

## Predicting to new environments: tools for visualising model behaviour and impacts on mapped distributions

by Damaris Zurell, Jane Elith and Boris Schröder

### SUPPORTING INFORMATION

#### Appendix S2 Simulated data example.

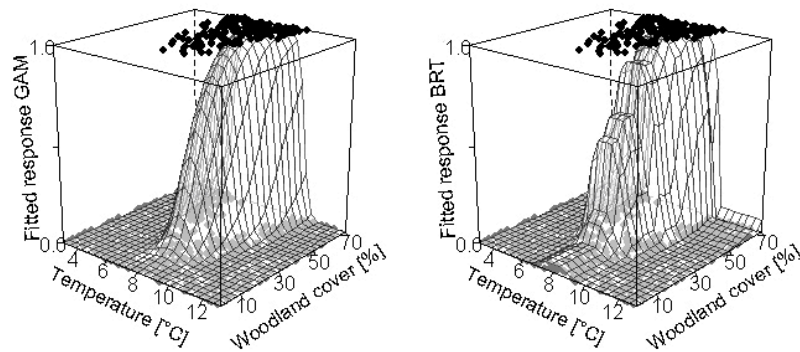
We simulated a virtual species (Zurell *et al.*, 2010) from a logistic regression model with two environmental variables, temperature and woodland cover. The virtual species showed a unimodal response to temperature and a linear positive response to percent woodland cover, in an additive model (cf. Fig. 2(a) in main article):

$$Y_i = \text{Bernoulli}(p_i)$$

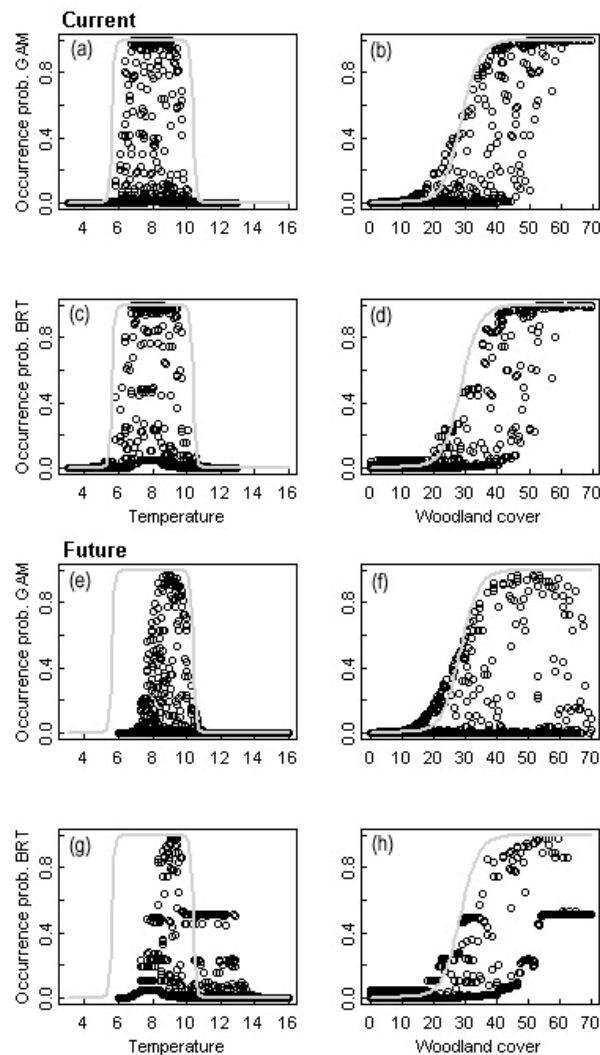
$$\ln\left(\frac{p_i}{1-p_i}\right) = a + b_1 \times \text{temperature}_i + b_2 \times (\text{temperature}_i)^2 + b_3 \times \text{woodland}_i + \varepsilon$$

$a = -170$		$\min(\text{temperature}) = 3$	
$b_1 = 40$		$\max(\text{temperature}) = 13$	
$b_2 = -2.5$	$i = 1 \quad 1000$	$\min(\text{woodland}) = 0$	$\varepsilon \sim N(\mu = 0, \sigma = 3)$
$b_3 = 0.35$		$\max(\text{woodland}) = 70$	

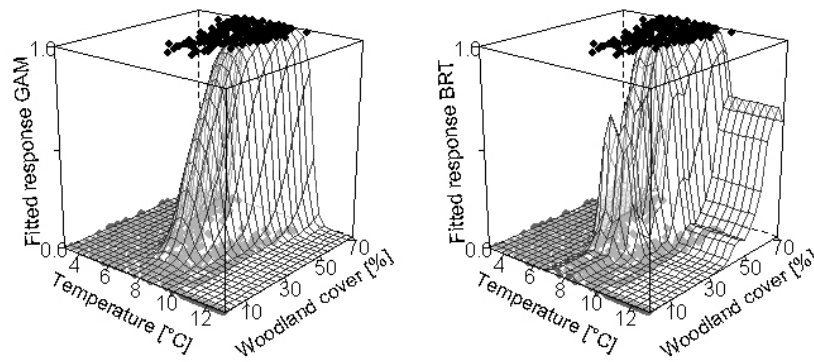
Three species were created whose niches were differently covered by data representing the situations of Fig. 1 (main article). (1) The niche of species 1 was entirely represented by data. Temperature and percent woodland cover at site  $i$  were drawn from uniform distributions between their minimum and maximum values (Fig. S1). (2) The niche of species 2 was truncated meaning that no samples existed for parts of the environmental niche of the species. Here, we assumed that high percent woodland cover would only occur within a specific temperature range (7-9 °C) and would linearly decline towards extreme temperatures (Fig. 2 in main article). (3) The niche of species 3 coincided with the edge of the sampled environmental space. In this case, we assumed zero percent woodland cover for all sites  $i$  with temperatures >10.5 °C (Fig. S3). Significant correlations between environmental variables were only found for species 3 ( $\rho = -0.46$ ). For all three cases, climate warming was simulated by increasing the temperature values of sites  $i$  in the training data by +3 °C. The entire code necessary to repeat the simulated data example is contained in Appendix S1.



**Figure S1.** Response surfaces for simulated species 1 whose niche is entirely encompassed within sample space. Both methods, GAM (left) and BRT (right), show congruent model behaviour.



**Figure S2.** Fitted values for simulated species 2 with truncated niche. Panels show the fitted values of GAM (a-b, e-f) and BRT (c-d, g-h) plotted along temperature gradient (left) and woodland cover gradient (right) for predictions to current environment (a-d) and for predictions under climate warming (+3 °C; e-h). Grey lines depict true response of species 2.



**Figure S3.** Response surfaces for simulated species 3 whose niche edge coincides with edge of sample space. BRT (right) shows erroneous extrapolation behaviour in those parts of the environment space where no samples exist beyond the species niche edge.

### Appendix S3 Method descriptions for visualisation tools.

Here, we describe the environmental overlap masks and inflated response curves in more details. Respective R codes are contained in Appendix S1.

#### Environmental overlap mask

*Function call:* `eo.mask()`

*Description:* The environmental overlap mask is an extension of MaxEnt's multivariate environmental similarity surface (MESS) maps (Elith *et al.*, 2010). The latter are BIOCLIM-style estimates that take the full gradients of the environmental variables that are covered by data, extend these into a hyper-dimensional rectangle and define all environmental conditions outside that rectangle or box as novel conditions. We could think of these boxes as 'implied sample space' because they pretend that all possible combinations of environmental variables within the box are represented in the sample data which may not be the case (Fig. 1 in main article). Now, the environmental overlap mask splits these hyper-dimensional rectangles or boxes into smaller bins. Then, it compares the bins in the test or prediction data set to the bins in the training data and defines all non-overlapping bins as novel environmental conditions. Thus, depending on the number of bins the box is split into, the environmental overlap mask will also allow identifying those parts of the box that are within the sampled ranges of each predictor variable treated individually but are nevertheless outside the sample space. Note that a bin number of one equates to the border that distinguishes novel space (negative values) in MESS maps.

*Arguments:* `eo.mask()` takes four arguments, two of which are obligatory.

- `traindata` must be a `data.frame` (obligatory)

- newdata must be a data.frame (obligatory)
- nbin determines the number of bins the environmental gradients are split into (default nbin=5). Thereby, nbin refers to the number of bins each univariate predictor is split into. The actual number of bins is then  $nbin^n$  where n is the number of predictors.
- type defines the output. This could either be a vector of length nrow(newdata) containing for each entry in newdata zeros for analogue environments and ones for novel environments (type='EO', the default) or a character vector returning for each entry in newdata a unique identifier for the bin the data point belongs to (type='ID'). The latter may help distinguishing data entries from sample space, 'implied sample space' or novel space beyond that.

*Required R packages:* base.

### Inflated response curves

*Function call:* inflated.response( )

*Description:* Inflated response curves are an abstracted 2D version of multidimensional response surfaces that help to visualise the combined effects of all variables in the model over their full range. They show the effect of a variable on the response while accounting not only for the average effects of the other variables but also for minimum and maximum (and median and quartile) values. Thus, the response plot for any one variable consists of many response curves representing all possible combinations of all other variables in the model. Because the number of combinations grows exponentially with the number of variables and restricts computational feasibility, we use Latin hypercube sampling to reduce dimensionality for large numbers of variables. This is simply a means to efficiently sample a representative subset from all possible combinations of environmental predictors (Carnell, 2009). Note that due to stochasticity different replicates of inflated response plots may differ depending on the number of variables and the Latin hypercube sample size.

*Arguments:* inflated.response( ) takes 17 arguments, two of which are obligatory.

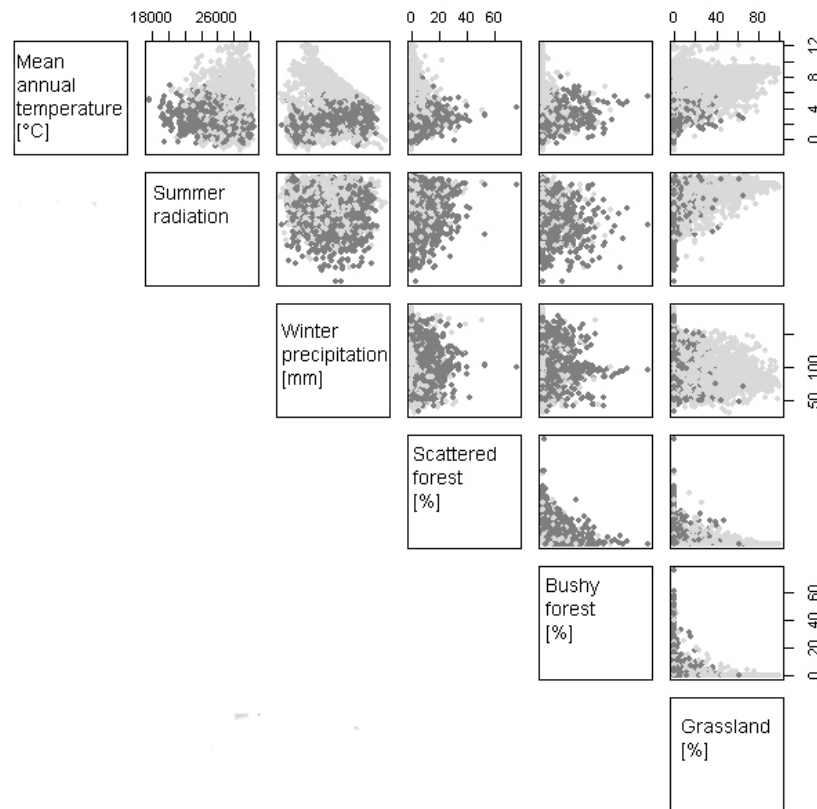
- object denotes the model object for which response plots are desired, e.g. a GAM model (obligatory)
- predictors is as data.frame containing the predictor variables of object (usually the training data) (obligatory)

- `select.columns` allows to determine a subset of predictors for which inflated response plots are to be drawn (defaults to `NULL` meaning that one plot is made for each column of predictors)
- `label` can be a character vector of names used to label the x axes of the plots (defaults to `NULL` meaning that column names of predictors are used for labelling)
- `ylab` takes a character string for labelling the y axis (defaults to 'Occurrence probabilities').
- `len` determines the number of data points used to draw the response curves (defaults to 50)
- `lhsample` determines the number of Latin hypercube samples (defaults to 100 meaning that 100 LH samples are drawn if more than 100 combinations of environmental predictor values exist)
- `method` determines the summary statistics that are used to characterise the predictors. 'mean' corresponds to conventional partial dependence plots. 'stat3' (the default) considers mean, minimum and maximum values of predictors. 'stat6' considers min, mean, median, max and quartiles. Thus,  $3^{n-1}$  combinations of environmental predictor values exist for 'stat3' and  $6^{n-1}$  for 'stat6' where n is the number of predictors in object.
- `disp` can take options 'all' (the default) or 'eo.mask'. In the latter case, `eo.mask( )` is used to mask all combinations outside the sample space as novel conditions. These parts of the response curves can then be displayed differently.
- `overlay.mean` takes a Boolean value determining whether the mean response should be overlayed over the inflated response curves (defaults to 'True')
- `col.curves` (defaults to 'grey'), `col.novel` (defaults to 'grey'), `col.mean` (defaults to 'black') are graphic parameters determining the color of the response curves, those parts of the curves that are identified as novel by `eo.mask( )`, and the overlayed mean response curve.
- `lwd` (defaults to 1), `lwd.mean` (defaults to 2), `lwd.known` (defaults to 2) are graphic parameters. If `disp='all'` then `lwd` determines the line width of all curves. If `disp='eo.mask'` then `lwd` determines the line width of the novel parts and `lwd.known` the line width for those predictor combinations present in the sample space. `lwd.mean` determines the line width of the overlayed mean response curve.
- `ylim` sets the y axis limits (defaults to `c(0,1)`)

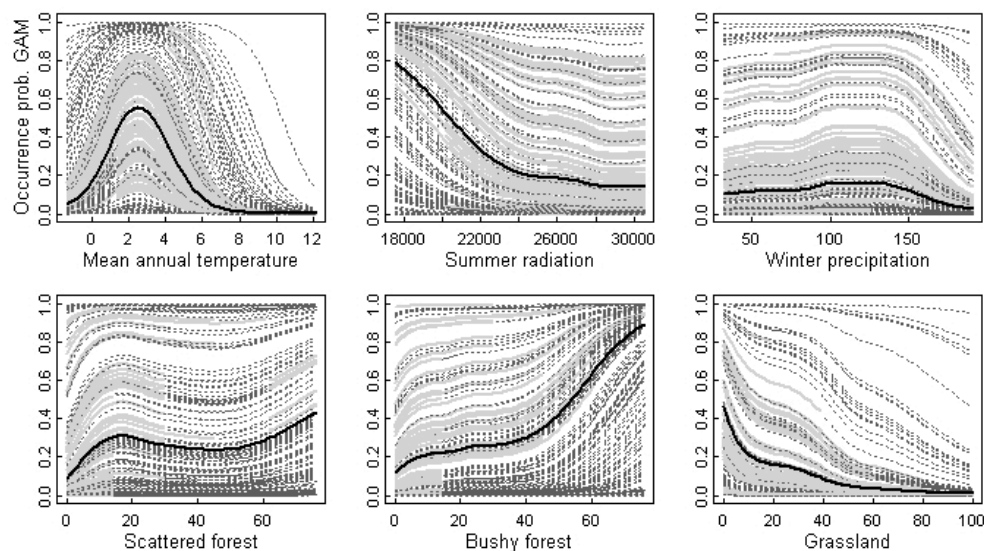
*Required R packages:* base, lhs (Latin Hypercube Sample package), any packages relating to specific SDM methods one wants to include (e.g. gam or gbm library), additional code for `eo.mask()`.

#### **Appendix S4** Swiss black grouse example.

A case study for Swiss black grouse was developed in Zurell *et al.* (2011) and we here give only a very brief description of the species and data. In the Swiss Alps, black grouse (*Tetrao tetrix*) mainly occur in treeline habitats at altitudes of up to 2500 m above sea level. Species distribution data at 1 km resolution were obtained from the Swiss Breeding Bird Atlas (Schmid *et al.*, 1998). Environmental predictors included climatic variables as well as land use and vegetation data at 1 km resolution. These were pre-selected so that bivariate Spearman correlations were below  $|r|=0.7$ . Climate data were derived from the BIOCLIM database (Swiss Federal Research Institute WSL). We here used only the most important climatic predictors, namely mean annual temperature, potential solar radiation in summer months (June-August) and winter precipitation (sums December-February). Land use and vegetation data were compiled from the land use and land cover database GEOSTAT (Swiss Federal Statistical Office). From these, we chose the three most important predictors namely percentage cover of scattered forest, bushy forest and grassland (Fig. S4). For illustrative purposes, we chose a rather extreme climate change scenario from the HadCM3 global circulation model which corresponded to the socio-economic scenario A1F1 for the year 2100 downscaled to a 1 km resolution (Swiss Federal Research Institute WSL). For this scenario, mean annual temperature increased by +7.7 °C and winter precipitation increased by +48 mm on average.

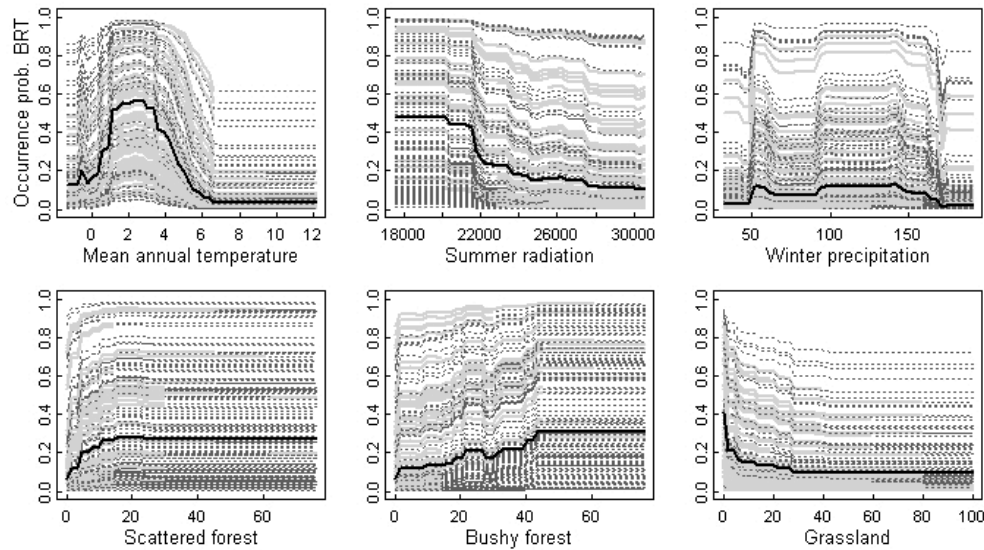


**Figure S4.** Scatterplots of black grouse presences (dark grey) and absences (light grey) along environmental gradients. Although the single environmental predictors cover large gradients, only portions of all possible combinations were present in the sample.



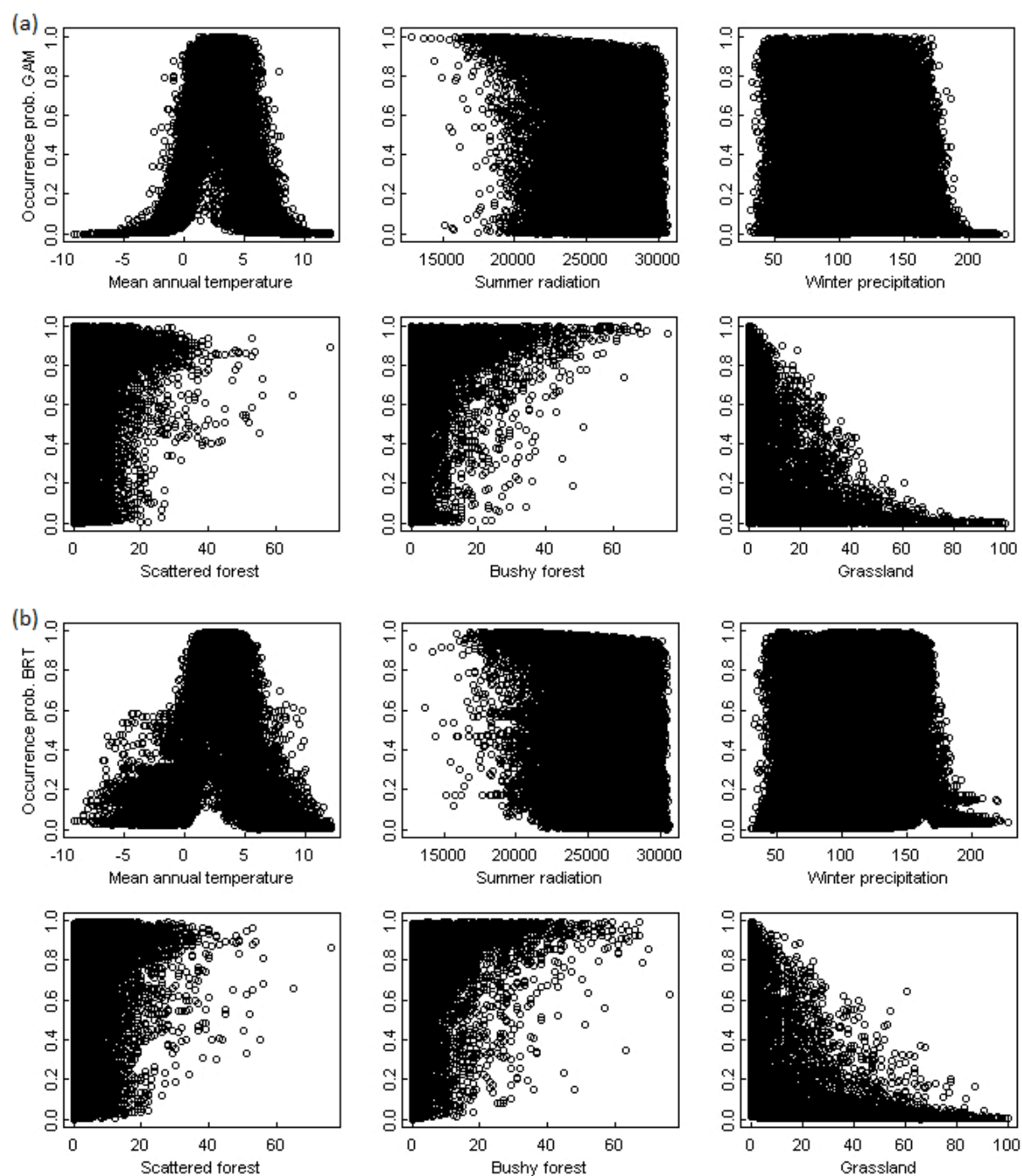
**Figure S5.** Inflated response curves for GAM along the sampled environmental gradients. Bold black lines show the mean effects as provided by conventional partial dependence plots. Light grey lines and dotted dark grey lines depict predictor effects over the full range of the other predictors (minimum, maximum, median, mean and quartiles). Light grey lines indicate combinations of environmental predictors that

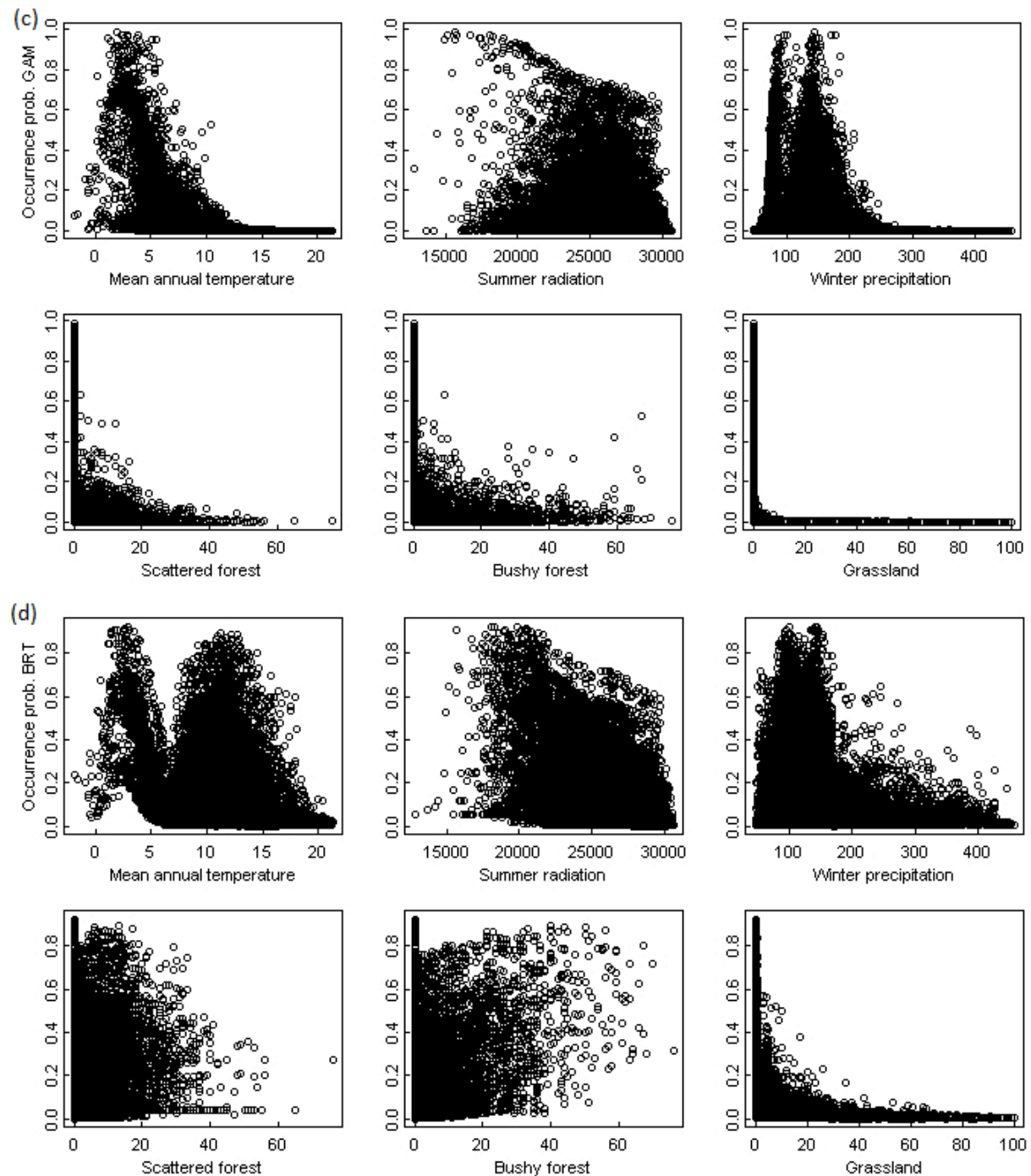
were observed in the sample space while dotted dark grey lines indicate extrapolations to novel, unsampled combinations. The plots represent  $n=150$  Latin hypercube samples from all possible combinations of environmental predictors.



**Figure S6. Inflated response curves for BRT along the sampled environmental gradients. Bold black lines show the mean effects as provided by conventional partial dependence plots. Light grey lines and dotted dark grey lines depict predictor effects over the full range of the other predictors (minimum, maximum, median, mean and quartiles). Light grey lines indicate combinations of environmental predictors that were observed in the sample space while dotted dark grey lines indicate extrapolations to novel, unsampled combinations. The plots represent  $n=150$  Latin hypercube samples from all possible combinations of environmental predictors.**







**Figure S7. Predictions for Swiss black grouse.** Panels show the fitted values of GAM (a, c) and BRT (b, d) plotted along the environmental gradients for predictions to current environment (a-b) and for predictions to climate change scenario for the year 2100 (A1F1 scenario from HadCM3 with average temperature increase of 7.7 °C and average precipitation increase of 48 mm for Switzerland). While fitted values are largely congruent for current predictions (a-b), we see marked differences in fitted values for the climate change scenario (c-d) most noticeably in the temperature and forest predictors. While GAM predicts low occurrence probabilities for high temperatures (c) as is also suggested by current fits (a) and which could imply a physiological limit to temperature, BRT predicts high occurrence probabilities even for high temperatures (d).

## References

- Carnell, R. (2009). lhs: Latin Hypercube Samples. R package version 0.5.
- Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330-342.
- Schmid, H., Luder, R., Naef-Daenzer, B., Graf, R. & Zbinden, N. (1998) Schweizer Brutvogelatlas. Verbreitung der Brutvögel in der Schweiz und im Fürstentum Liechtenstein 1993-1996. Swiss Ornithological Institute, Sempach, Switzerland.
- Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbaß, N., Pagel, J., Reineking, B., Schröder, B. & Grimm, V. (2010) The virtual ecologist approach: simulating data and observers. *Oikos*, **119**, 622-635.
- Zurell, D., Grimm, V., Rossmannith, E., Zbinden, N., Zimmermann, N. E. & Schröder, B. (2011) Uncertainty in predictions of range dynamics: Black Grouse climbing the Swiss Alps. *Ecography*. DOI: 10.1111/j.1600-0587.2011.07200.x.