Species distribution models – modelling cycle

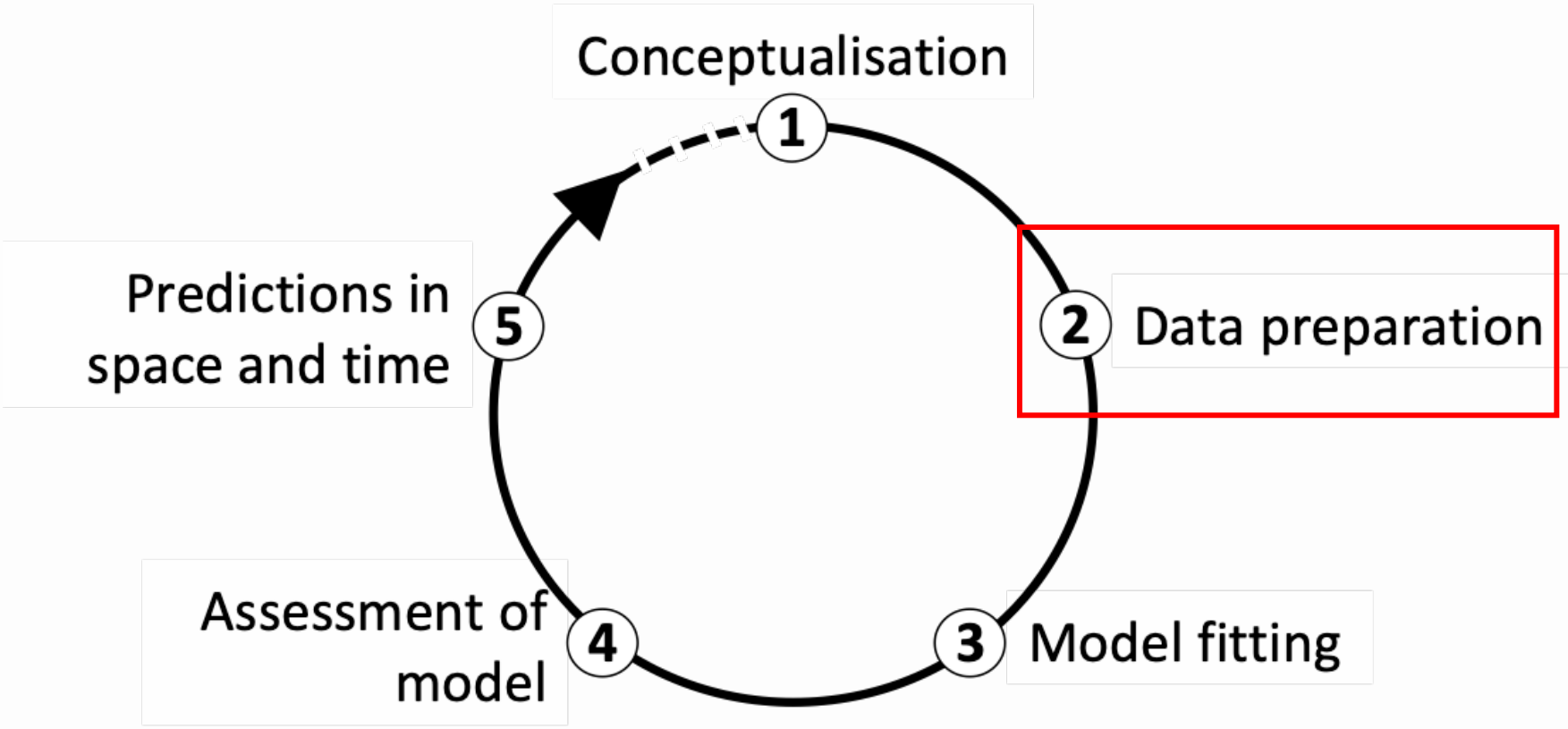
Prof. Dr. Damaris Zurell

Ecology & Macroecology Lab

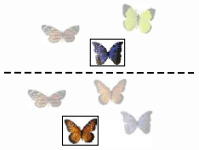
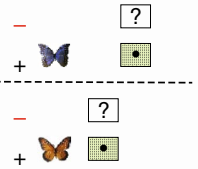
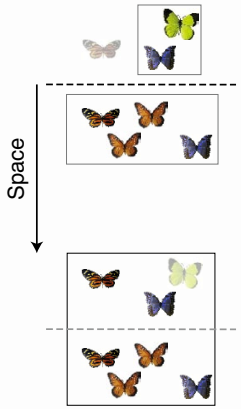
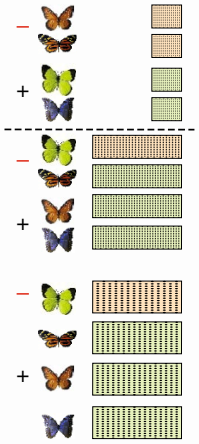
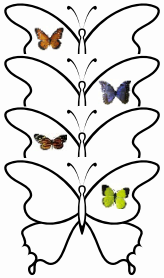
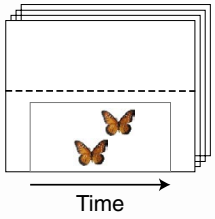
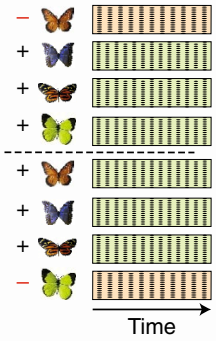
https://damariszurell.github.io

@ZurellLab

SDMs – model building steps



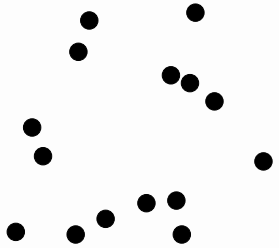
Species data types

Data type	Sampling Method	Taxonomic scope	Spatiotemporal scope	Output	Characteristics	Examples
Incidental records		?			• Single species • Spatiotemporally specific • Absences unknown	• Museum records • Many amateur observations • GBIF points
Small-area inventories					• Multi-species • Spatiotemporally specific • Absences often reliable	• Surveys with protocol, gridded atlas efforts • Relevé or forest plots • Visual, acoustic sensors • Trap-/trawl-based surveys
Large-area inventories					• Multi-species • Spatiotemporally unspecific • Absences somewhat reliable	• Regional checklists • National Park inventories
Expert synthesis maps					• Single species focus • Spatiotemporally unspecific • Absences somewhat reliable	• Expert ranger maps, following manual or regional delineation

Jetz et al. (2019) Nature Ecol Evol 3: 539-551.

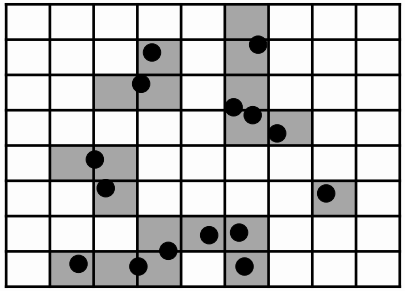
Species data types

Point occurrences



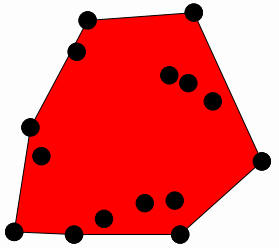
e.g. GBIF, OBIS

Inventories: grid maps



e.g. UK breeding bird atlas

Range map



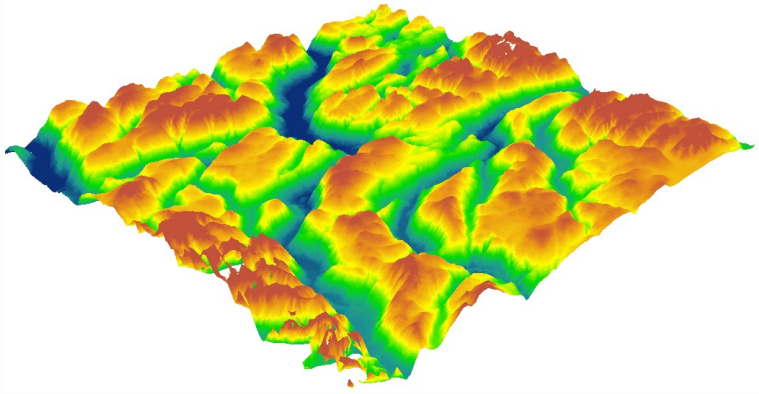
e.g. IUCN

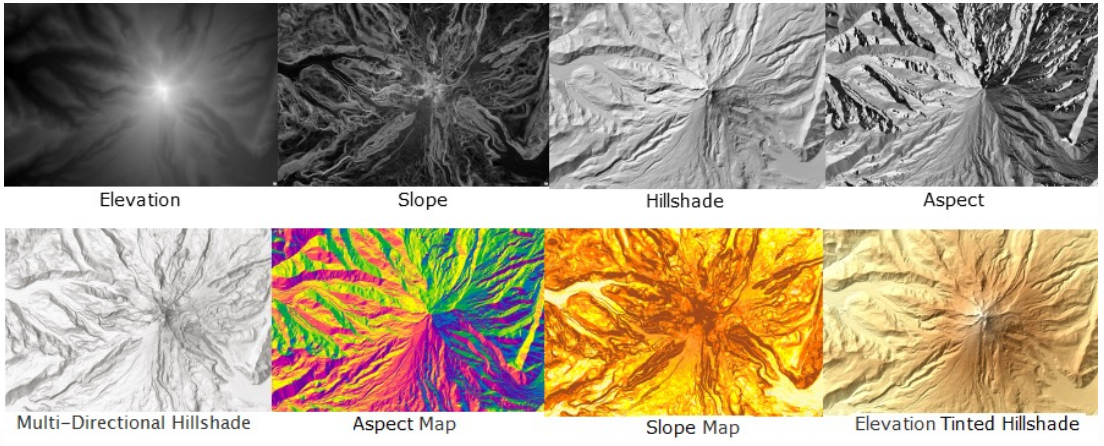
Environmental data types

- Digital elevation data and derivatives
- Climate data (incl. climate scenarios)
- Land use data
- Remote sensing

Digital elevation model (DEM)

- A three-dimensional representation of a terrains‘s surface
- Useful for deriving topographic predictors like altitude, slope, and aspect etc.
- Free global sources, e.g. SRTM (Space Shuttle Radar Topography Mission; 30 m)



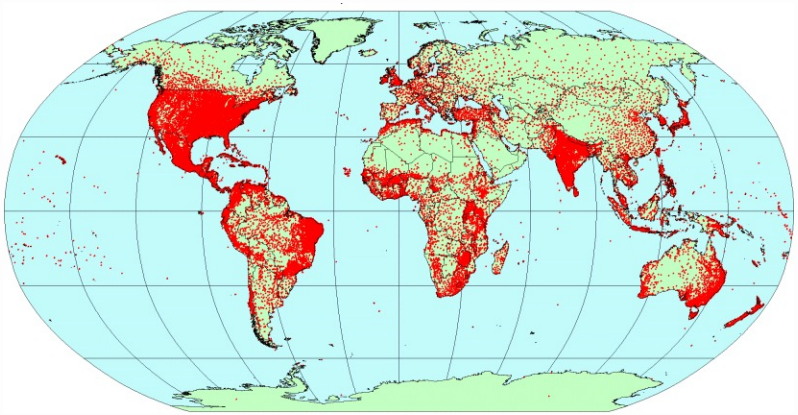


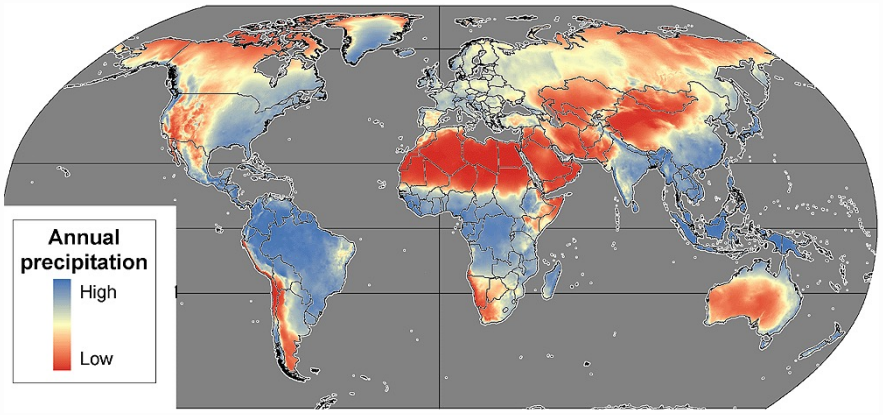
https://gisgeography.com/

https://desktop.arcgis.com/

Climate data

- Spatial interpolation of climate values from long-term weather station data
- E.g. Worldclim, CHELSA
- Temperature and precipitation, and derived bioclimatic variables





https://databasin.org/

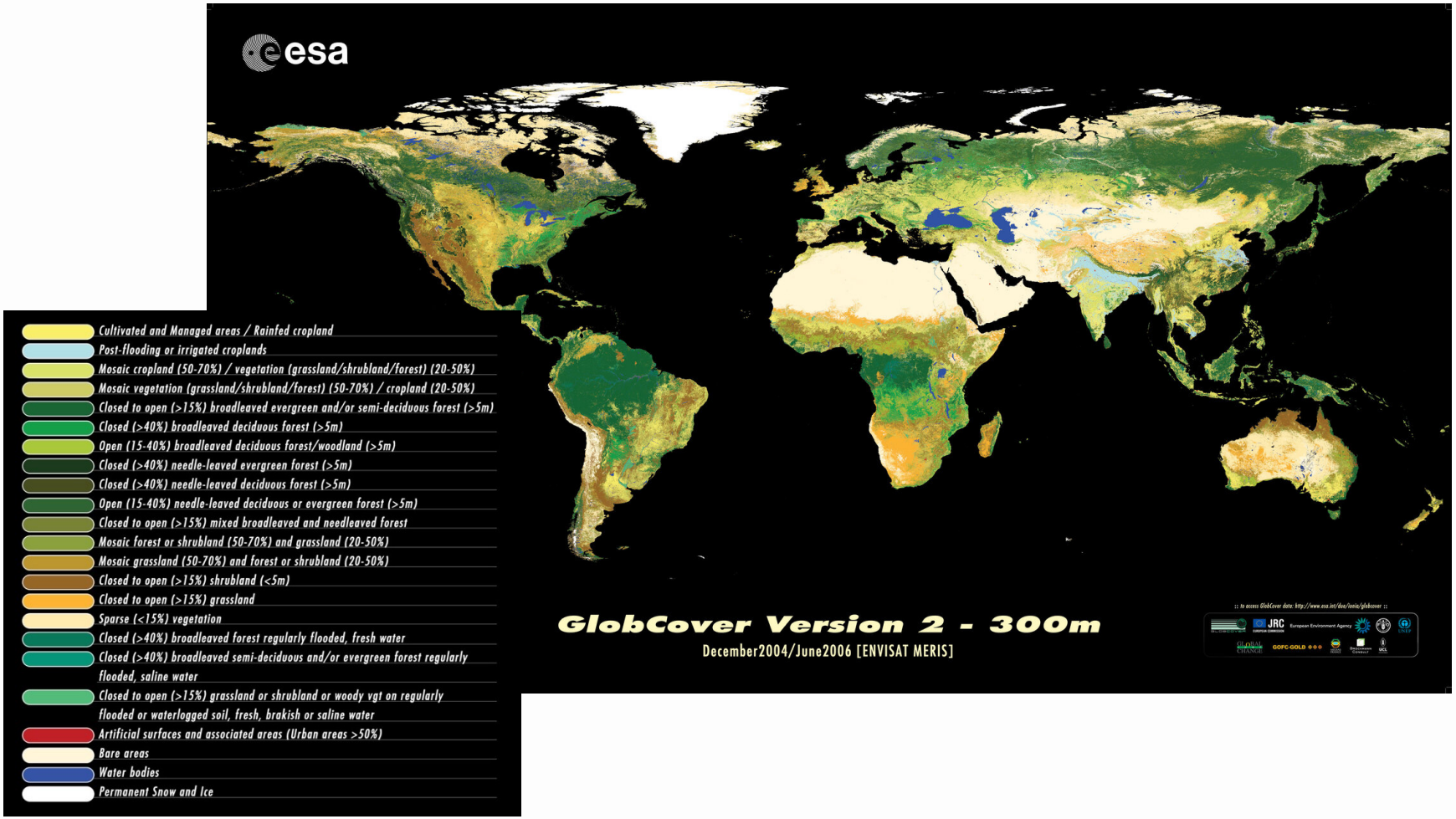
Bioclimatic variables

- https://www.worldclim.org/data/bioclim.html
- Derived from monthly temperature and rainfall values ➔ annual trends, seasonality and extremes

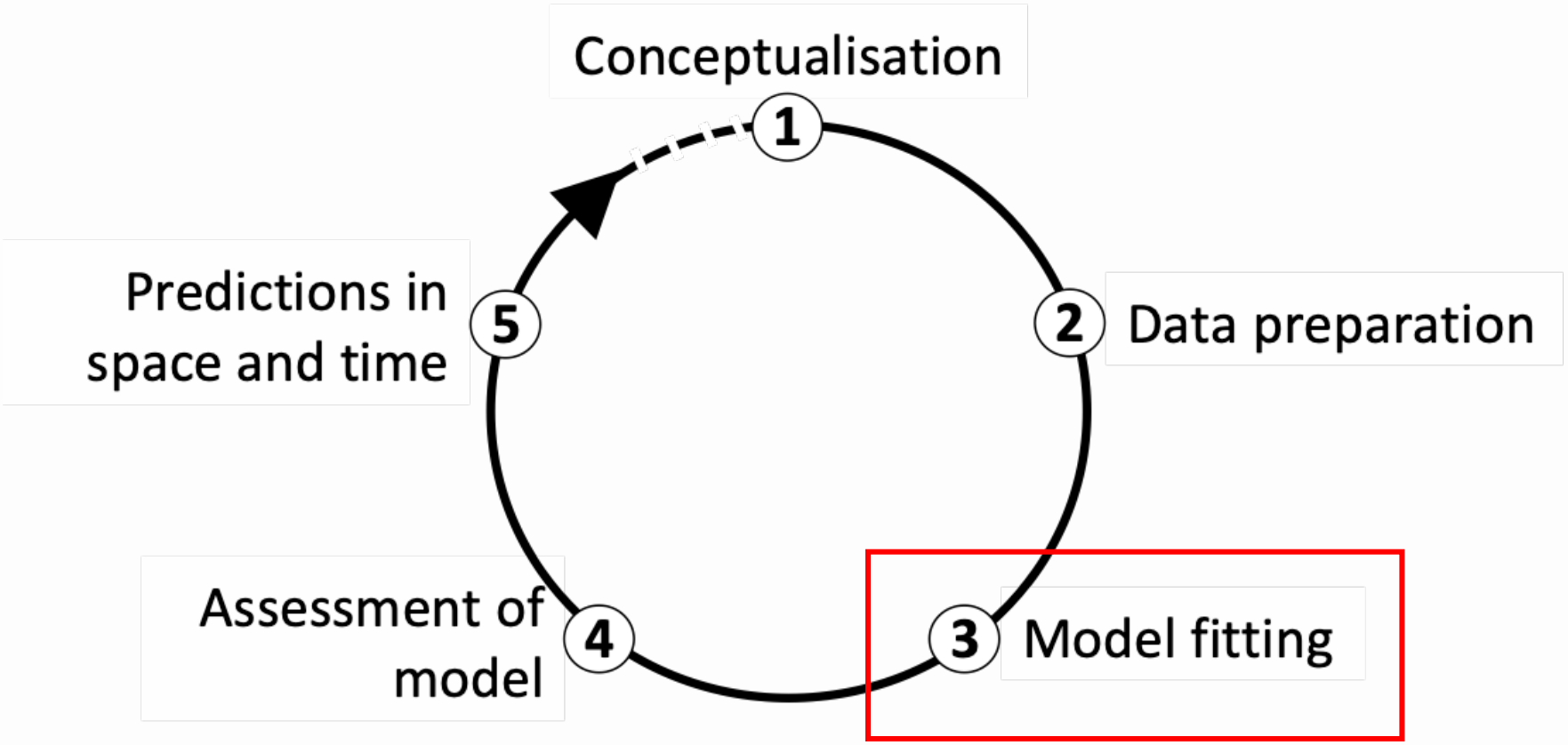
BIO1 = Annual Mean Temperature
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3 = Isothermality (BIO2/BIO7) (* 100)
BIO4 = Temperature Seasonality (standard deviation *100)
BIO5 = Max Temperature of Warmest Month
BIO6 = Min Temperature of Coldest Month
BIO7 = Temperature Annual Range (BIO5-BIO6)
BIO8 = Mean Temperature of Wettest Quarter
BIO9 = Mean Temperature of Driest Quarter
BIO10 = Mean Temperature of Warmest Quarter
BIO11 = Mean Temperature of Coldest Quarter
BIO12 = Annual Precipitation
BIO13 = Precipitation of Wettest Month
BIO14 = Precipitation of Driest Month
BIO15 = Precipitation Seasonality (Coefficient of Variation)
BIO16 = Precipitation of Wettest Quarter
BIO17 = Precipitation of Driest Quarter
BIO18 = Precipitation of Warmest Quarter
BIO19 = Precipitation of Coldest Quarter

Land cover / land use data

- Global land cover maps derived from remote sensing



SDMs – model building steps



SDM algorithms

Many different algorithms available for SDMs:

- **Profile methods** only consider species presences; use simple statistical techniques such as environmental distance to known sites
  - e.g. BIOCLIM, DOMAIN, Mahalonobis distance
- **Regression-based** techniques and **machine-learning** algorithms use presence and absence (or background) data to contrast used and unused sites
  - **Regression**: e.g. generalised linear model (GLM), generalised additive model (GAM), multivariate adapative regression splines (MARS), ...
  - **Machine-learning**: e.g. classification and regression tree (CART), artificial neural network (ANN), generalised boosted model/boosted regression trees (GBM/BRT), random forest (RF), maximum entropy (Maxent), genetic algorithms, ...
  - **Most of these can also be used for other response types, e.g. abundance, richness etc.**

SDM algorithms

- There is no single best approach for SDMs. (I have no favourite)
- Model choice should be guided by model purpose, available data, scale, ...
- More complex models tend to better fit current species-environment relationship. Yet, it is highly debated whether more complex models make better **predictions under global change**.
- For global change analyses, the IUCN recommends to use **at least three algorithms that are as independent as possible**.

Araujo & Rahbek (2006) Science 313: 1396-1397.
iUCN Standards and Petitions Committee (2019) http://www.iucnredlist.org/documents/RedListGuidelines.pdf

Generalised linear models (GLMs)

- Parametric regression method based on maximum likelihood estimation
- Allow error distributions different from Normal distribution
- The linear predictor is related to the response variable by a link function
- For presence-absence data we typically use the logit link:
  - The link function is used to transform the response to normality

$$E(Y|X) = \pi(X) = \frac{e^{\beta X+\epsilon}}{1 + e^{\beta X+\epsilon}}$$

  - The logit *g(X)* is linear in its parameters

$$g(X) = ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta X + \epsilon$$

Fitting GLMs in R

- Example with Ring Ouzel in UK

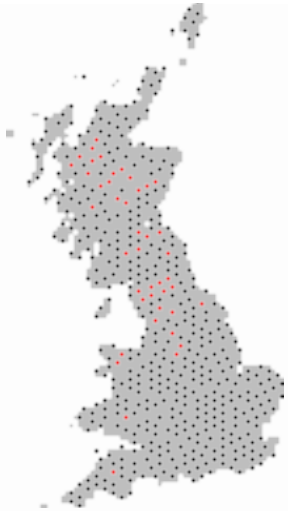
```
m1 <- glm(Turdus_torquatus ~ bio11, family="binomial", data= sp_dat)
```

Response variable

Predictor variable

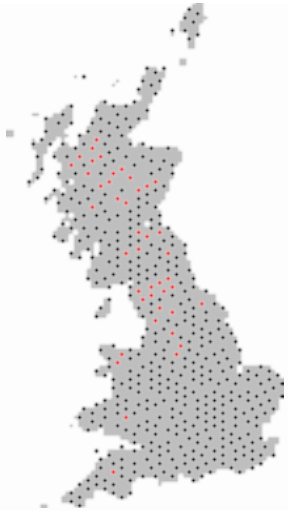
Link function

Data



Fitting GLMs in R

- Example with Ring Ouzel in UK

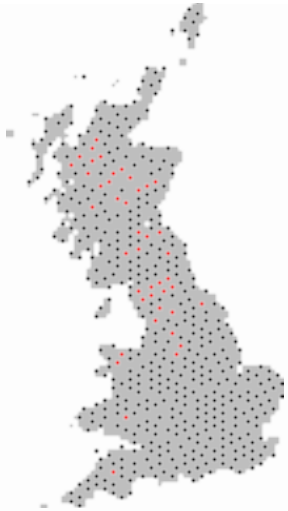


```
summary(m1)
```

```
Call:
glm(formula = Turdus_torquatus ~ bio11, family = "binomial",
    data = sp_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8183  -0.2813  -0.1803  -0.1157   3.1911

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.60045    0.44867   3.567 0.000361 ***
bio11       -0.16071    0.01978  -8.126 4.44e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 294.43  on 506  degrees of freedom
Residual deviance: 176.39  on 505  degrees of freedom
AIC: 180.39

Number of Fisher Scoring iterations: 7
```

Fitting GLMs in R

- Example with Ring Ouzel in UK

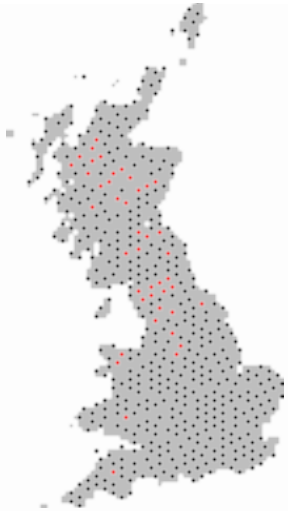


```
summary(m1)
```

```
Call:
glm(formula = Turdus_torquatus ~ bio11, family = "binomial",
    data = sp_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8183  -0.2813  -0.1803  -0.1157   3.1911

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.60045    0.44867   3.567 0.000361 ***
bio11       -0.16071    0.01978  -8.126 4.44e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 294.43  on 506  degrees of freedom
Residual deviance: 176.39  on 505  degrees of freedom
AIC: 180.39

Number of Fisher Scoring iterations: 7
```

Fitting GLMs in R

- Example with Ring Ouzel in UK



```
summary(m1)
```

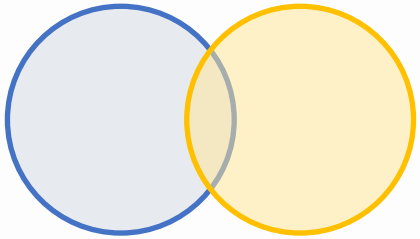
```
Call:
glm(formula = Turdus_torquatus ~ bio11, family = "binomial",
    data = sp_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8183  -0.2813  -0.1803  -0.1157   3.1911

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.60045    0.44867   3.567 0.000361 ***
bio11       -0.16071    0.01978  -8.126 4.44e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 294.43  on 506  degrees of freedom
Residual deviance: 176.39  on 505  degrees of freedom
AIC: 180.39

Number of Fisher Scoring iterations: 7
```

Consideration for model fitting

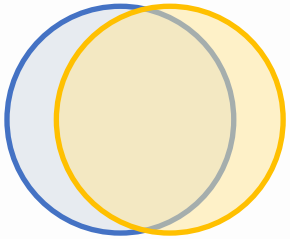
- How to deal with multicollinearity in the environmental data?
- How many variables should be included in the model (without overfitting) and how should we select these?
- Which model settings should be used?
- When multiple model algorithms or candidate models are fitted, how to select the final model or average the models?
- Do we need to test or correct for non-independence in the data (spatial or temporal autocorrelation, nested data)?
- Do we want to threshold the predictions, and which threshold should be used?

What is multicollinearity

- Predictor variables are not independent from each other



Weak correlation



Strong correlation

- Makes it difficult to identify the more meaningful predictor
- Lead to inflated erros
- Most problematic when extrapolating

How to deal with multicollinearity

- *select07* method as simple and effective way to reduce multicollinearity – other approaches here:

Research

EDITOR'S CHOICE

Ecography 36: 027–046, 2013
doi: 10.1111/j.1600-0587.2012.07348.x

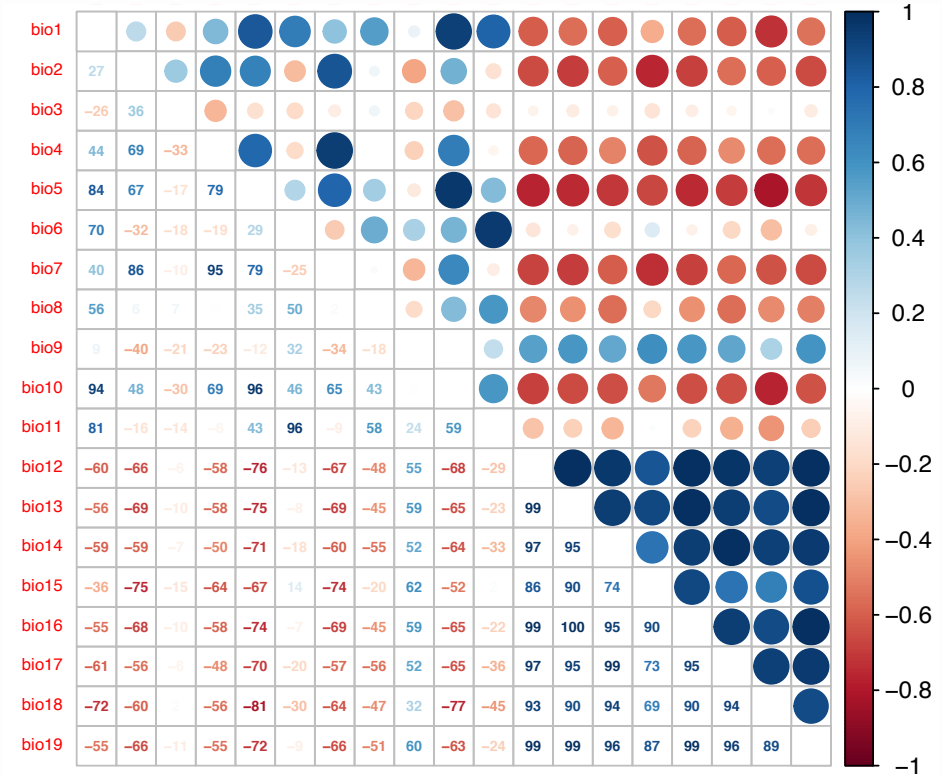
Subject Editor: Marti Jane Anderson. Accepted 24 February 2012

Collinearity: a review of methods to deal with it and a simulation study evaluating their performance

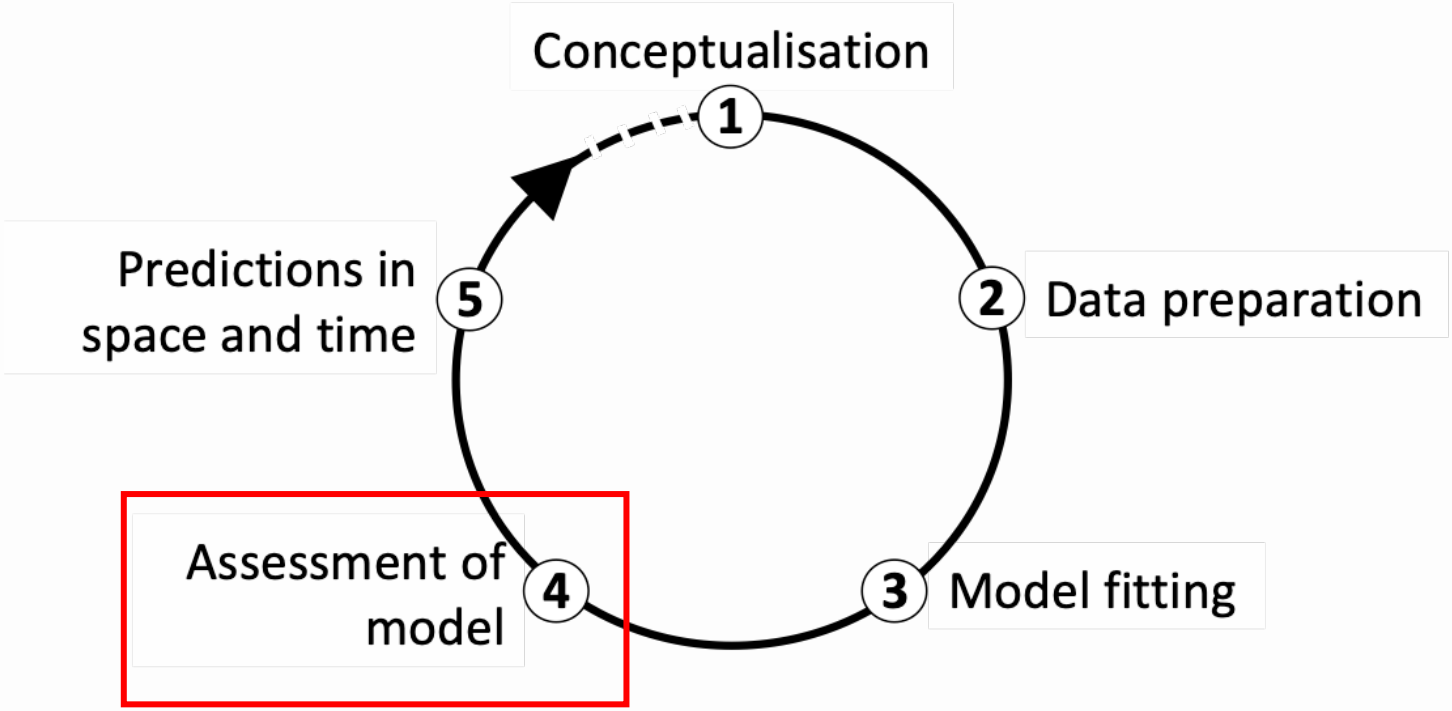
Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell and Sven Lautenbach

How to deal with multicollinearity

- *select07:* inspect pairwise correlations between the 19 bioclimatic variables, determine univariate importance of all predictors, from pairs of highly correlated variables remove the less important one



SDM – model building steps

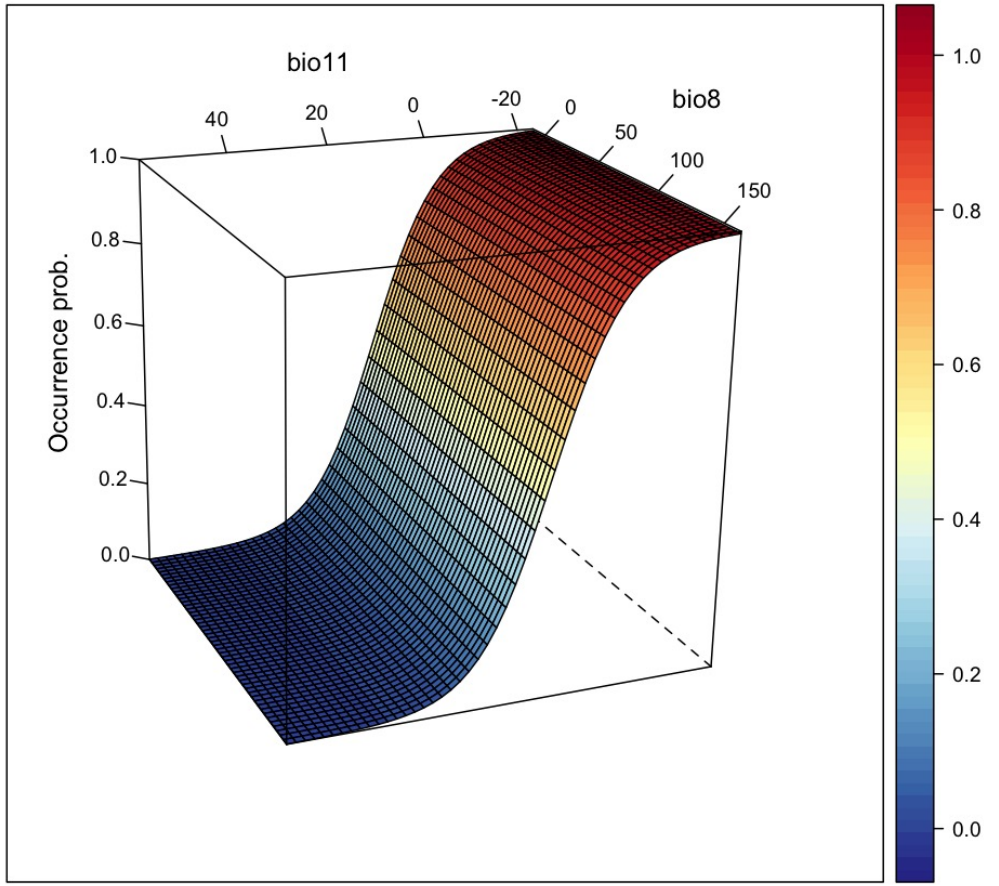


- How does the species-environment relationship look like?
- How well is my model supported by data?
- How well does my model predict to independent data?

SDM – visualise response

Response surfaces

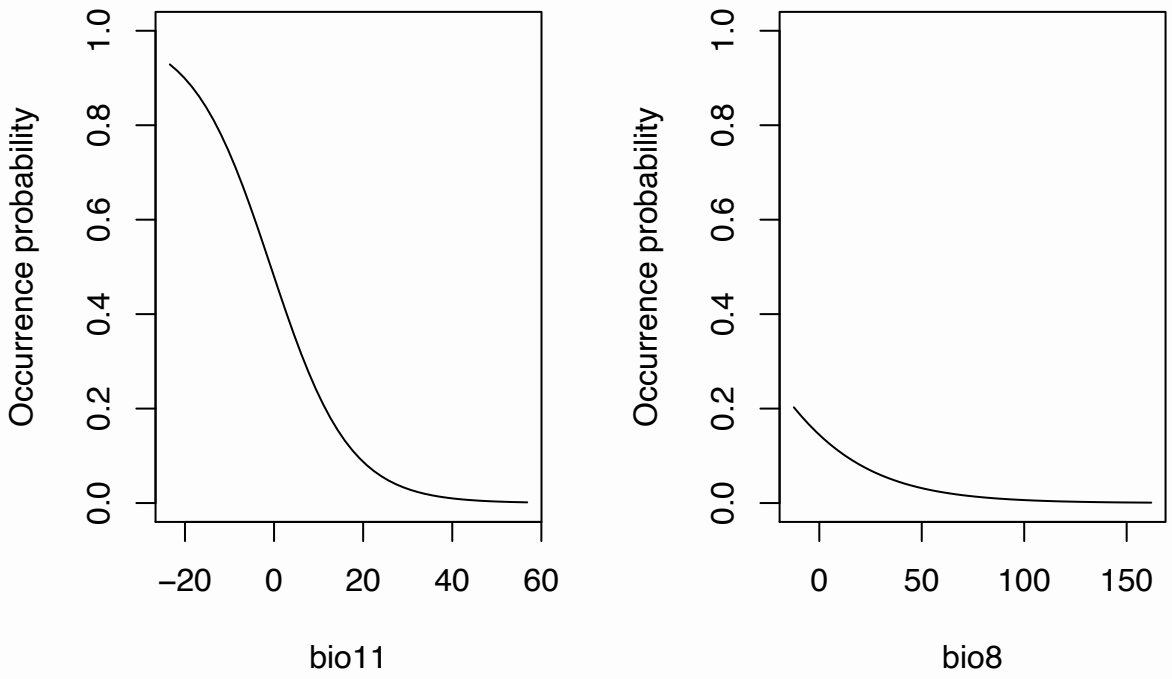
- Purpose: visualise model predictions along two environmental gradients



SDM – visualise response

Partial response plots

- Purpose: abstract model behaviour to 2D
- Approach: plot response curve for each predictor separately while keeping the other predictors constant at their mean



SDM – visualise response

Partial response plots

- Purpose: abstract model behaviour to 2D
- Approach: plot response curve for each predictor separately while keeping the other predictors at their mean

```r
for (i in 1:2){

  # Create new data frame with predictor variables
  xyz <- data.frame(
    # Sequence for one predictor:
    seq(min(sp_dat[,my_preds[i]]),max(sp_dat[,my_preds[i]]),length=50),

    # Calculate the mean of the other predictor:
    mean(sp_dat[,my_preds[-i]])
    )

  # Make sure the new data frame contains the correct predictor names
  names(xyz) <- c(my_preds[i], my_preds[-i])

  # Make prediction to new data
  xyz$z <- predict(m1, newdata=xyz, type='response')

  # Plot the response curve
  plot(xyz[,my_preds[i]],xyz$z,type='l', ylim=c(0,1), xlab=my_preds[i], ylab='Occurrence probability')
}
```

Make „artificial“ new data frame with one (equal-spaced) environmental gradient and mean of other variable(s)

Make prediction to the new data frame

SDM – visualise response

Partial response plots

- Purpose: abstract model behaviour to 2D
- Approach: plot response curve for each predictor separately while keeping the other predictors at their mean

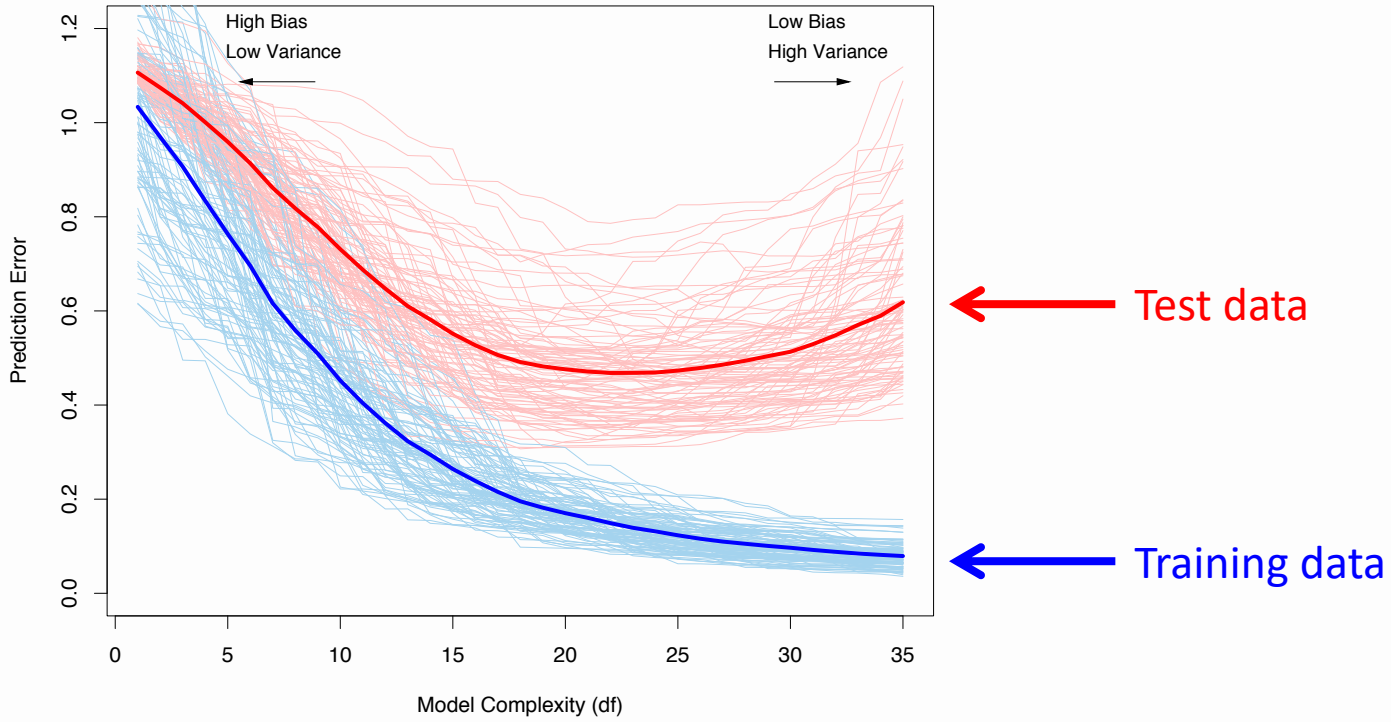
```
# Plot the partial responses
partial_response(m1, predictors = sp_dat[,my_preds])
```

Simplified function to draw partial response plots

SDM – assessing model performance

The Bias-Variance trade-off:

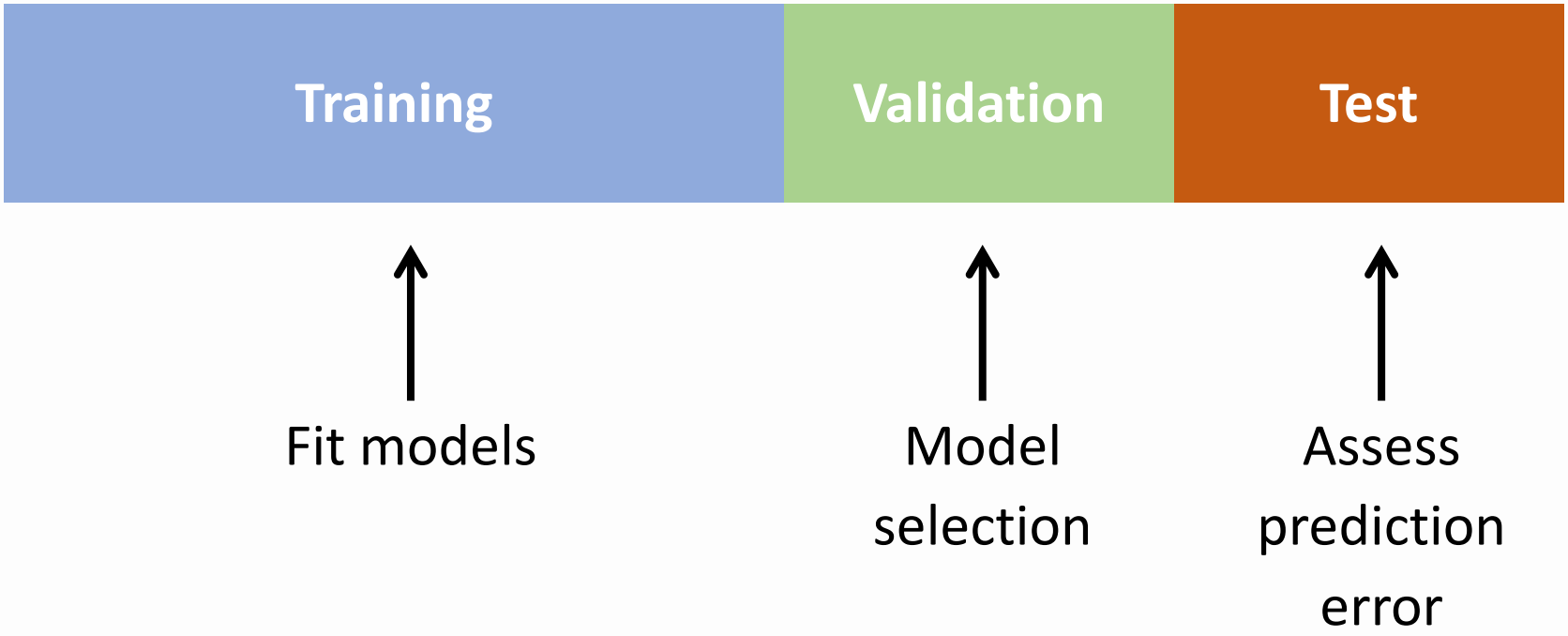
- Bias is caused by simplifying model assumptions
- Variance is caused by fluctuations in the data



Hastie et al. (2009) Elements of statistical learning. Springer

SDM – assessing model performance

Ideally, predictive performance should be validated on independent test data.



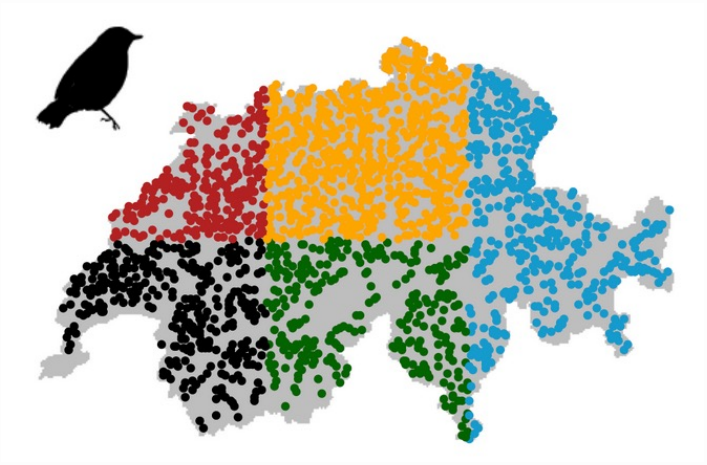
Hastie et al. (2009) Elements of statistical learning. Springer

SDM – assessing model performance

Ideally, predictive performance should be validated on independent test data (= external validation).

Typical approaches when independent test data missing:

- **Internal validation**: only on training data
- **Split-sample**, e.g. 70% training - 30% test
- **k-fold cross-validation**: (k-1)/k proportion training – 1/k proportion test, repeat k times
- **k-fold block cross-validation**:
  - Spatial blocks, or
  - Environmental blocks



SDM – assessing model performance

Goodness-of-fit: typically derived from the log-likelihood

- Example: explained deviance D^2

Log-likelihood:

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n} (y_i \times ln[\pi(x_i)] + (1 - y_i) \times ln[1 - \pi(x_i)])$$

Deviance:

$$D = -2 \times L$$

Explained deviance:

$$D^2 = 1 - \frac{D(model)}{D(Null.model)}$$

SDM – assessing model performance

Discrimination: in how far can model distinguish between presences and absences?

- Threshold-dependent measures:
  - Sensitivity (true positive rate)
  - Specificity (true negative rate)
  - TSS (true skill statistic)
  - Kappa

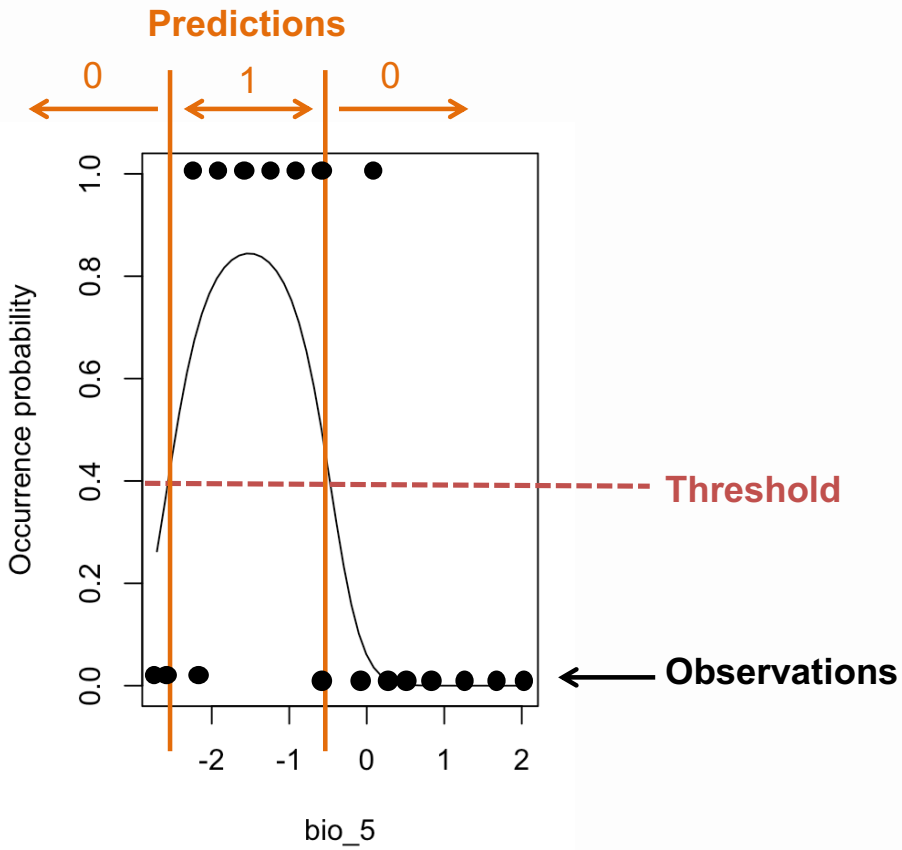
- Threshold-independent measures:
  - AUC (area under the receiver operating characteristic curve)

SDM - performance measures

Threshold-dependent measures:

Derived from confusion matrix

Predictions \ Observations	1	0
1	a	b
0	c	d

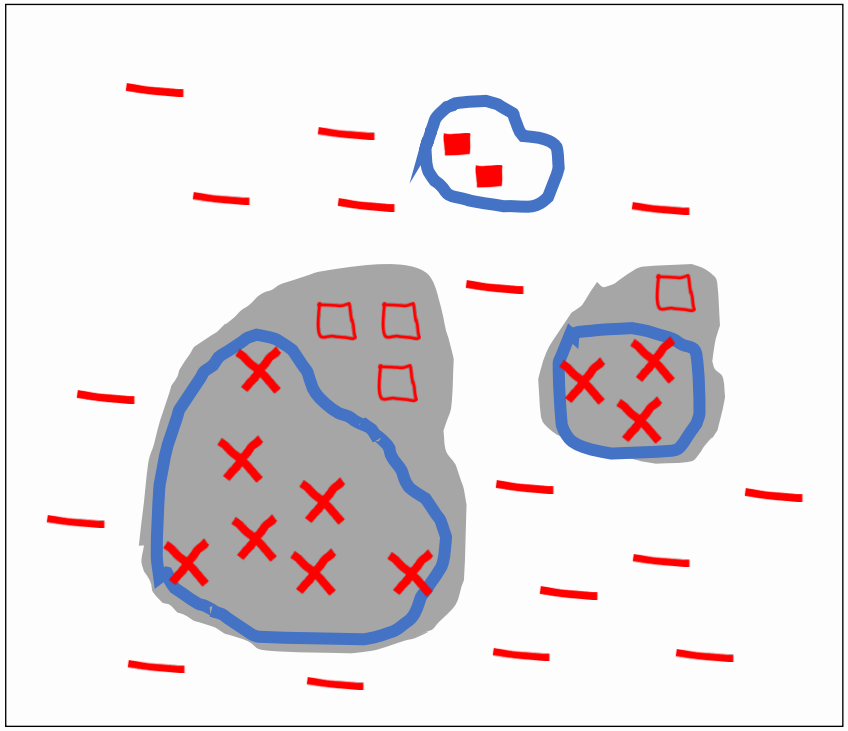


SDM - performance measures

Threshold-dependent measures:

Derived from confusion matrix

Predictions \ Observations	1	0
1	a	b
0	c	d



Actual distribution

SDM prediction

True negative (d)

True positive (a)

False positive (b)

False negative (c)

SDM - performance measures

Threshold-dependent measures:

Derived from confusion matrix

		Observations	
		1	0
Predictions	1	a	b
	0	c	d

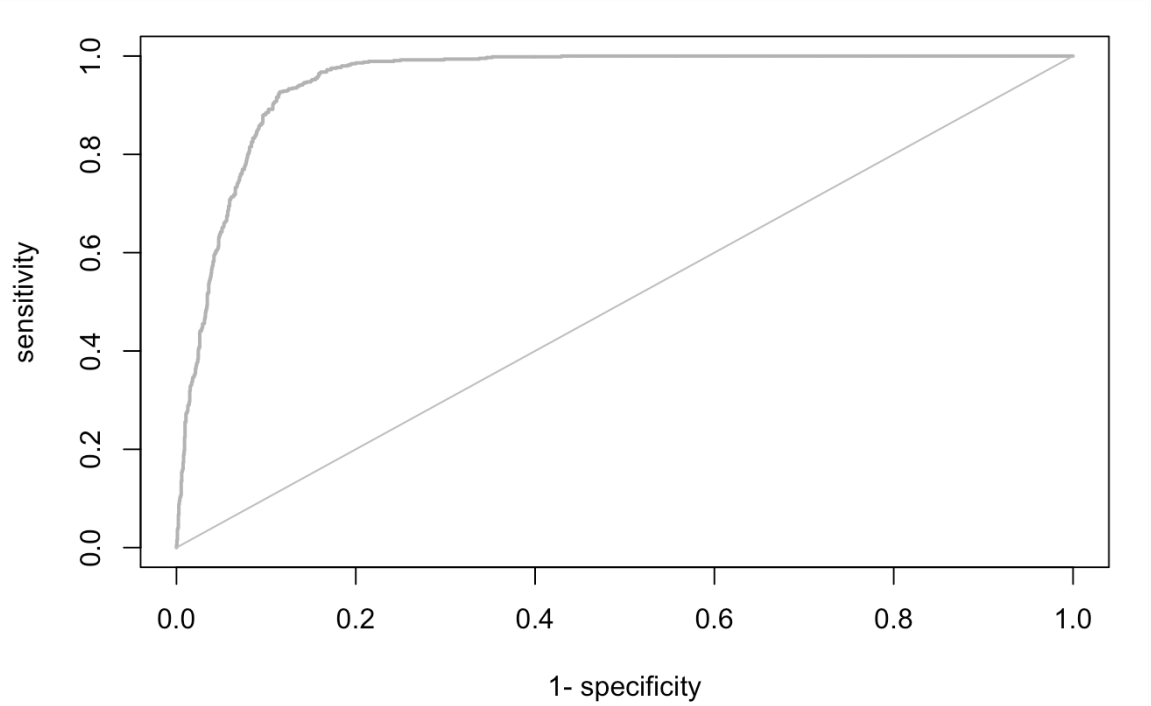
Measure	Formula	Fair prediction?
Overall accuracy	$\frac{a+d}{n}$	
Sensitivity	$\frac{a}{a+c}$	Sens > 0.75
Specificity	$\frac{d}{b+d}$	Spec > 0.75
Kappa statistic	$\frac{\left(\frac{a+d}{n}\right) - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}{1 - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}$	Kappa > 0.4
TSS	sensitivity + specificity – 1	TSS > 0.5

Allouche et al. (2006) J Applied Ecology 43: 1223-1232.

SDM - performance measures

Threshold-independent measures:

- AUC (area under the receiver operating characteristic curve)
- The probability to predict a higher probability of occurrence for a true presence than for a true absence

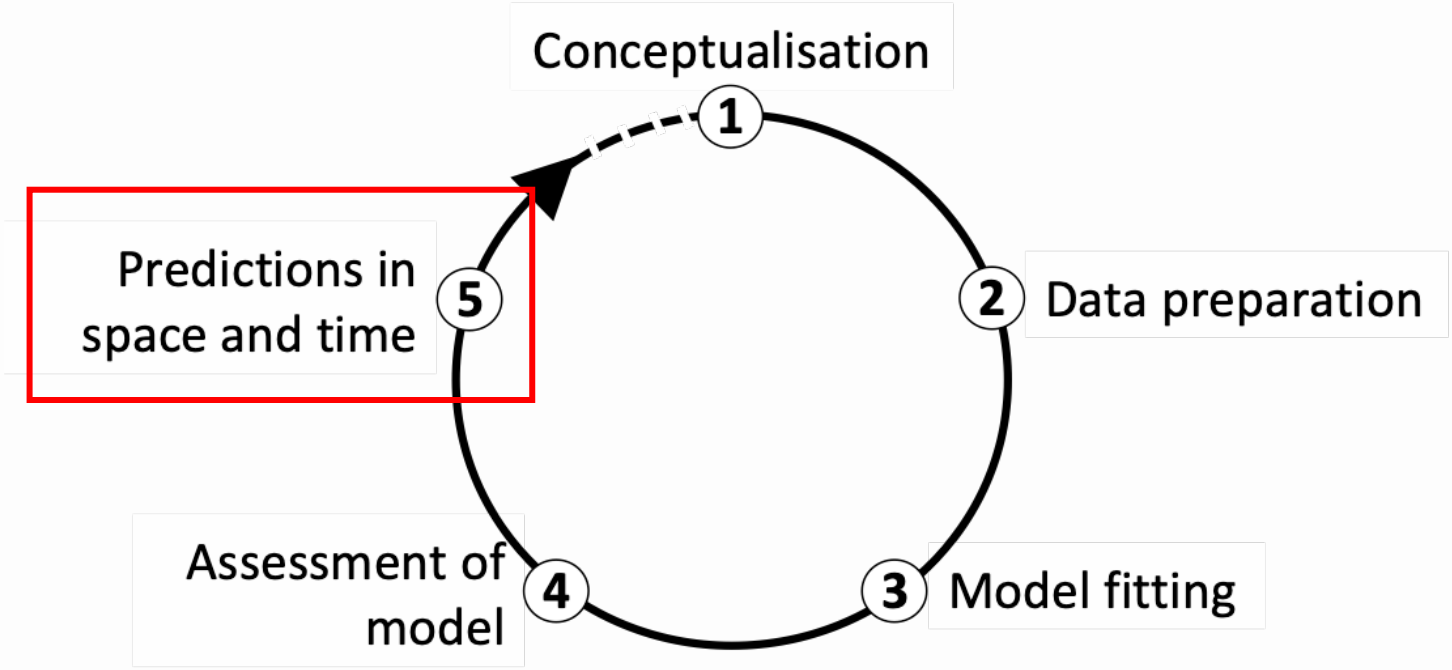


Fair prediction?

AUC > 0.7

SDM – model building steps

- What is the potential distribution of the species?
- How certain is this prediction?

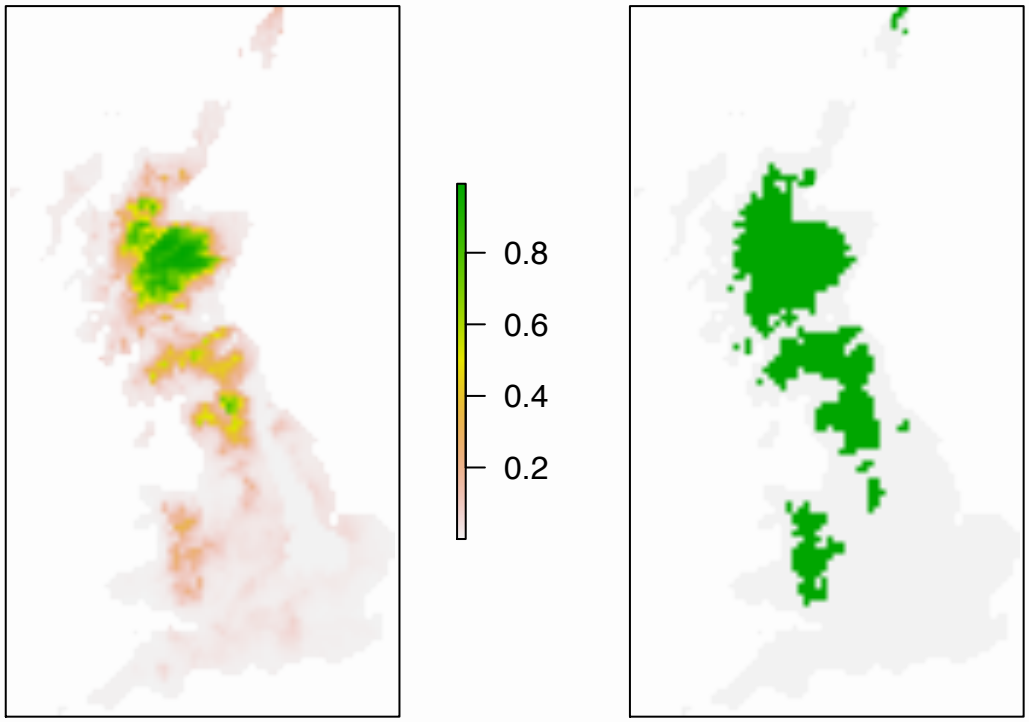


SDM predictions

Spatiotemporal predictions:

```
predict(m1, newdata= bio_fut_df, type="response")
```

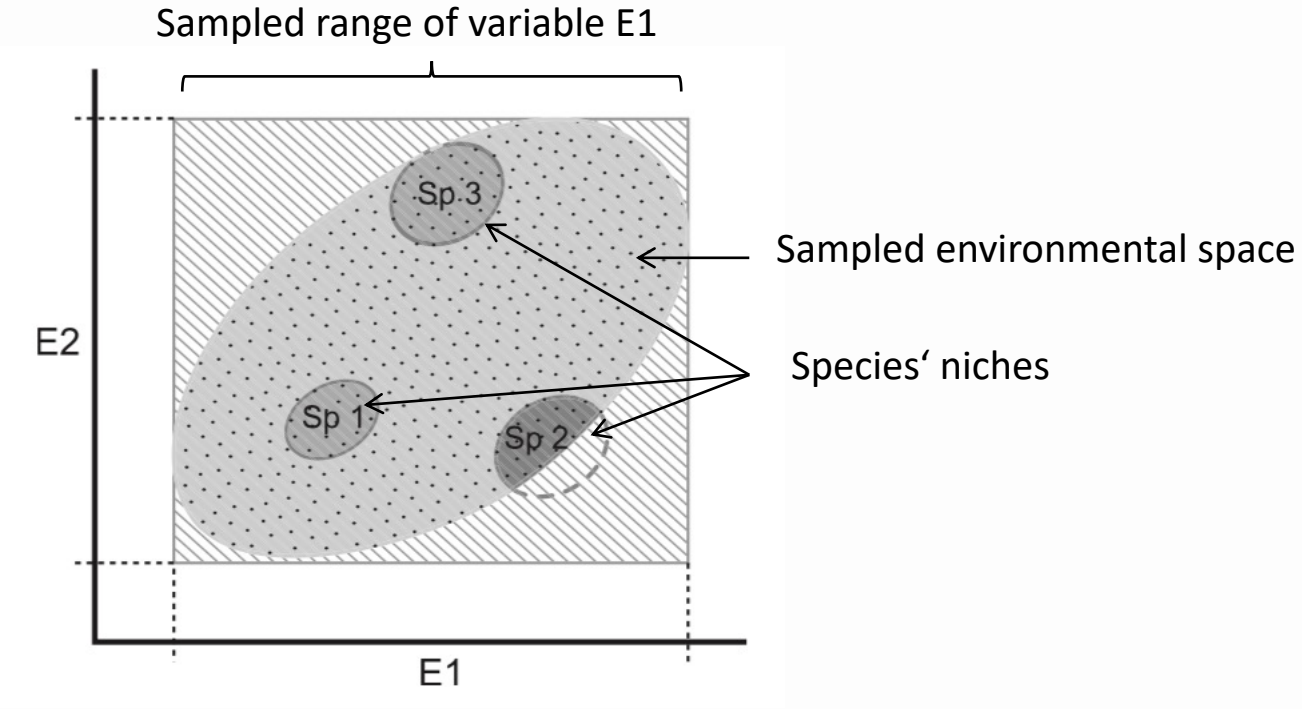
Predicted occurrence probabilities → Threshold → Binary predictions



SDM predictions

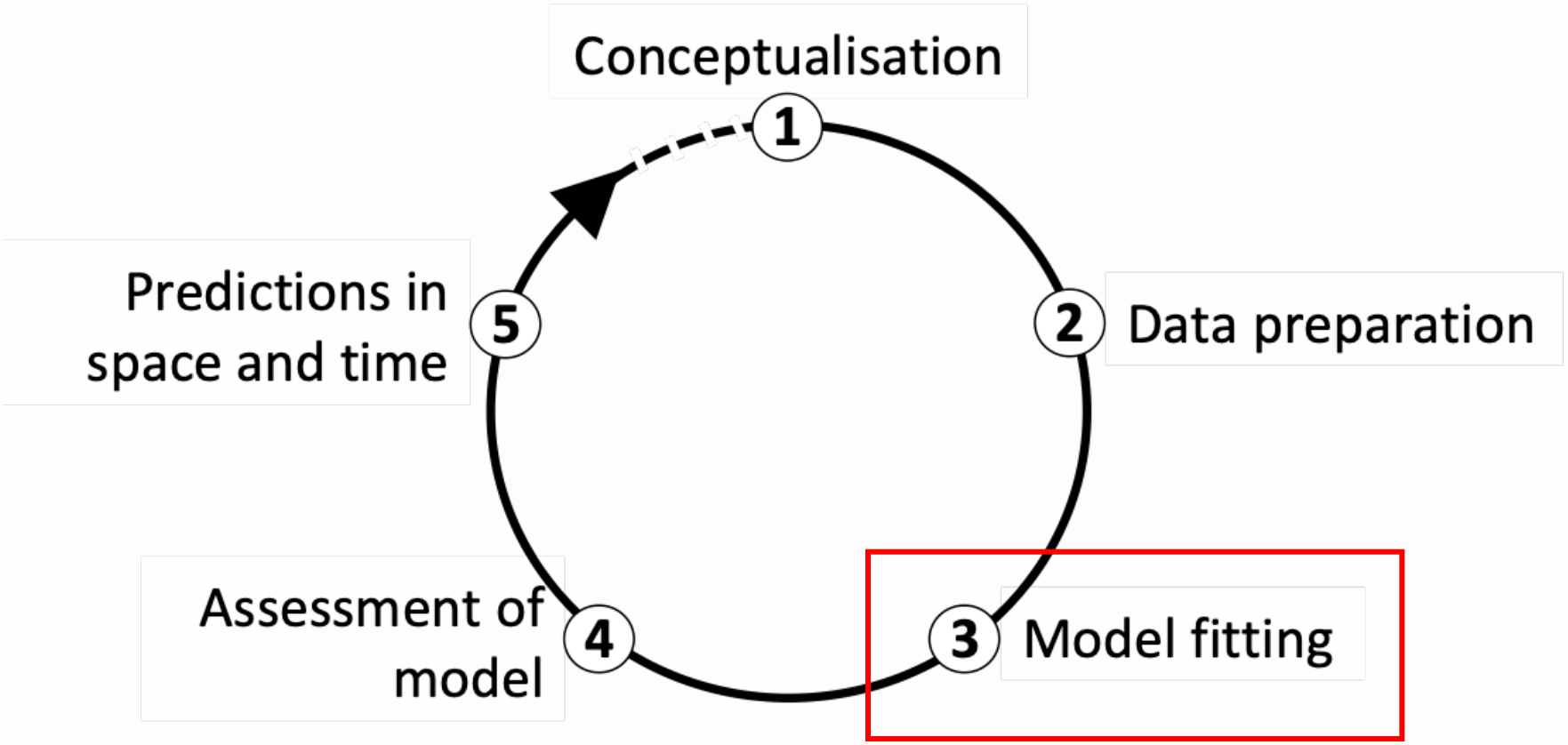
Spatiotemporal predictions:

- Interpolation: within sampled environmental space
- Extrapolation: beyond sampled environmental space



Zurell et al. (2012) Diversity and Distributions 18: 628-634.

SDMs – model building steps

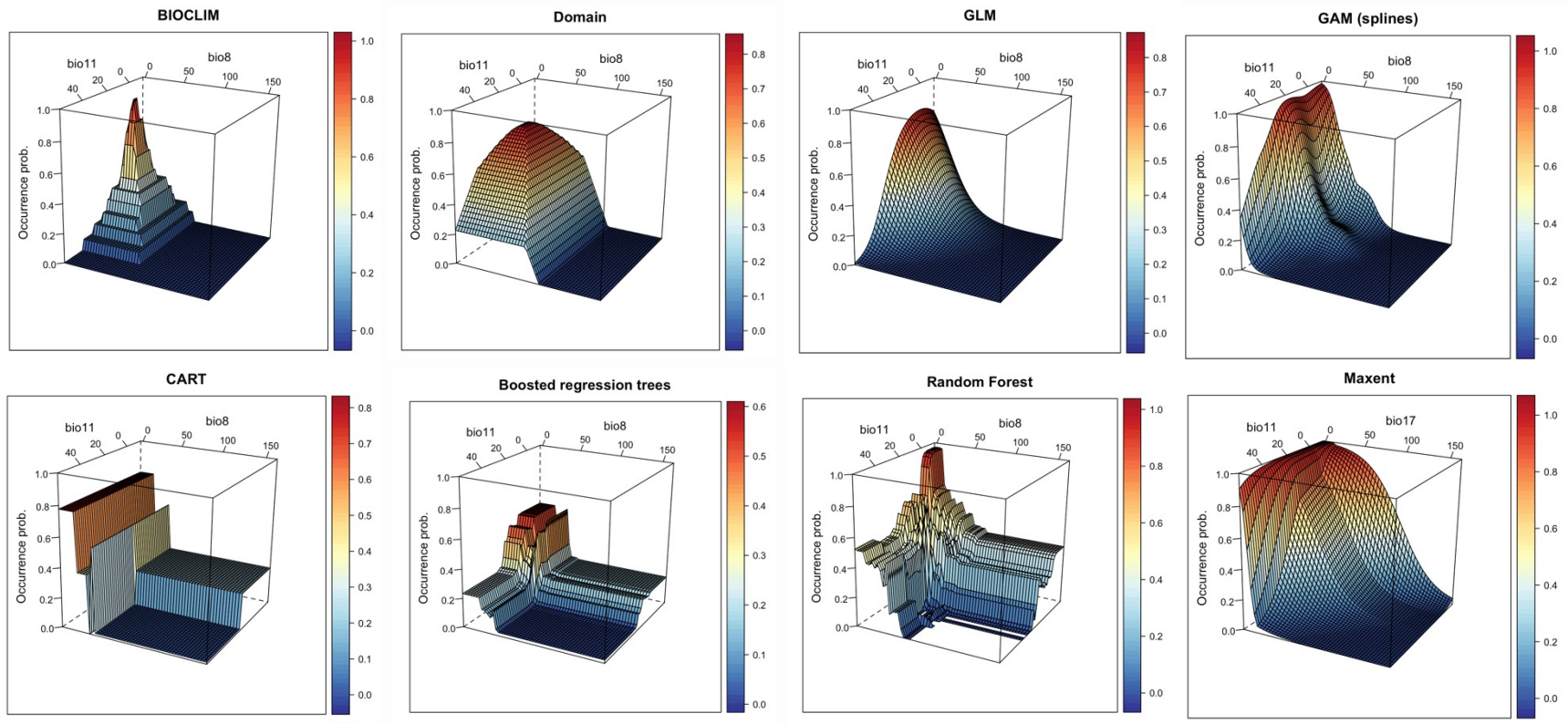


SDM algorithms

Many different algorithms available for SDMs:

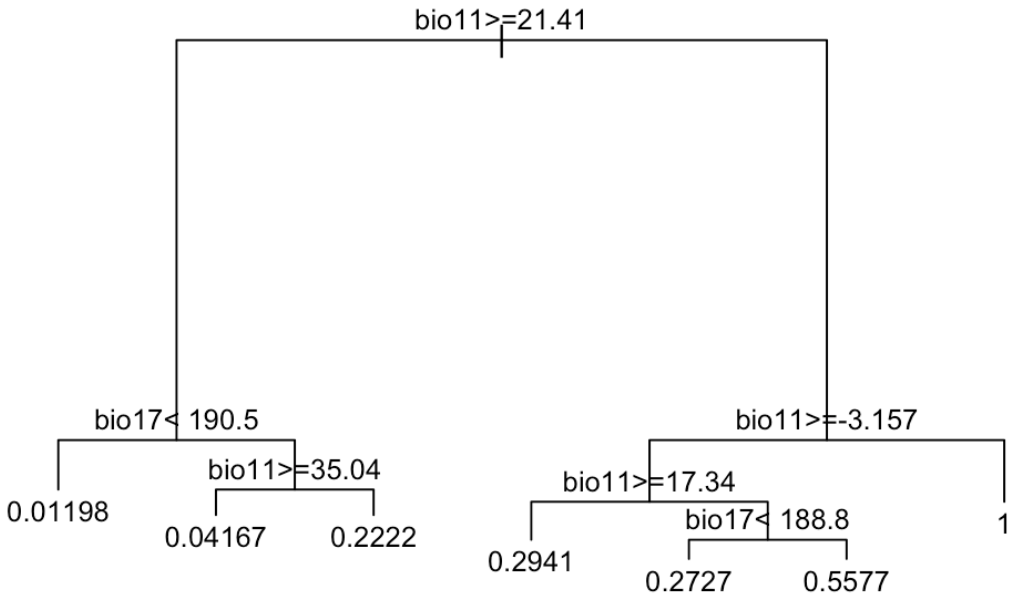
- **Profile methods** only consider species presences; use simple statistical techniques such as environmental distance to known sites
  - e.g. BIOCLIM, DOMAIN, Mahalonobis distance
- **Regression-based** techniques and **machine-learning** algorithms use presence and absence (or background) data to contrast used and unused sites
  - **Regression**: e.g. generalised linear model (GLM), generalised additive model (GAM), multivariate adapative regression splines (MARS), ...
  - **Machine-learning**: e.g. classification and regression tree (CART), artificial neural network (ANN), generalised boosted model/boosted regression trees (GBM/BRT), random forest (RF), maximum entropy (Maxent), genetic algorithms, ...
  - **Most of these can also be used for other response types, e.g. abundance, richness etc.**

SDM algorithms



Machine-learning: CART

- Classification and regression trees (CARTs)
- Recursive partitioning method to divide the data into homogeneous subgroups
- Find splits (nodes) that best separate the observations
- Interactions between variables fitted automatically



Machine-learning: CART extensions

- CARTs sensitive to noise: typically show low bias and high variance
- One solution: model averaging

  - Bagging = bootstrap aggregation: fit many CARTs to bootstrapped samples of data and average results

    ➔ Random Forest

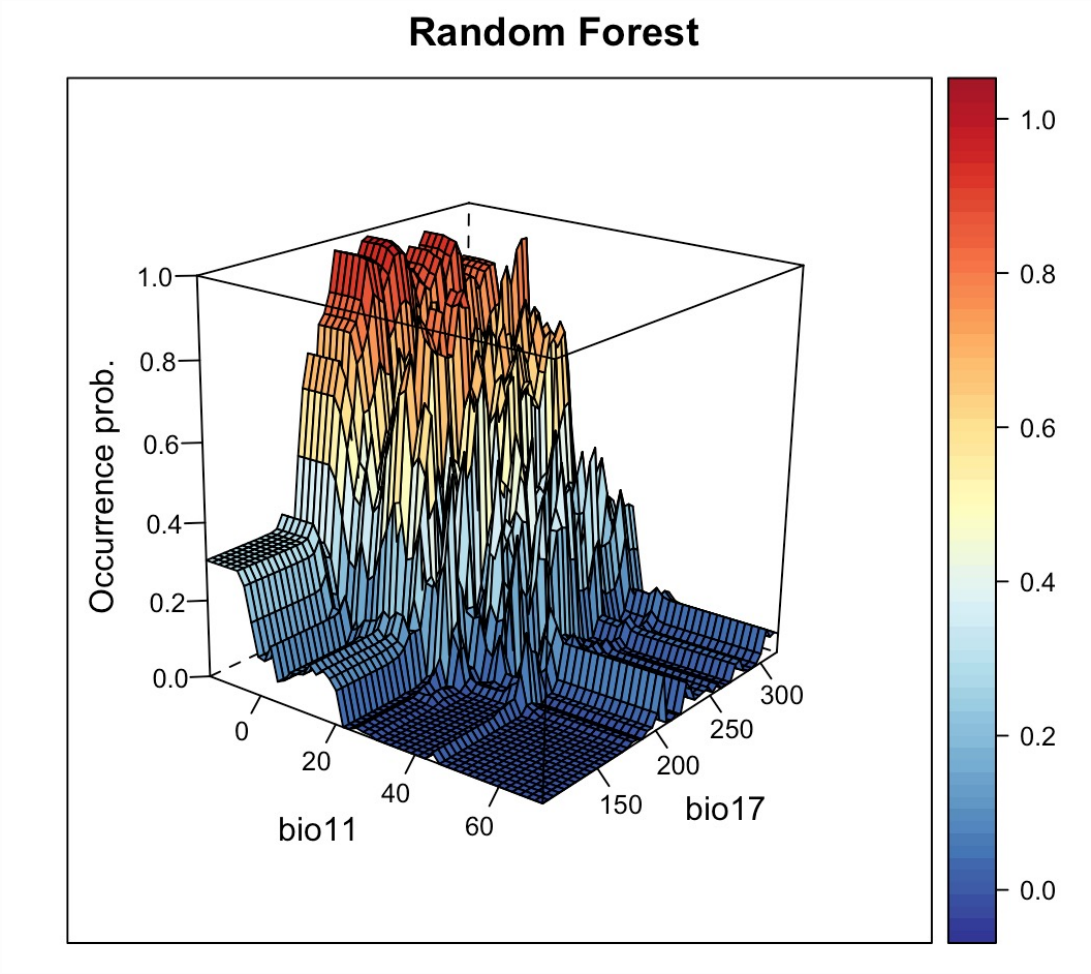
  - Boosting: fit relatively simple CARTs sequentially in adaptive way = each model depends on the previous ones

    ➔ Boosted regression trees

Hastie et al. (2009) Elements of statistical learning. Springer

Machine-learning: random forest

- R package „randomForest“



Machine-learning: random forest

- R package „randomForest“

Data frame of predictors

Response

```
m_rf <- randomForest( x=sp_train[,my_preds], y=sp_train$Turdus_torquatus,
    ntree=1000, importance =T)
```

How many trees to grow?

Should variable importance be computed?

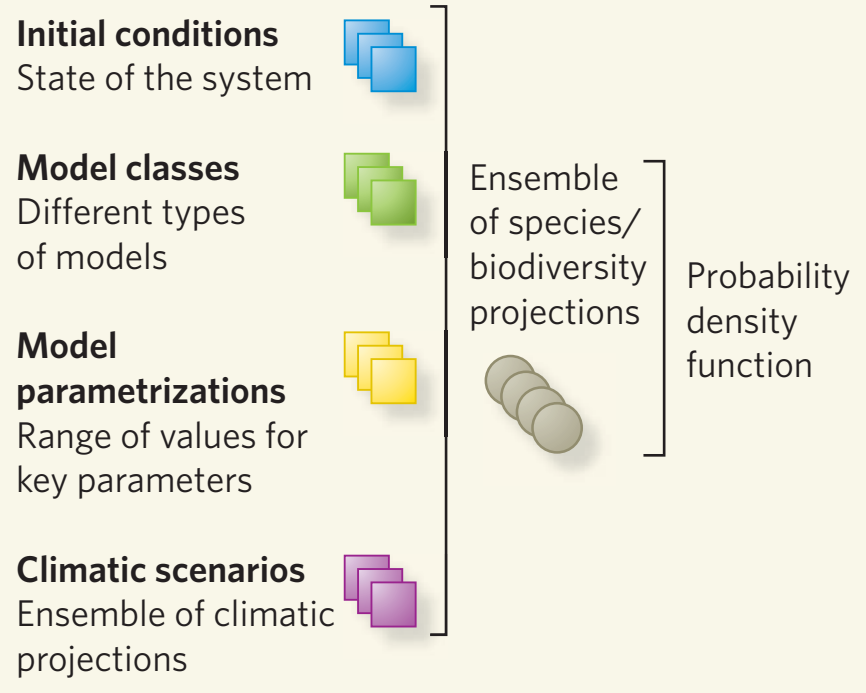
Response type: probabilities

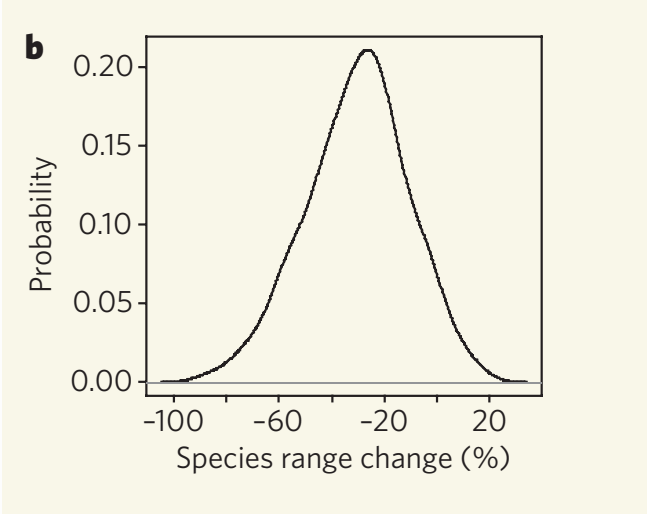
```
predict(m_rf, xyz, type='response')
```

Data frame with predictor variables

SDM ensembles

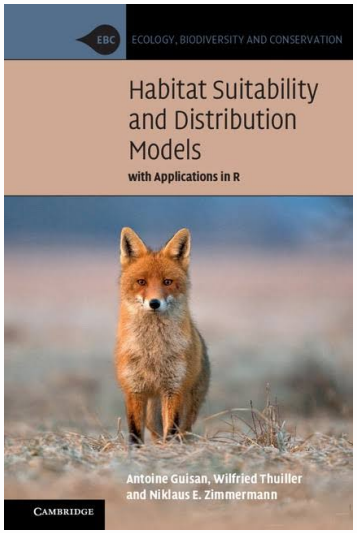
- Ensembles of forecasts are produced by making multiple simulations across more than one set of initial conditions (data), model classes, model parameterisations, and boundary conditions (scenarios)



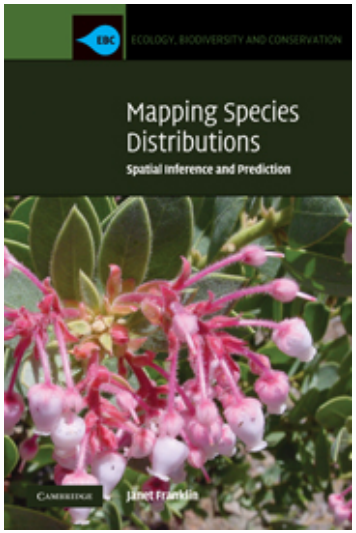


Araujo & New (2007) Trends in Ecology & Evolution 22: 42-47.
Thuiller (2007) Nature 448: 550-552.

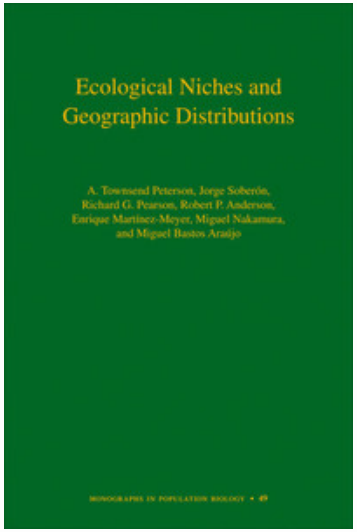
Literature



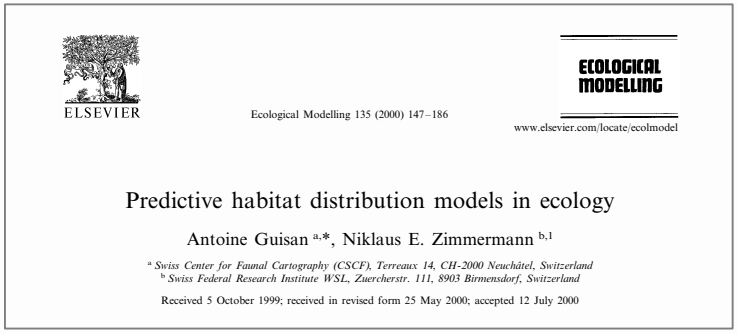
DOI: 10.1017/9781139028271



DOI: 10.1017/CBO9780511810602



DOI: 10.23943/princeton/9780691136868.001.0001


ELSEVIER

Ecological Modelling 135 (2000) 147–186

ECOLOGICAL MODELLING

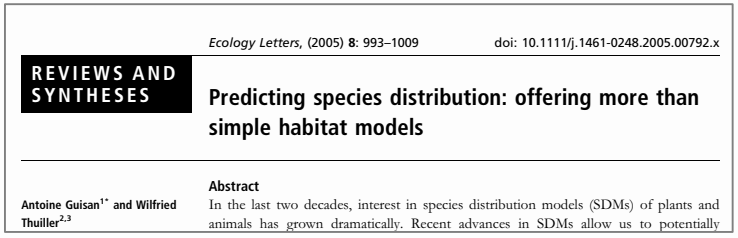
www.elsevier.com/locate/ecolmodel

Predictive habitat distribution models in ecology

Antoine Guisan [a,*], Niklaus E. Zimmermann [b,1]

[a] Swiss Center for Faunal Cartography (CSCF), Terreaux 14, CH-2000 Neuchâtel, Switzerland
[b] Swiss Federal Research Institute WSL, Zuercherstr. 111, 8903 Birmensdorf, Switzerland

Received 5 October 1999; received in revised form 25 May 2000; accepted 12 July 2000


Ecology Letters, (2005) **8**: 993–1009 doi: 10.1111/j.1461-0248.2005.00792.x

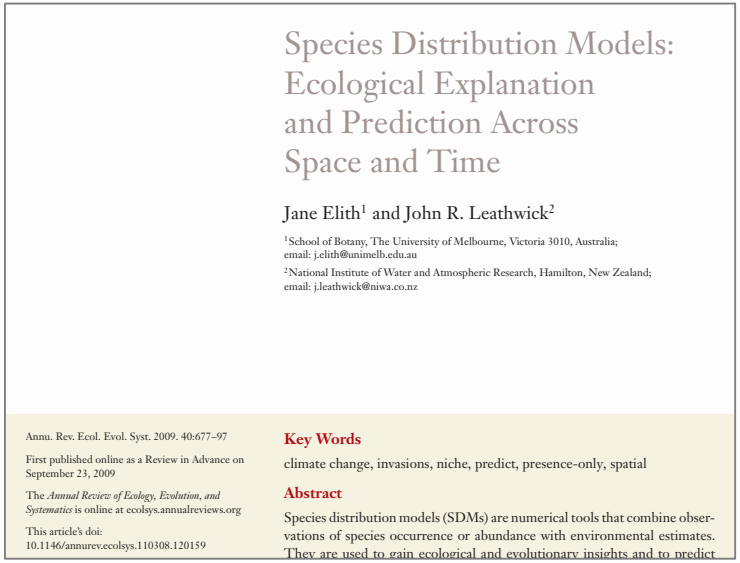
REVIEWS AND SYNTHESES

Predicting species distribution: offering more than simple habitat models

Antoine Guisan[1*] and Wilfried Thuiller[2,3]

Abstract

In the last two decades, interest in species distribution models (SDMs) of plants and animals has grown dramatically. Recent advances in SDMs allow us to potentially


Species Distribution Models: Ecological Explanation and Prediction Across Space and Time

Jane Elith[1] and John R. Leathwick[2]

[1]School of Botany, The University of Melbourne, Victoria 3010, Australia; email: j.elith@unimelb.edu.au

[2]National Institute of Water and Atmospheric Research, Hamilton, New Zealand; email: j.leathwick@niwa.co.nz

Annu. Rev. Ecol. Evol. Syst. 2009. 40:677–97

First published online as a Review in Advance on September 23, 2009

The *Annual Review of Ecology, Evolution, and Systematics* is online at ecolsys.annualreviews.org

This article's doi: 10.1146/annurev.ecolsys.110308.120159

Key Words

climate change, invasions, niche, predict, presence-only, spatial

Abstract

Species distribution models (SDMs) are numerical tools that combine observations of species occurrence or abundance with environmental estimates. They are used to gain ecological and evolutionary insights and to predict