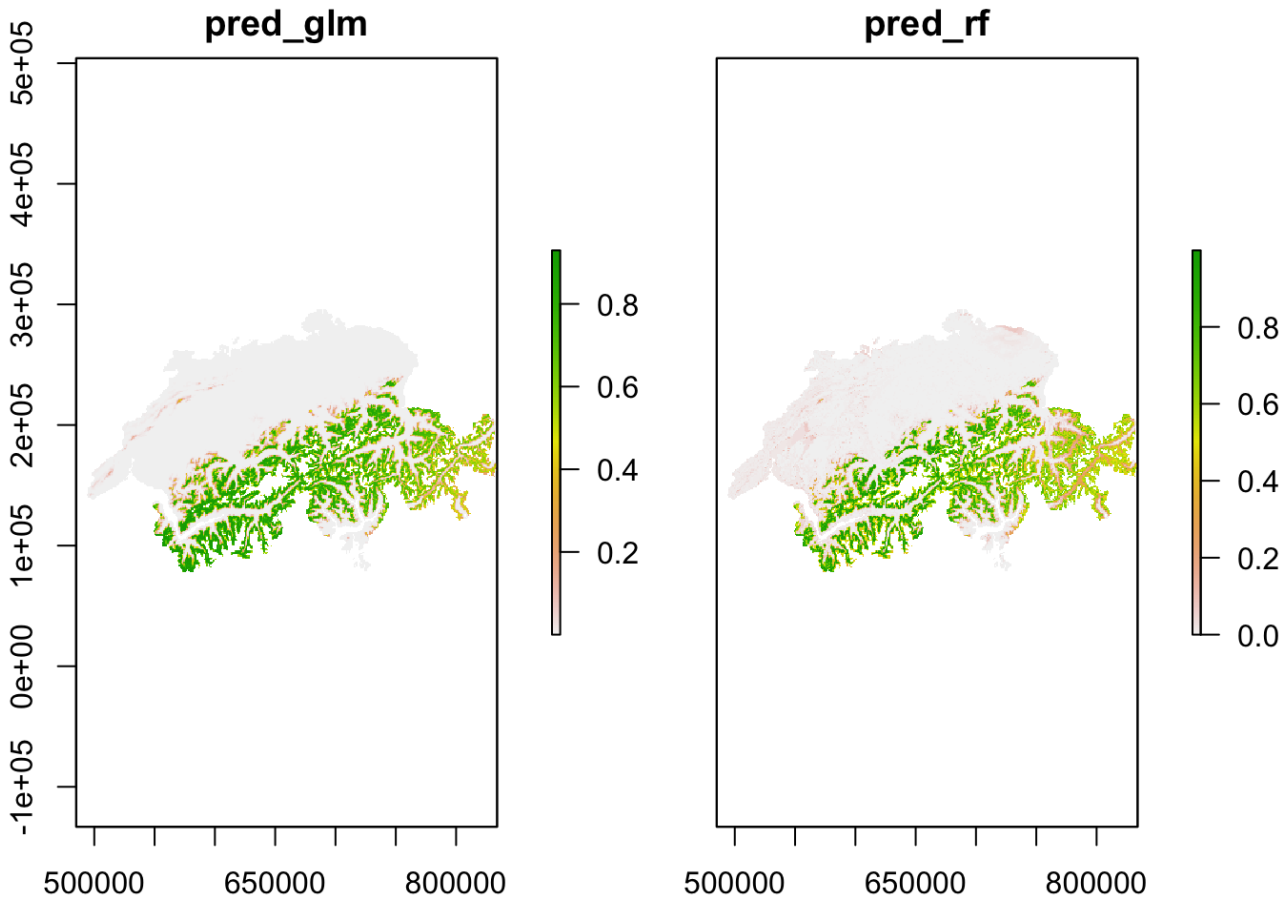



```
# Make predictions to future climate:
bio_fut_df$pred_glm <- mecofun::predictSDM(m_glm, bio_fut_df)
bio_fut_df$pred_rf <- mecofun::predictSDM(m_rf, bio_fut_df)
bio_fut_df$pred_ens <- apply(bio_fut_df[,-c(1:5)],1,median)

# Make binary predictions:
bio_fut_df$bin_glm <- ifelse(bio_fut_df$pred_glm > eval_glm$thresh, 1, 0)
bio_fut_df$bin_rf <- ifelse(bio_fut_df$pred_rf > eval_rf$thresh, 1, 0)
bio_fut_df$bin_ens <- ifelse(bio_fut_df$pred_ens > eval_ens$thresh, 1, 0)

# Make raster stack of predictions:
r_pred_fut <- rasterFromXYZ(bio_fut_df[,-c(3:5)])
plot(r_pred_fut[[-1]])
```

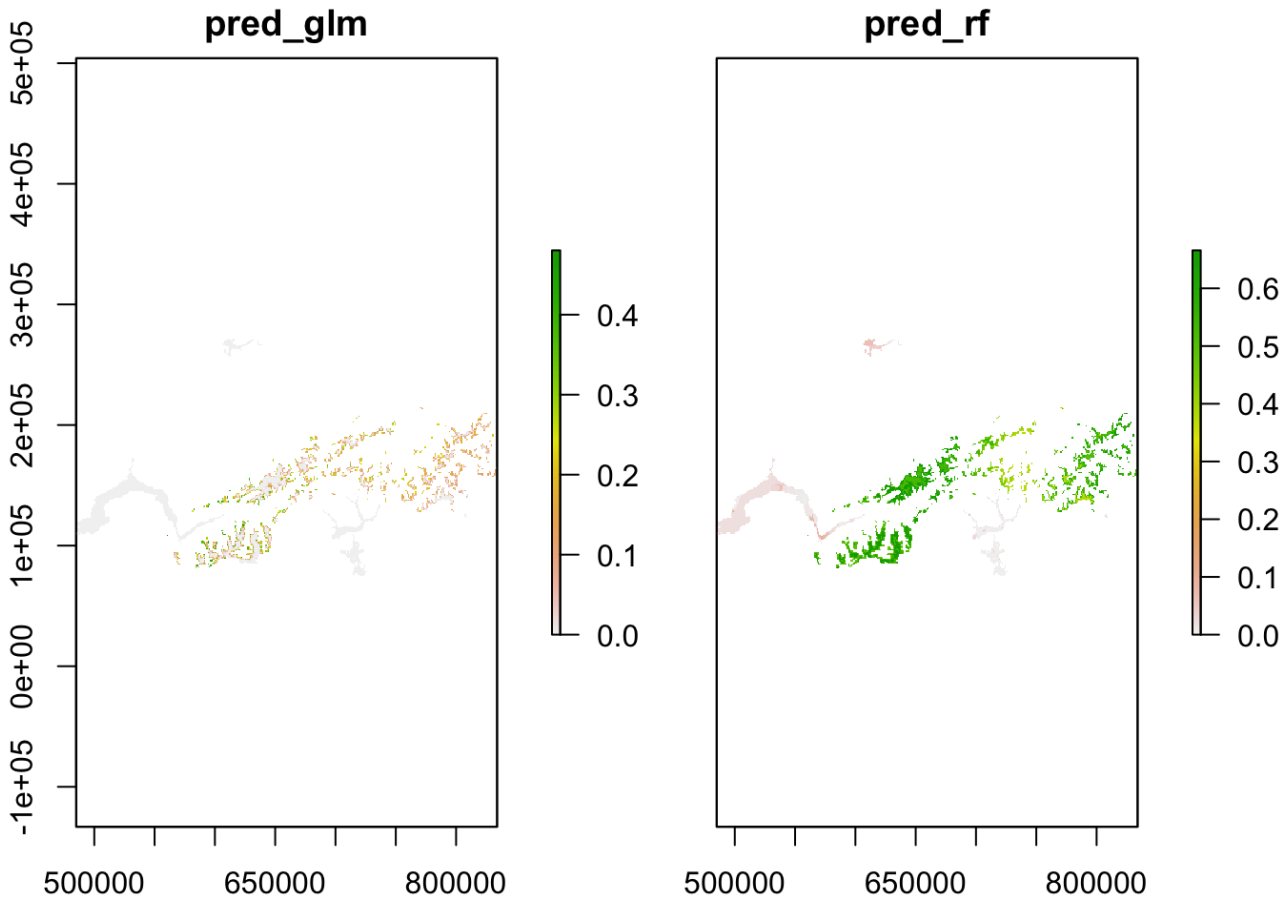


We can now explore model projections for analogous climates versus novel climates:

```
# Predictions to analogous climates:
bio_analog_df <- bio_fut_df[,c('x','y','pred_glm','pred_rf')]
bio_analog_df[bio_fut_df$eo_mask>0,c('pred_glm','pred_rf')] <- NA
plot(rasterFromXYZ(bio_analog_df))
```




```
# Predictions to novel climates:
bio_novel_df <- bio_fut_df[,c('x','y','pred_glm','pred_rf')]
bio_novel_df[bio_fut_df$eo_mask==0,c('pred_glm','pred_rf')] <- NA
plot(rasterFromXYZ(bio_novel_df))
```



Questions:

- Is the Ring Ouzel range expected to increas or decrease under future climate change?
- Where do Ring Ouzel lose and gain habitat under future climate change?
- Why is it important to quantify environmental novelty?
- With all your insights on model behaviour, model performance and predictions to future and novel environments, do you feel confident to decide which algorithm is better? Why?

3 Challenges and perspectives

As pointed out above, you should always critically assess the underlying assumptions of your SDM study and have to be aware of potential limitations. This short course is only meant to give a very brief introduction to SDMs and can by no means be exhaustive in all modelling aspects. If you are planning to use SDMs, I highly recommend consulting dedicated guides and textbooks on the topic.

As a last point, I want to point out that although SDMs constitute the mist widely used modelling approach in global change research, they also remain highly criticised in this context. This is largely to do with the underlying assumptions. Foremost, SDMs as introduced here assume that species are at equilibrium with environment, that we have sampled both the species and the environmental data perfectly, and that we have incorporated all major factors determining species range limits. These aspects are questionable for several reasons. First, species respond dynamically to global change, so they will almost certainly show transient dynamics. Important processes affecting biodiversity response to global change are physiology, demography, dispersal, interspecific interactions, adaptation, and the change in the environmental drivers (Urban et al. 2016; Zurell 2017). And, of course, all of these processes are acting on the species also here and now. Thus, ignoring these

processes when making transfers to other time periods or geographic areas, may considerably bias our predictions. Additionally, the observation process can bias our ability to detect certain signals (Guillera-Arroita 2017). There are several incentives to improve modelling approaches and better account for these processes, but still a lot of progress needs to be done (IPBES 2016; Guillera-Arroita 2017).

But to end on a positive note: SDMs are extremely useful tools when used carefully!

References

Allouche, Omri, Asaf Tsoar, and Ronen Kadmon. 2006. “Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (Tss).” *Journal of Applied Ecology* 43: 1223–32.

Araújo, M. B., R. P. Anderson, A. M. Barbosa, C. M. Beale, C. F. Dormann, R. Early, R. A. Garcia, et al. 2019. “Standards for Distribution Models in Biodiversity Assessments.” *Science Advances* 5: eaat4858.

Austin, M. P. 1980. “Searching for a Model for Use in Vegetation Analysis.” *Vegetatio* 42 (October). Springer Nature: 11–21.

Bruno, J. F., J. J. Stachowicz, and M. D. Bertness. 2003. “Inclusion of Facilitation into Ecological Theory.” *Trends in Ecology & Evolution* 18 (3): 119–25.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carre, J. R. Garcia Marquez, et al. 2013. “Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance.” *Ecography* 36: 27–46.

Dowle, M., and A. Srinivasan. 2019. *Data.table: Extension of ‘Data.frame‘*. https://CRAN.R-project.org/package=data.table (https://CRAN.R-project.org/package=data.table).

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, et al. 2006. “Novel Methods Improve Prediction of Species’ Distribution from Occurence Data.” *Ecography* 29: 129–51.

Elith, J., and J. R. Leathwick. 2009. “Species Distribution Models: Ecological Explanation and Prediction Across Space and Time.” *Annual Review of Ecology, Evolution, and Systematics* 40: 677–97.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. “A Working Guide to Boosted Regression Trees.” *Journal of Animal Ecology* 77: 802–13.

Elith, J., S. J. Phillips, T. Hastie, M. Dudik, Y. E. Chee, and C. J. Yates. 2011. “A Statistical Explanation of Maxent for Ecologists.” *Diversity and Distributions* 17: 43–57.

Franklin, J. 2010. *Mapping Species Distributions: Spatial Inference and Prediction*. Cambride University Press.

Freeman, E. A., and G. Moisen. 2008. “PresenceAbsence: An R Package for Presence Absence Analysis.” *Journal of Statistical Software* 23: 1–31.

Guillera-Arroita, G. 2017. “Modelling of Species Distributions, Range Dynamics and Communities Under Imperfect Detection: Advances, Challenges and Opportunities.” *Ecography* 40: 281–95.

Guisan, A., and W. Thuiller. 2005. “Predicting Species Distribution: Offering More Than Simple Habitat Models.” *Ecology Letters* 8: 993–1009.

Guisan, A., W. Thuiller, and N. E. Zimmermann. 2017. *Habitat Suitability and Distribution Models with Applications in R*. Cambride University Press.

Guisan, A., and N. E. Zimmermann. 2000. “Predictive Habitat Distribution Models in Ecology.” *Ecological Modelling* 135: 147–86.

Hanski, I. 1998. “Metapopulation Dynamics.” *Nature* 396: 41–49.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. Springer.

Hijmans, R. J. 2019. *Raster: Geographic Data Analysis and Modeling*. https://CRAN.R-project.org/package=raster (https://CRAN.R-project.org/package=raster).

Hosmer, David W., and Stanley Lemeshow. 2013. *Applied Logistic Regression*. 3rd ed. John Wiley & Sons, Inc.

Hutchinson, G. E. 1957. “Concluding Remarks, Cold Spring Harbor Symposium.” *Quantitative Biology* 22: 415–27.

IPBES. 2016. *The Methodological Assessment Report on Scenarios and Models of Biodiversity and Ecosystem Services*. Edited by S. Ferrier, K. N. Ninan, P. Leadley, R. Alkemade, L. A. Acosta, H. R. Akcakaya, L. Brotons, et al. Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity; Ecosystem Services, Bonn, Germany.

Liaw, A., and M. Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2: 18–22. https://CRAN.R-project.org/doc/Rnews/ (https://CRAN.R-project.org/doc/Rnews/).

Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. “Selecting Thresholds of Occurrence in the Prediction of Species Distributions.” *Ecography* 28: 385–93.

Merow, C., M. J. Smith, T. C. Edwards Jr, A. Guisan, S. M. McMahon, S. Normand, W. Thuiller, R. O. Wueest, N. E. Zimmermann, and J. Elith. 2014. “What Do We Gain from Simplicity Versus Complexity in Species Distribution Models?” *Ecography* 37: 1267–81.

Merow, C., M. J. Smith, and J. A. Silander Jr. 2013. “A Practical Guide to Maxent for Modeling Species’ Distributions: What It Does, and Why Inputs and Settings Matter.” *Ecography* 36: 1058–69.

Neuwirth, E. 2014. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer (https://CRAN.R-project.org/package=RColorBrewer).

Peterson, A. T., J. Soberon, R.G. Pearson, R. P. Anderson, E. Martinez-Meyer, M. Nakamura, and M. B. Araujo. 2011. *Ecological Niches and Geographic Distributions*. Princeton University Press.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. “Maximum Entropy Modeling of Species Geographic Distributions.” *Ecological Modelling* 190: 231–59.

Pulliam, H. R. 1988. “Sources, Sinks and Population Regulation.” *American Naturalist* 132: 652–61.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (https://www.R-project.org/).

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S: Hauenstein, et al. 2017. “Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure.” *Ecography* 40: 913–29.

Sarkar, D. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer. http://lmdvr.r-forge.r-project.org (http://lmdvr.r-forge.r-project.org).

Schmid, H., R. Luder, B. Naef-Daenzer, R. Graf, and N. Zbinden. 1998. *Schweizer Brutvogelatlas. Verbreitung Der Brutvoegel Inder Schweiz Und Im Fuerstentum Liechtenstein 1993-1996*. Swiss Ornithological Institute, Sempach, Switzerland.

Soberon, J. 2007. “Grinellian and Eltonian Niches and Geographic Distributions of Species.” *Ecology Letters* 10: 1115–23.

Thuiller, W., M. Guéguen, J. Renaud, D. N. Karger, and N. E. Zimmermann. 2019. “Uncertainty in Ensembles of Global Biodiversity Scenarios.” *Nature Communications* 10: 1446.

Thuiller, W., B. Lafourcade, R. Engler, and M. B. Araujo. 2009. “BIOMOD - a Platform for Ensemble Forecasting of Species Distributions.” *Ecography* 32: 369–73.

Urban, M. C., G. Bocedi, A. P. Hendry, J.-B. Mihoub, G. Pe’er, A. Singer, J. R. Bridle, et al. 2016. “Improving the Forecast for Biodiversity Under Climate Change.” *Science* 353: aad8466.

Zurell, D. 2017. “Integrating Demography, Dispersal and Interspecific Interactions into Bird Distribution Models.” *Journal of Avian Biology* 48: 1505–16.

Zurell, Damaris, Niklaus E. Zimmermann, Helge Gross, Andri Baltensweiler, Thomas Sattler, and Rafael O. Wüest. 2020. “Testing Species Assemblage Predictions from Stacked and Joint Species Distribution Models.” *Journal of Biogeography* 47 (1): 101–13. https://doi.org/10.1111/jbi.13608 (https://doi.org/10.1111/jbi.13608).

Zurell, D., J. Elith, and B. Schroeder. 2012. “Predicting to New Environments: Tools for Visualising Model Behaviour and Impacts on Mapped Distributions.” *Diversity and Distributions* 18: 628–34.

Zurell, D., and J.O. Engler. 2019. “Ecological Niche Modelling.” In *Effects of Climate Change on Birds*, edited by P.O. Dunn and A.P. Moller, 60–73. Oxford University Press.