Modélisation de niche écologique

Rodrigue Idohou, PhD

Modélisation de niche écologique

1. Cadre théorique: habitat, niche, distribution
2. Types et sources de données en modélisation
3. Algorithmes de modélisation
4. Evaluation de la performance des modèles
5. Etude de cas avec Maxent

1. Cadre théorique

Why model species ranges?

We need to know where species occur and why they occur where they do:

- we want to predict where a particular species occurs;
- we want to know more about organism-environment relationships.

Used in response to

- increasing rates of habitat, and species loss,
- incomplete (spatial and temporal) distribution info for a large number of taxa,
- existing distribution data collected in an ad hoc fashion.

Given the rate of species loss, it is unlikely that we will get the distribution data that we need in time if we rely on conventional survey techniques.

Atlases are an invaluable data source and cover very few taxa but they are very important for model development and calibration.

Distribution models have been used to predict

- species richness (Jetz & Rahbeck 2002)
- centres of endemism (Johnson, Hay & Rogers 1998),
- the occurrence of particular species assemblages (Neave, Norton & Nix 1996),
- the occurrence of individual species (Gibson *et al*. 2004),
- the location of unknown populations (Raxworthy *et al.* 2004)

Distribution models have been used to predict

- the location of suitable breeding habitat (Osborne, Alonso & Bryant 2001),
- breeding success (Paradis *et al*. 2000),
- abundance (Jarvis & Robertson 1999),
- genetic variability of species (Scribner *et al*. 2001)

They have also been used to

- help target field surveys (Engler, Guisan & Rechsteiner 2004),
- aid in the design of reserves (Li *et al*.1999),
- inform wildlife management outside protected areas (Milsom *et al*. 2000)
- guide mediatory actions in human–wildlife conflicts (Sitati *et al*. 2003).
- monitor declining species (Osborne, Alonso & Bryant 2001),
- predict range expansions of recovering species (Corsi, Dupre & Boitani 1999),

They have also been used to

- estimate the likelihood of species' long-term persistence in areas considered for protection (Cabeza *et al*. 2004)
- identify locations suitable for introduction (Debeljak *et al*., 2001)
- identify locations suitable for reintroductions (Glenz *et al*., 2001).
- identify sites vulnerable to local extinction (Gates & Donald 2000)
- identify sites vulnerable to species invasion (Kriticos *et al*. 2003),
- explore the potential consequences of climate change (Erasmus *et al*. 2002).

KEY CONCEPT

Every organism has a habitat and a niche.

- A habitat is **all** aspects of the area in which an organism lives.
  - biotic (living) factors
  - Abiotic (non-living) factors

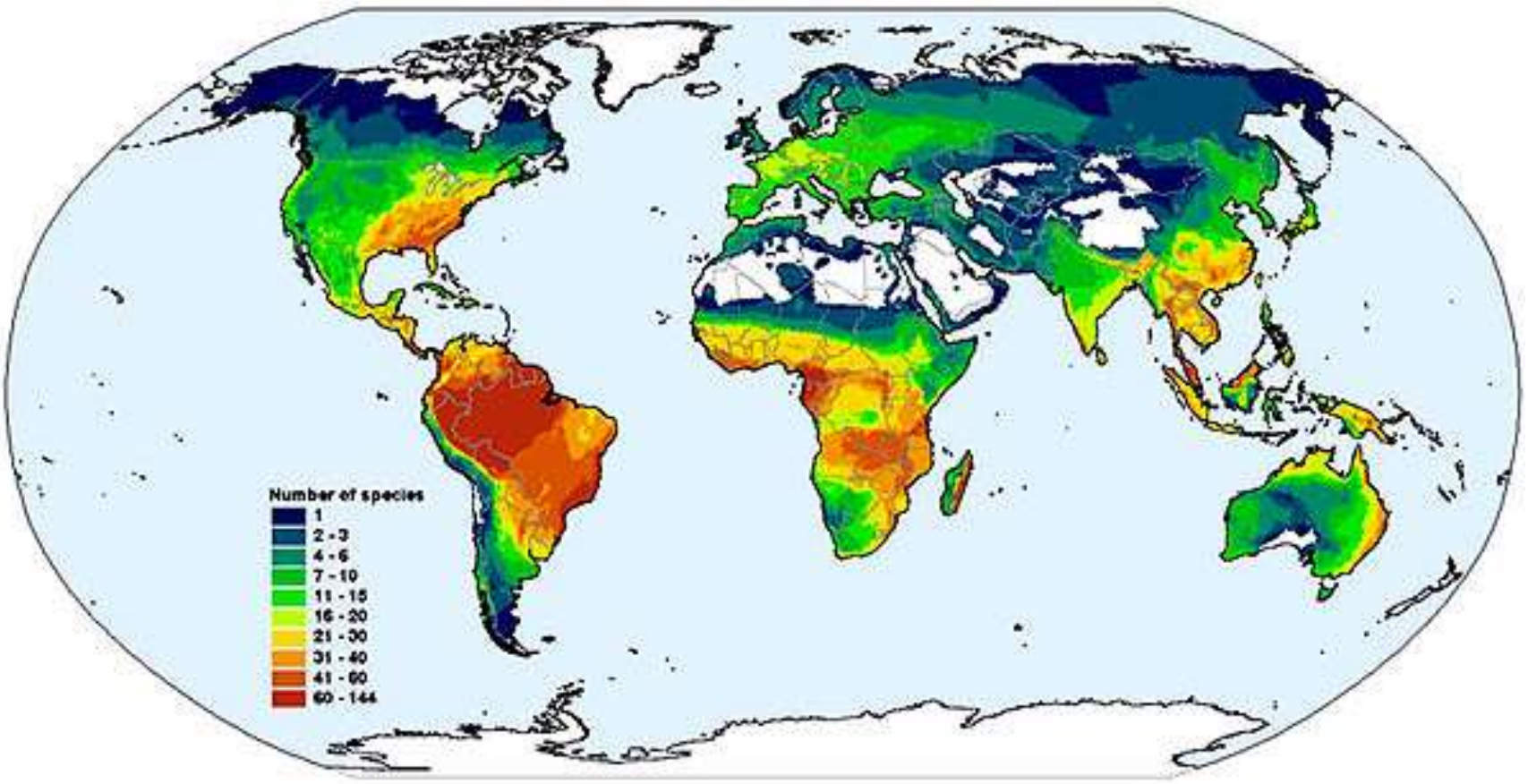
    Ex: ALL aspects of the habitat including grass, trees, and watering hole

- A niche includes all of the physical, chemical, and biological factors that a species needs to survive, stay healthy, and reproduce.
  - food
  - Abiotic conditions (temp, water)
  - Behavior (time of day its active, when it reproduces)

You can think of a habitat as *where* a species lives

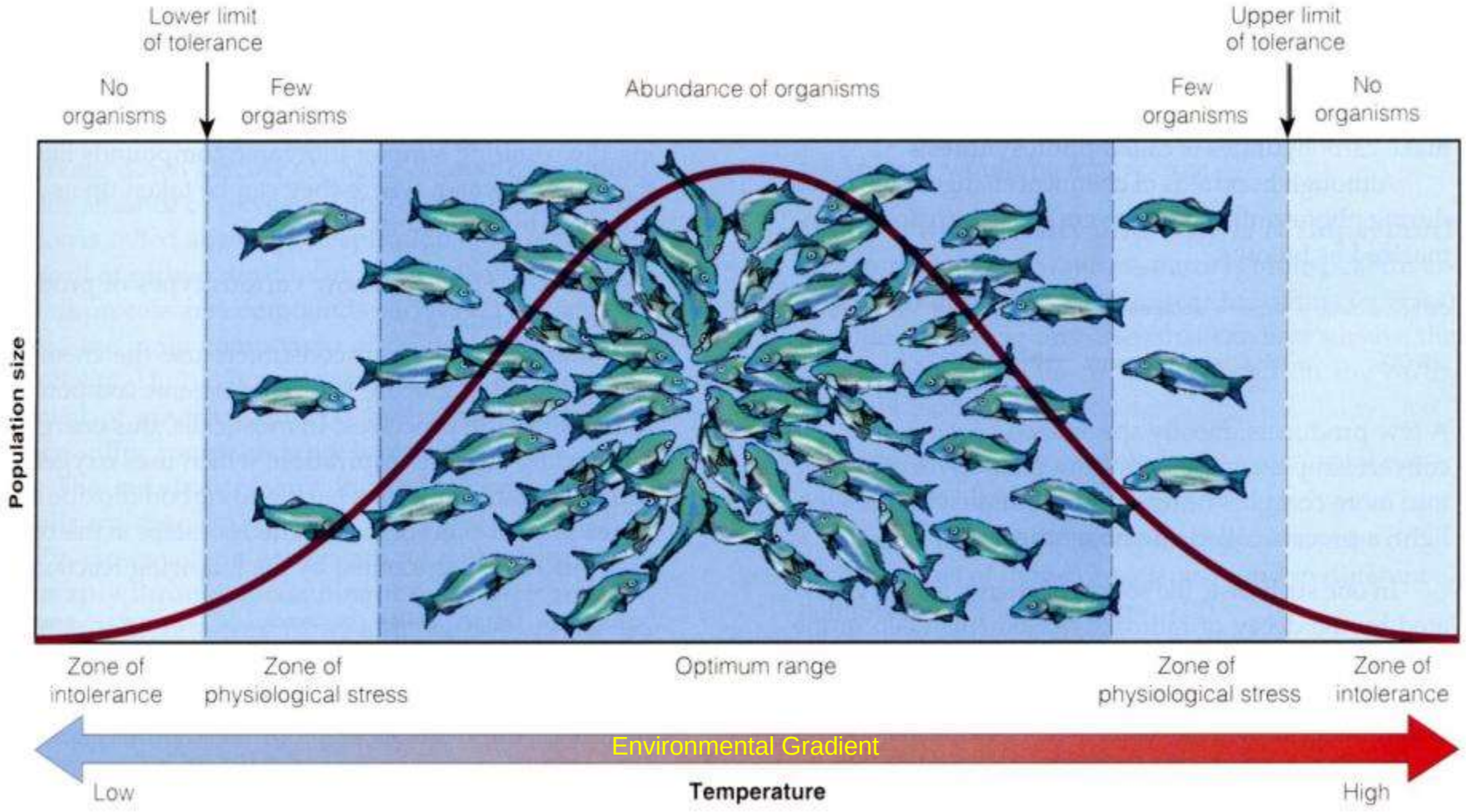
and a niche as *how it lives* within its habitat.

What drives species distributions?



All species have tolerance limits for environmental factors beyond which individuals cannot **survive**, **grow**, or **reproduce**

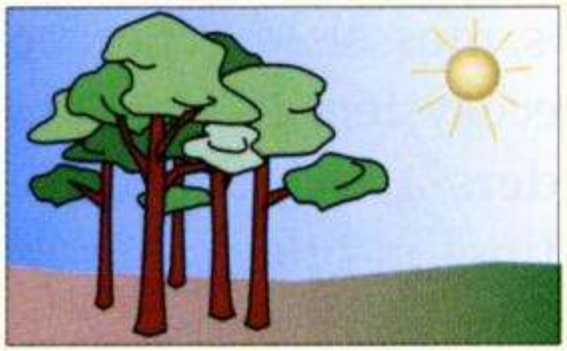
Tolerance Limits and Optimum Range



Tolerance limits exist for all important environmental factors

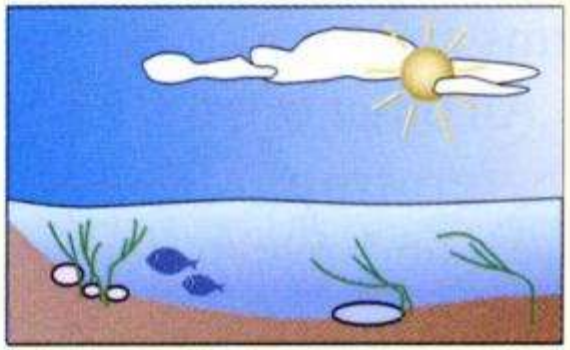
Critical factors and Tolerance Limits

Terrestrial Ecosystems



- **Sunlight**
- **Temperature**
- **Precipitation**
- **Wind**
- **Latitude** (distance from equator)
- **Altitude** (distance above sea level)
- **Fire frequency**
- **Soil**

Aquatic Life Zones



- **Light penetration**
- **Water currents**
- **Dissolved nutrient concentrations** (especially N and P)
- **Suspended solids**

Fundamental versus realized niche

Fundamental (theoretical) niche

- is the full spectrum of environmental factors that can be potentially utilized by an organism

Realized (actual) niche

- represent a subset of a fundamental niche that the organism can actually utilize restricted by:
    - historical factors (dispersal limitations)
    - biotic factors (competitors, predators)
    - realized environment (existent conditions)

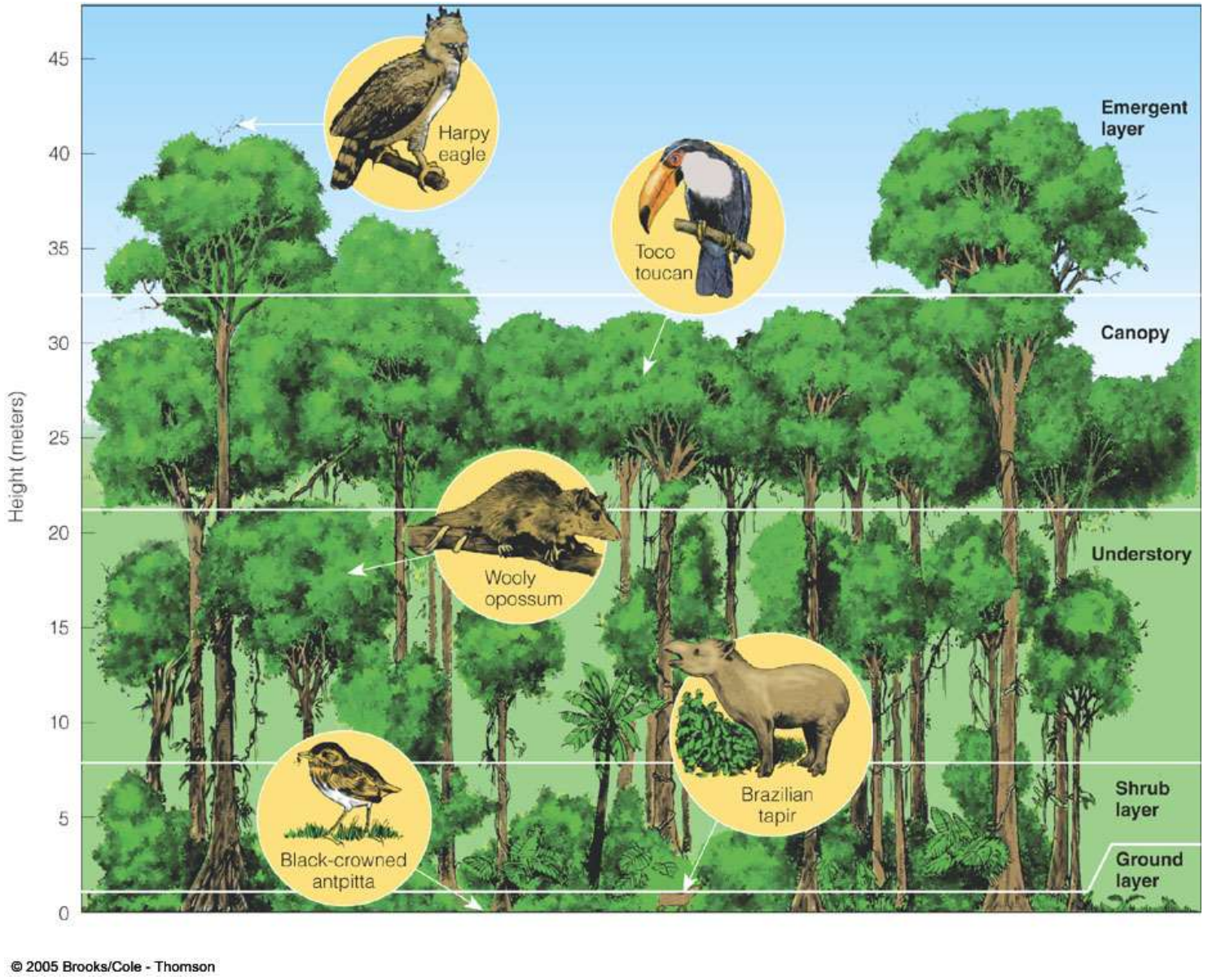
The principle of competitive exclusion

"Two species requiring approximately the same resources are not likely to remain long evenly balanced in numbers in the same habitat."

J. Grinnell (1915)

In consequence, the loser is excluded, at least locally, ***unless…***

1. There are refuges from competition; the potential loser hangs on in marginal habitats; or
2. The loser can re-immigrate from elsewhere; or
3. Disturbances in the environment prevent the winner from gaining a complete monopoly.



Stratification of niches, habitats allows many different species to coexist. This is biodiversity.

Niches

- Resource partitioning avoids competition;
- Realized niches divide resources (insects) among several species
  - woodpeckers, nuthatches, & creepers.
- Each species evolved & adapted to specialized diet.

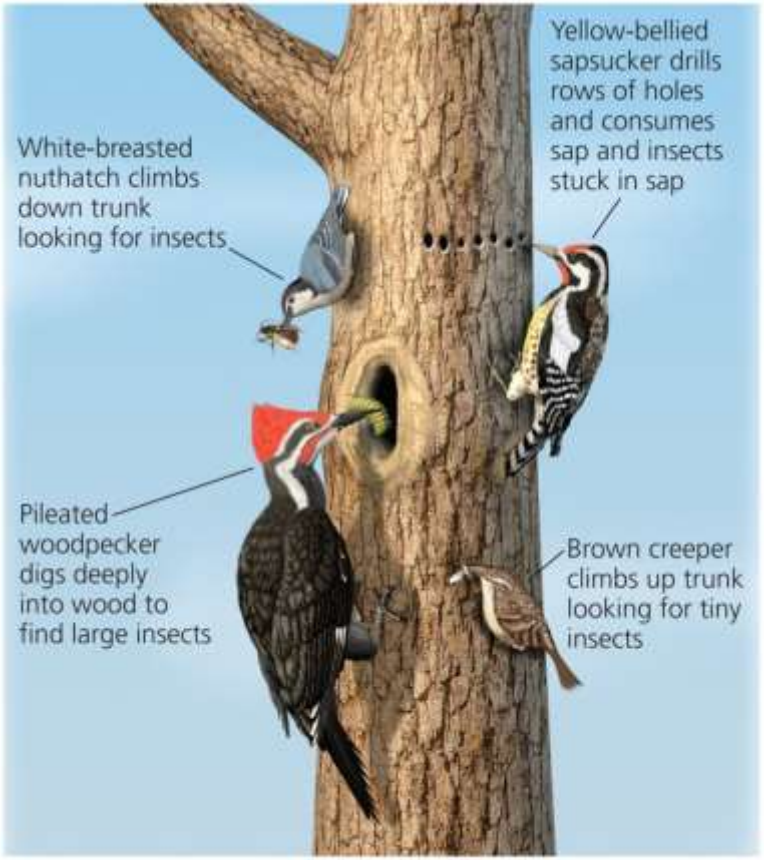


Diagram illustrating the relationship between species' position in geographical space and environmental space

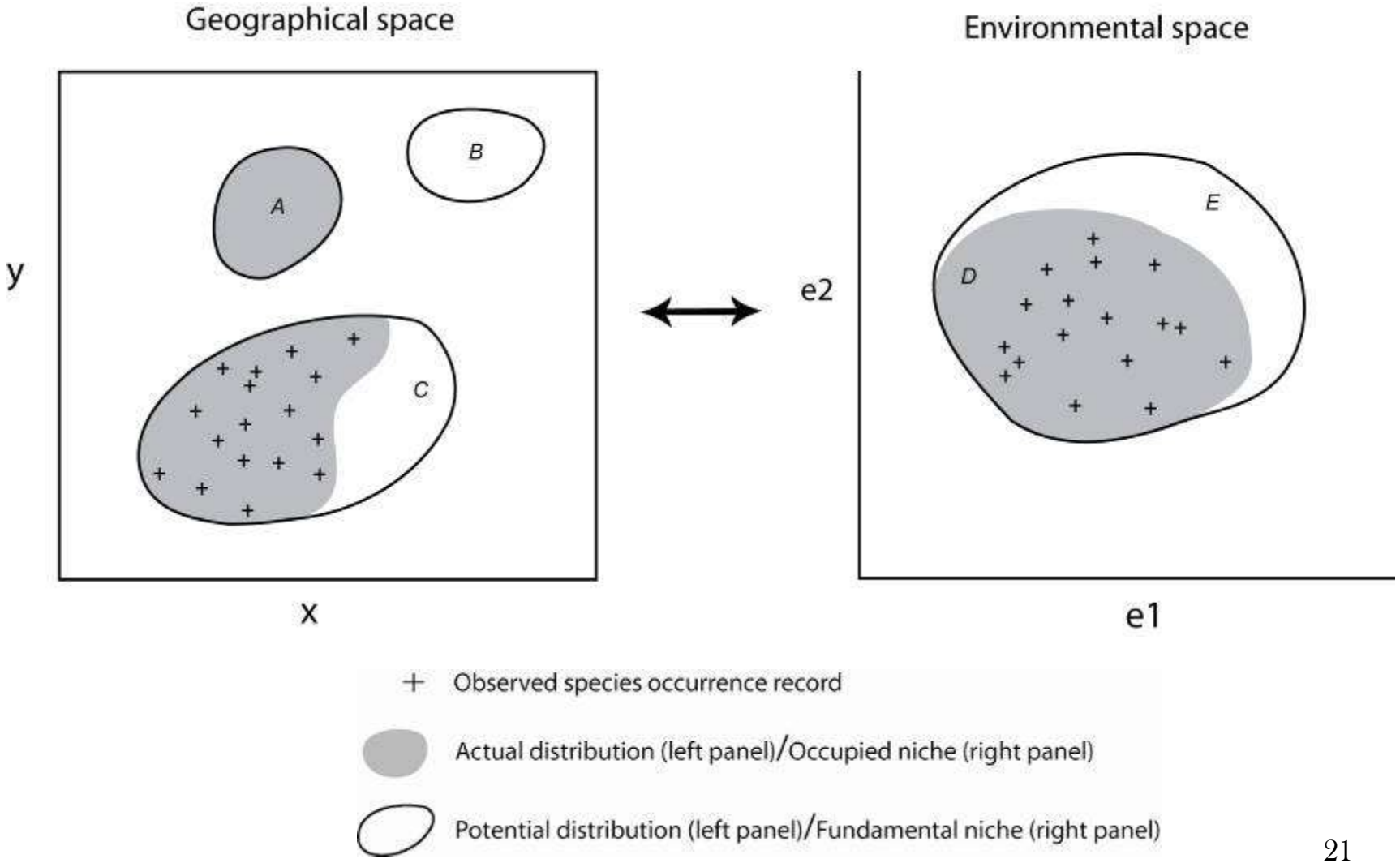
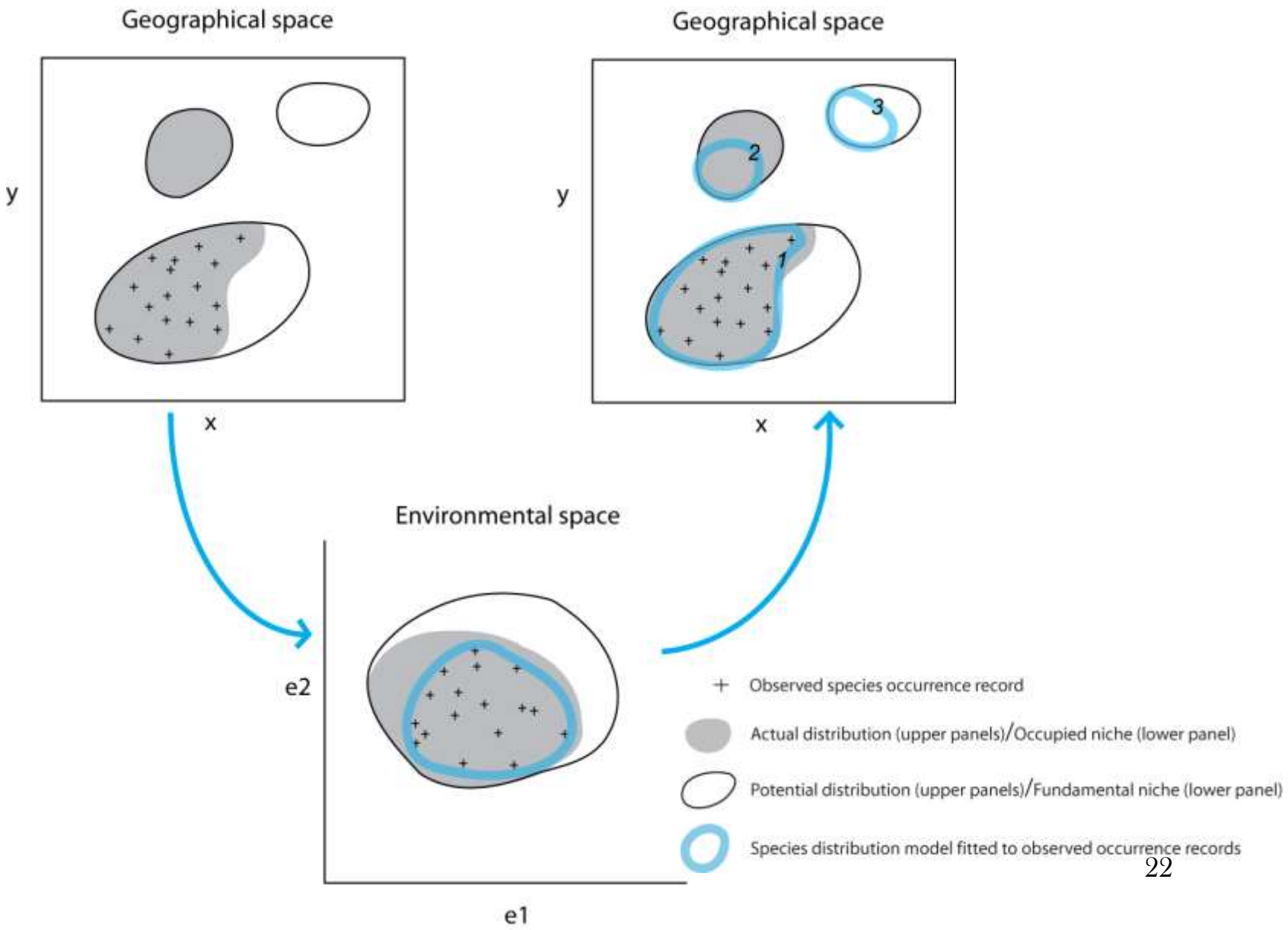


Illustration of the general species' distribution modeling approach

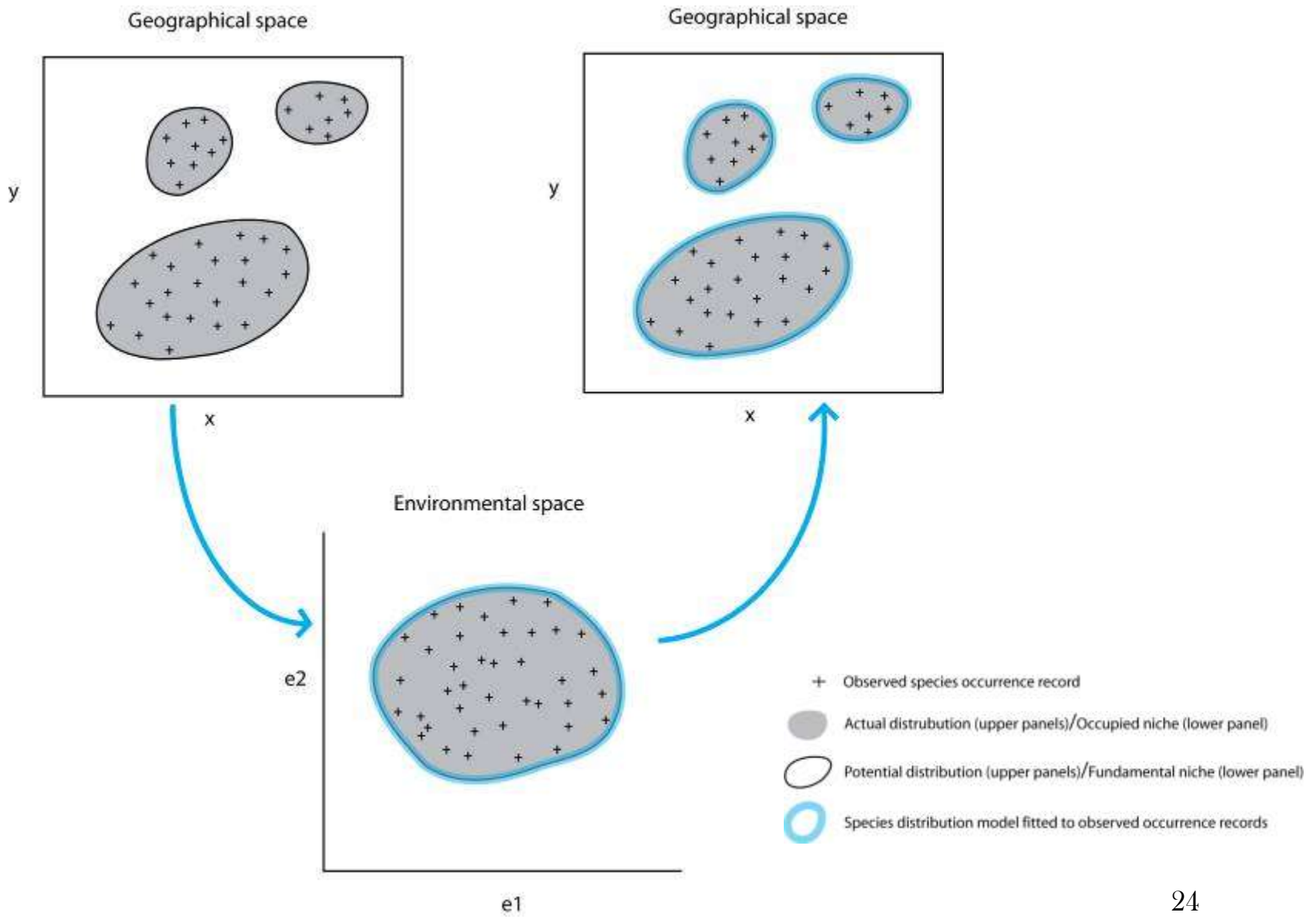


Two key factors determining the degree to which observed localities can be used to estimate the niche or distribution:

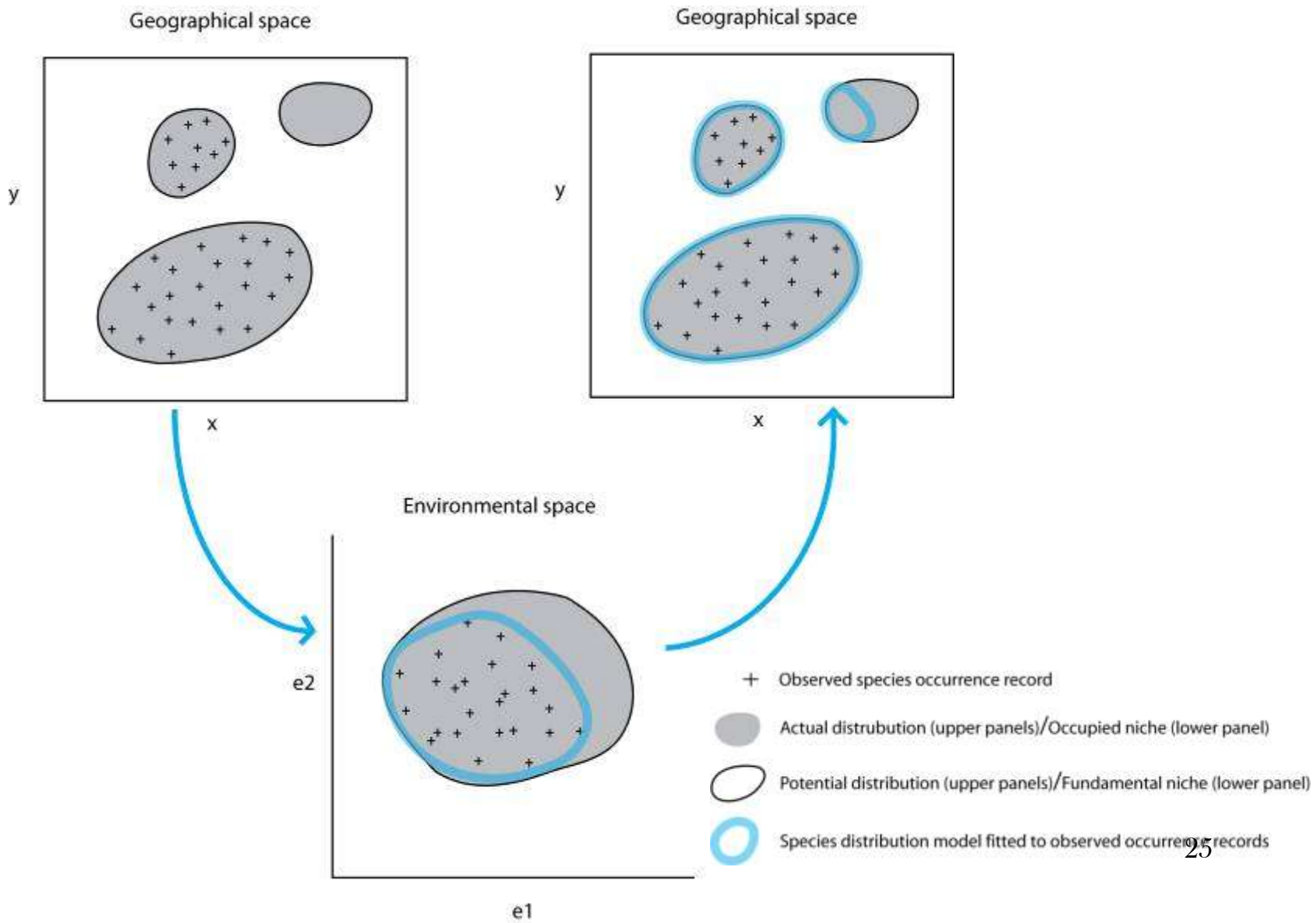
- *Equilibrium*: A species is said to be at equilibrium with current environmental conditions if it occurs in all suitable areas, whilst being absent from all unsuitable areas. The degree to which a species is at equilibrium depends both on biotic interactions (e.g. competitive exclusion from an area) and dispersal ability.

- *Sampling adequacy*: The extent to which the observed occurrence records provide a sample of the environmental space.

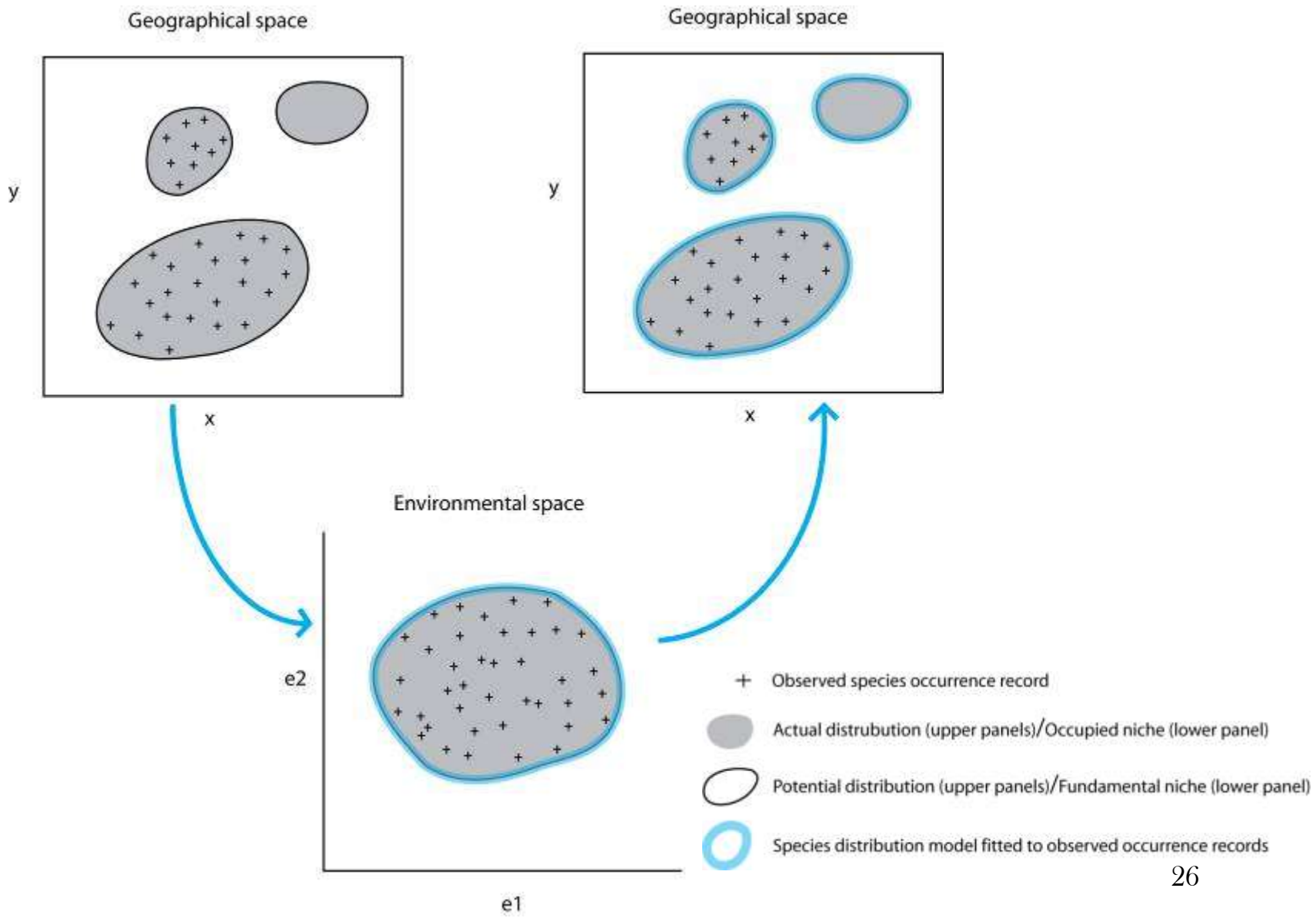
Suppose equilibrium and good sampling: the ideal scenario



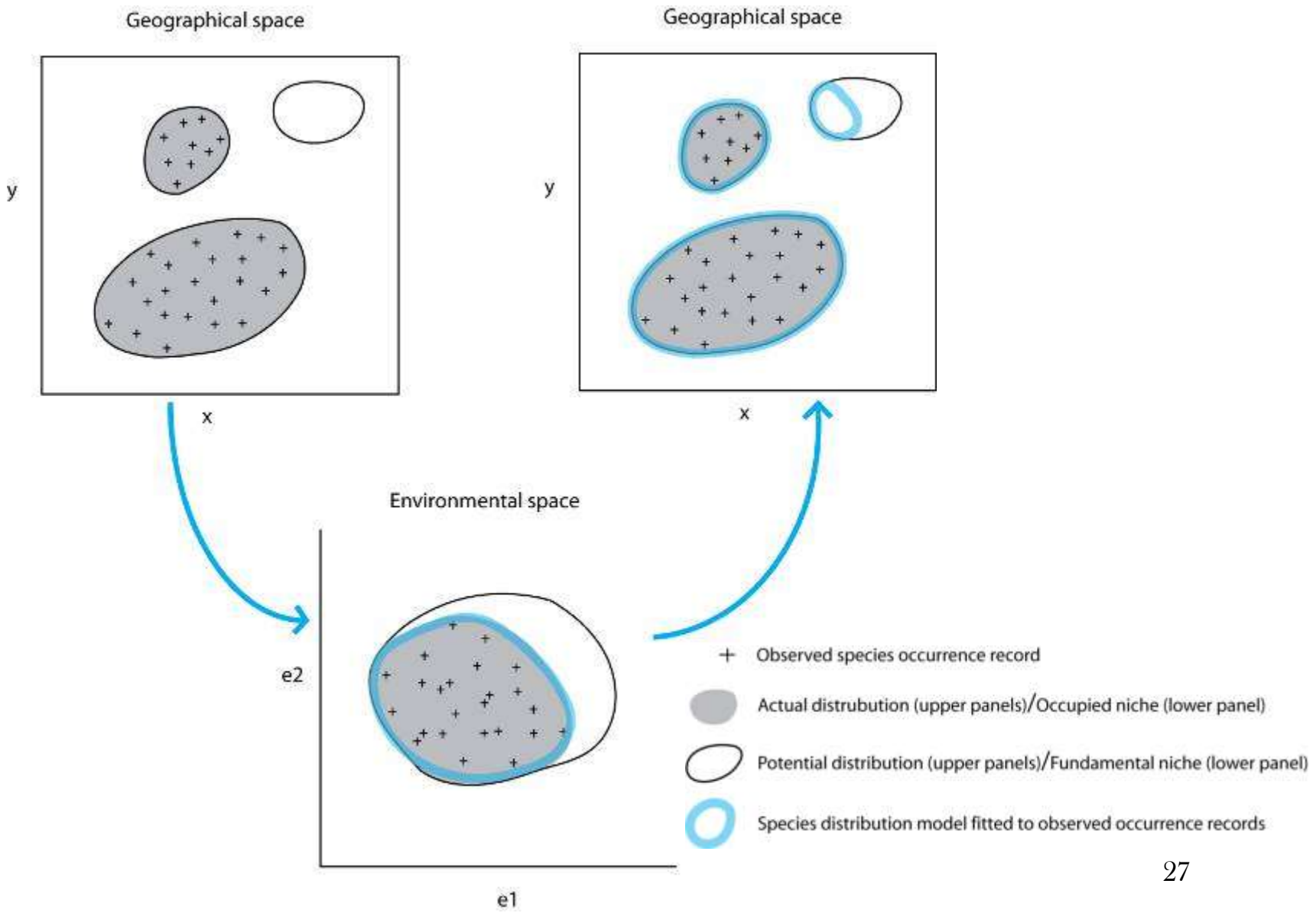
Suppose high equilibrium but poor sampling (in both geographical and environmental space)



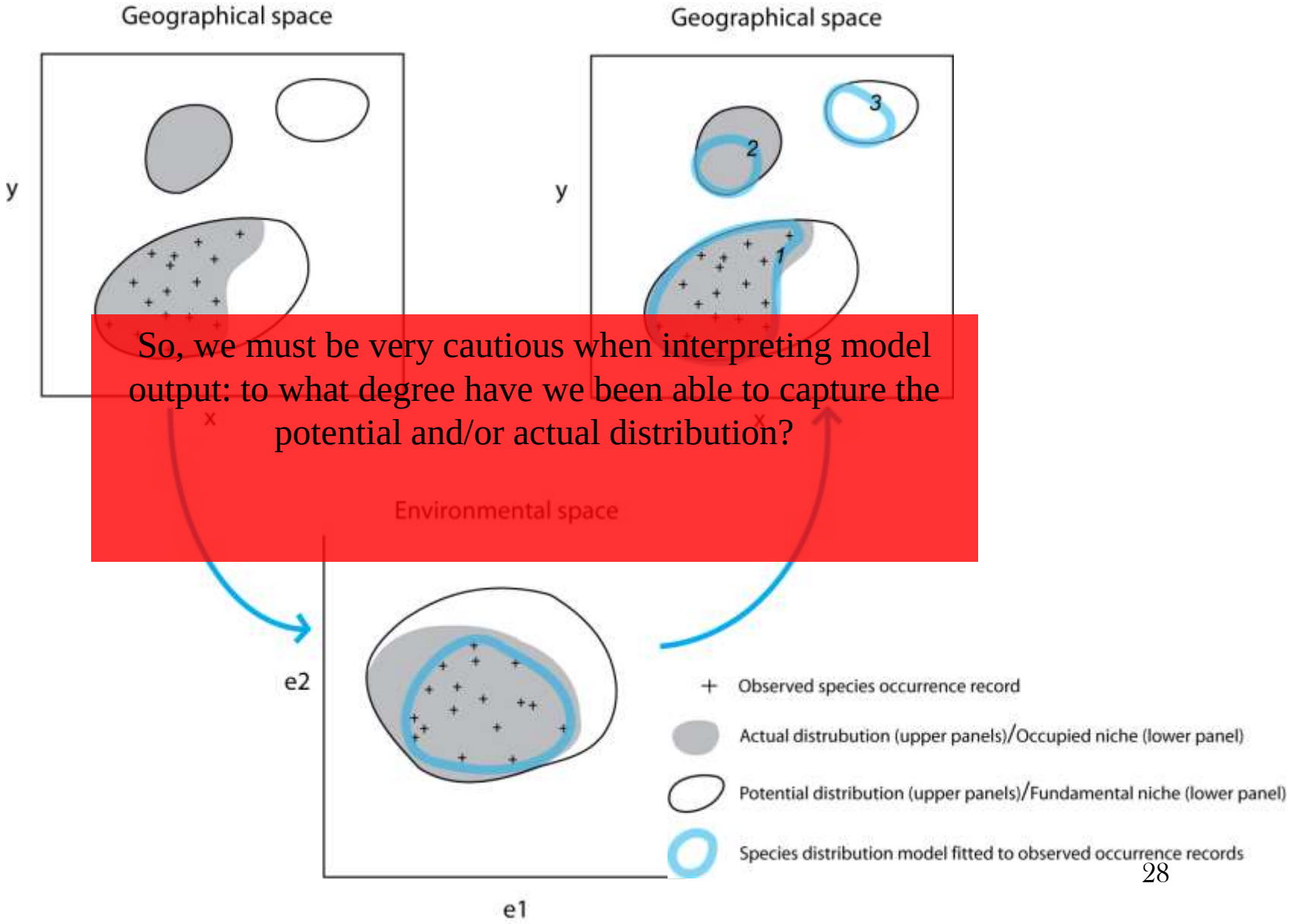
Suppose high equilibrium and poor sampling in geographical space, but good sampling in environmental space



Suppose low equilibrium but good sampling



Back to our first example: in reality we have a combination of dis-equilibrium and incomplete sampling



So, we must be very cautious when interpreting model output: to what degree have we been able to capture the potential and/or actual distribution?

The role of GIS

- The large datasets of biological and environmental data that are used in distribution modeling are ideally suited to being stored and formatted in a GIS.

- GIS is also crucial for visualizing model results and carrying out additional processing of model output.

- However, the distribution modeling itself is usually undertaken outside the GIS framework.

- Some GIS platforms incorporate distribution modeling tools (e.g. DIVA-GIS, IDRISI) or have add-in scripts that enable distribution models to be run (e.g. BIOCLIM scripts for ArcView).

Niche-based modelling – assumptions

- Environmental factors drive species distribution
- Species are in equilibrium with their environment
- Limiting variables – are they really limiting?
- Coincidence with climate or climate shift
- Evidence for species dying/not reproducing due to climate
- Collinearity of variables
- Assumption of assembly rules: niche assembly vs dispersal assembly
- Static vs dynamic approaches: data snapshot or time series response?

Caution! The use and misuse of models

- *Garbage in, garbage out*: a model is only as good as the data put into it.
- *Model extrapolation*: should be treated with a great deal of caution.
- *The lure of complicated technology*: Remember that a model can only be useful if the theoretical underpinnings on which it is based are sound, regardless of how advanced the data and technology are.

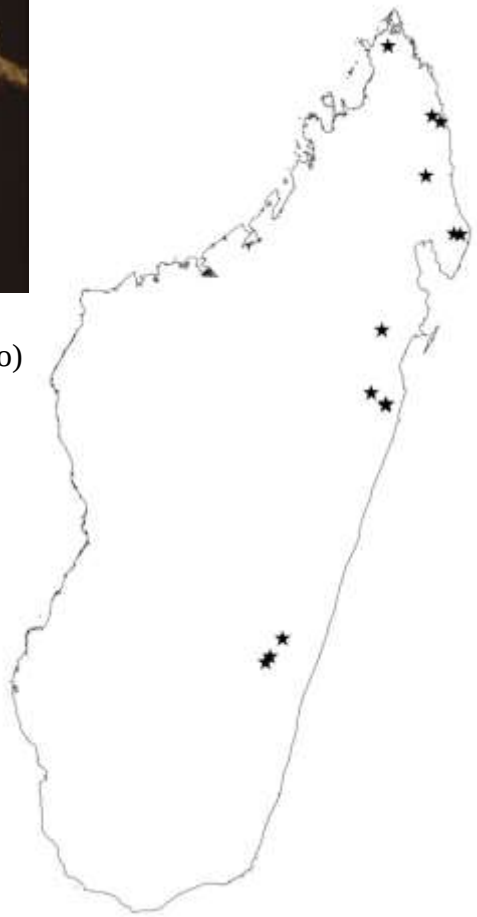
2. Types et sources de données en modélisation

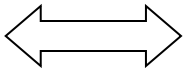
Two types of data for model input:

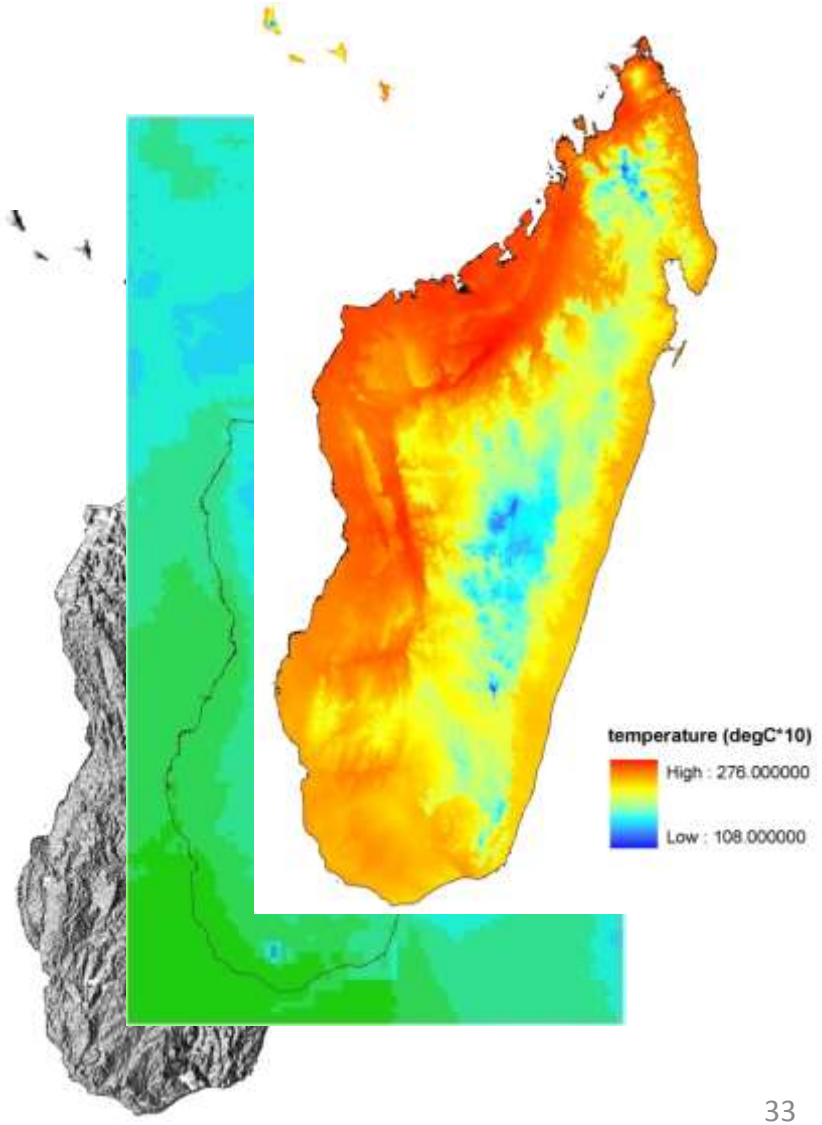
Species' distribution data

Environmental data

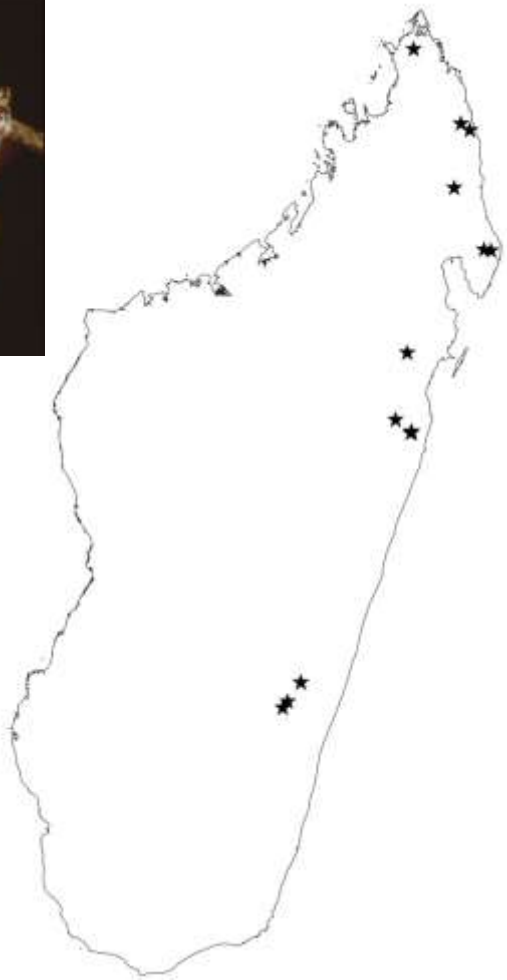
Uroplatus sp.
(leaf-tailed gecko)

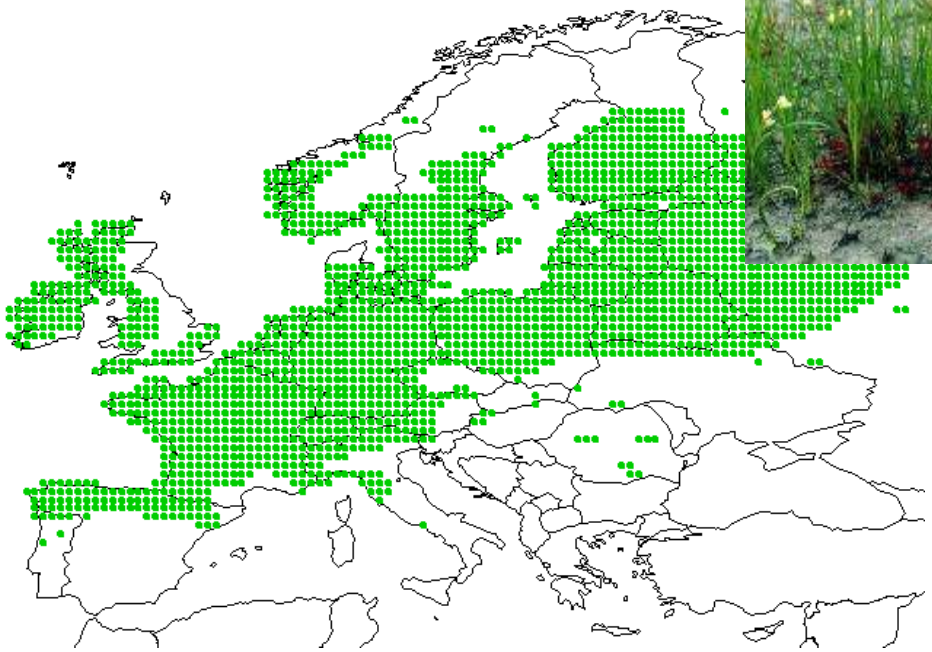






Species' distribution data: presence-only or presence/absence?





Presence and (assumed) absence data for a plant in Europe

Presence-only data for a gecko in Madagascar

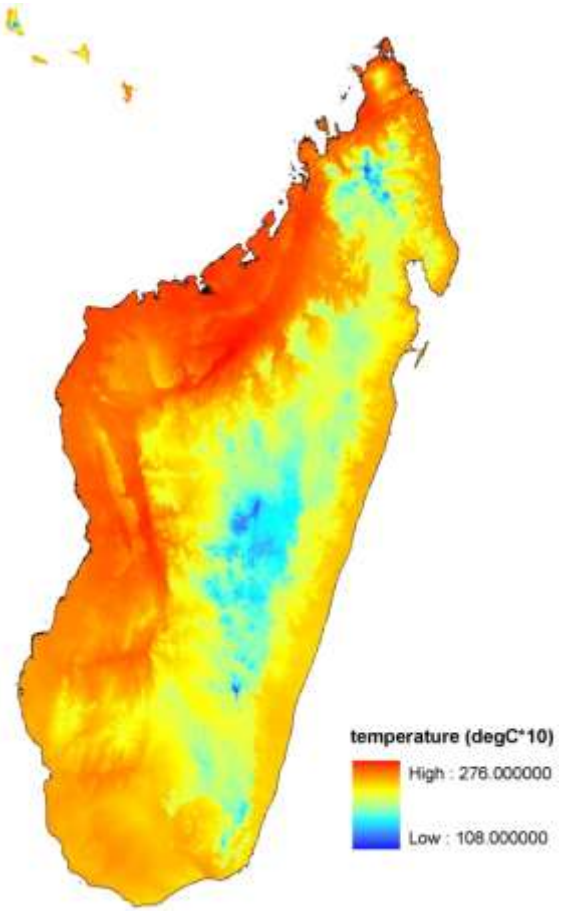
When is an observed species absence really an absence of suitable conditions?

A species may be classified as ‘absent’ for a number of reasons, but this does not necessarily denote absence of suitable conditions:

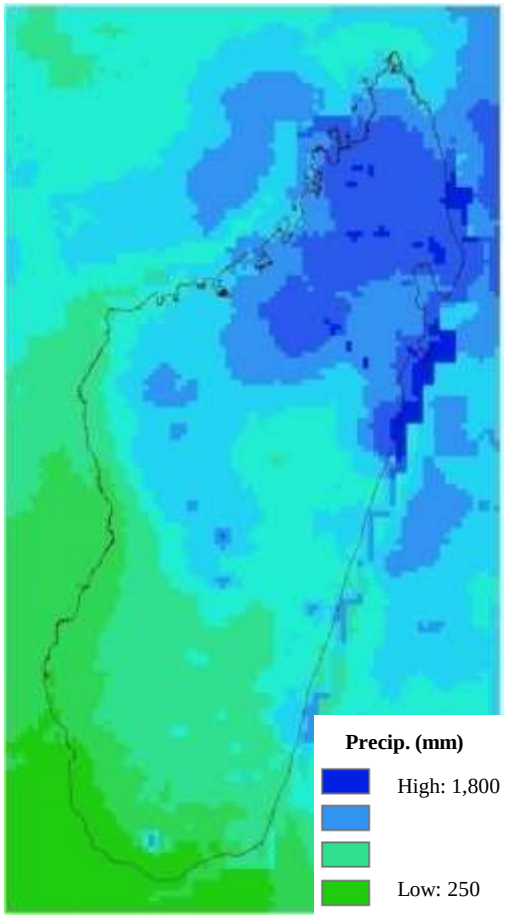
1. The species could not be detected, even though it was present — 'False absence'
2. The species was absent, even though the environment is suitable (e.g. due to dispersal limitation, or metapopulation dynamics) — 'False absence'
3. The environment is truly unsuitable for the species — 'True absence'

Take care when using ‘absence’ data

Environmental Data: common types



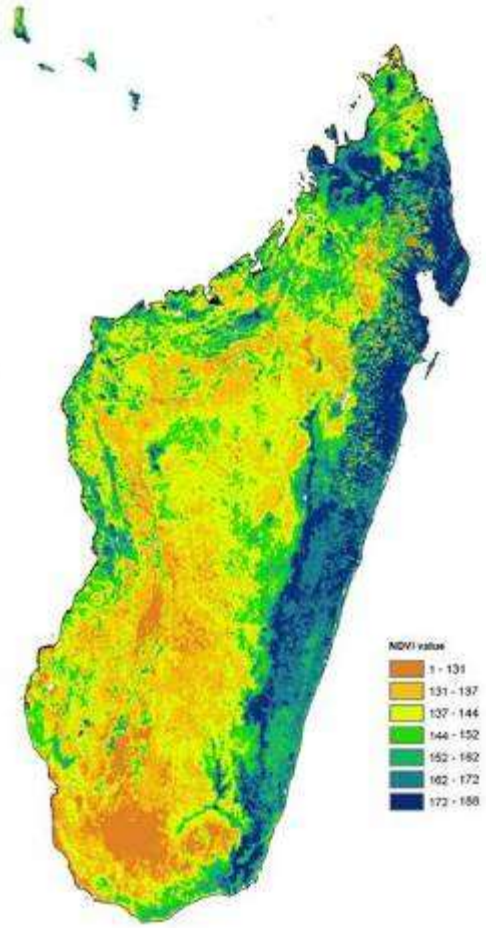
Mean annual temperature
Source: WorldClim



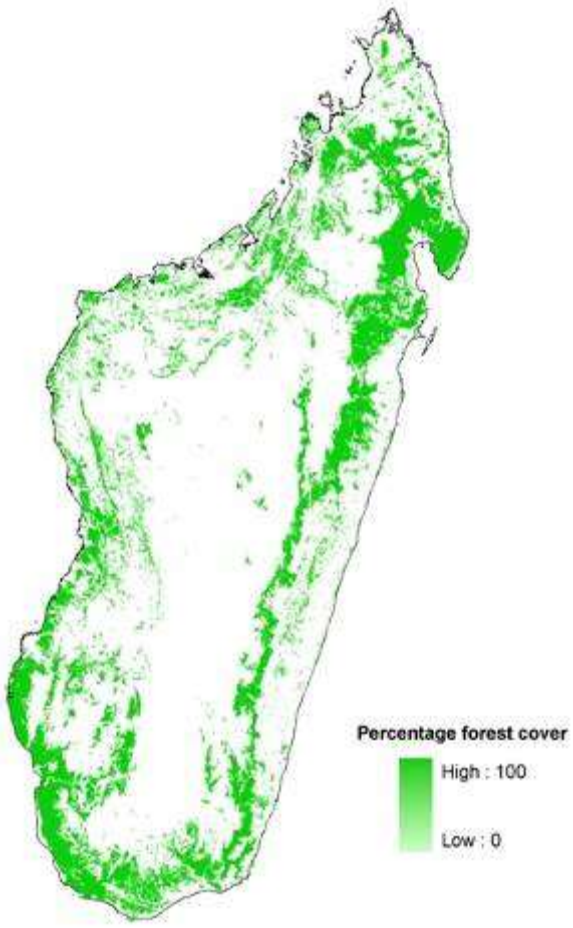
Mean annual precipitation
Source: NOAA FEWS

Aspect: East-West
Source: USGS Hydro1k

Environmental Data: common types



AVHRR NDVI
Source: NASA



% forest cover
Source: IEFN and CI

Geology
Source: Kew gardens

(note: categorical data)

Some example sources of biological and environmental data for use in species' distribution modeling

Type of data	Source
Species' distributions	
- Data for a wide range of organisms in many regions of the world	Global Biodiversity Information Facility (GBIF): www.gbif.org
- Data for a range of organisms, mostly rare and/or endangered, and primarily in North America	NatureServe: www.NatureServe.org
Climate	
- Interpolated climate surfaces for the globe at 1km resolution	WorldClim: http://www.worldclim.org/
- Scenarios of future climate change for the globe	Intergovernmental Panel on Climate Change (IPCC): http://ipcc-ddc.cru.uea.ac.uk/
- Reconstructed palaeoclimates	NOAA: http://www.ncdc.noaa.gov/paleo/paleo.html
Topography	
- Elevation and related variables for the globe at 1km resolution	USGS: http://edc.usgs.gov/products/elevation/gtopo30/hydro/index.html
Remote sensing (satellite)	
- Various land cover datasets	Global Landcover Facility: http://glcf.umiacs.umd.edu/data/
- Various atmospheric and land products from the MODIS instrument	NASA: http://modis.gsfc.nasa.gov/data/
Soils	
- Global soil types	UNEP: http://www.grid.unep.ch/data/data.phpcategory=lithosphere
Marine	
- Various datasets describing the world's oceans	NOAA: www.nodc.noaa.gov

Key bioclimatic parameters

Variable Number	Variable	Min temp (°C)	Max temp (°C)	Rainfall (mm month^{-1})	Radiation (W m^{-2}d^{-1})	Pan evaporation (mm d^{-1})
Bio01	Annual mean temperature (°C)	×	×			
Bio02	Mean diurnal temperature range (mean(period max-min)) (°C)	×	×			
Bio03	Isothermality (Bio02 ÷ Bio07)	×	×			
Bio04	Temperature seasonality (C of V)	×	×			
Bio05	Max temperature of warmest week (°C)		×			
Bio06	Min temperature of coldest week (°C)	×				
Bio07	Temperature annual range (Bio05-Bio06) (°C)	×	×			

Key bioclimatic parameters

Variable Number	Variable	Min temp (°C)	Max temp (°C)	Rainfall (mm month^{-1})	Radiation (W m^{-2}d^{-1})	Pan evaporation (mm d^{-1})
Bio08	Mean temperature of wettest quarter (°C)	×	×	×		
Bio09	Mean temperature of driest quarter (°C)	×	×	×		
Bio10	Mean temperature of warmest quarter (°C)	×	×			
Bio11	Mean temperature of coldest quarter (°C)	×	×			
Bio12	Annual precipitation (mm)			×		
Bio13	Precipitation of wettest week (mm)			×		
Bio14	Precipitation of driest week (mm)			×		

Key bioclimatic parameters

Variable Number	Variable	Min temp (°C)	Max temp (°C)	Rainfall (mm month^{-1})	Radiation (W m^{-2}d^{-1})	Pan evaporation (mm d^{-1})
Bio15	Precipitation seasonality (C of V)			×		
Bio16	Precipitation of wettest quarter (mm)			×		
Bio17	Precipitation of driest quarter (mm)			×		
Bio18	Precipitation of warmest quarter (mm)	×	×	×		
Bio19	Precipitation of coldest quarter (mm)	×	×	×		
Bio20	Annual mean radiation (W m^{-2})				×	
Bio21	Highest weekly radiation (W m^{-2})				×	

Key bioclimatic parameters

Variable Number	Variable	Min temp (°C)	Max temp (°C)	Rainfall (mm month^{-1})	Radiation (W m^{-2}d^{-1})	Pan evaporation (mm d^{-1})
Bio22	Lowest weekly radiation (W m^{-2}				×	
Bio23	Radiation seasonality (C of V)				×	
Bio24	Radiation of wettest quarter (W m^{-2})			×	×	
Bio25	Radiation of driest quarter (W m^{-2})			×	×	
Bio26	Radiation of warmest quarter (W m^{-2})	×	×		×	
Bio27	Radiation of coldest quarter (W m^{-2})	×	×		×	
Bio28	Annual mean moisture index			×		×

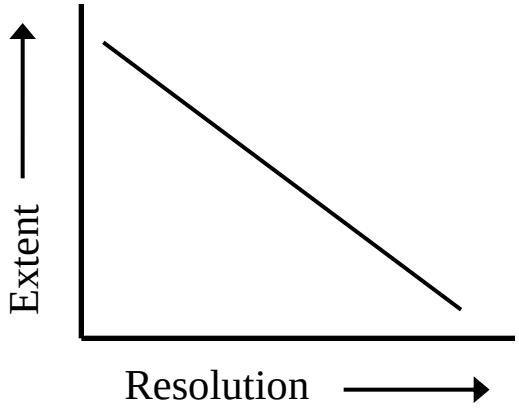
Key bioclimatic parameters

Variable Number	Variable	Min temp (°C)	Max temp (°C)	Rainfall (mm month^{-1})	Radiation (W m^{-2}d^{-1})	Pan evaporation (mm d^{-1})
Bio29	Highest weekly moisture index			×		×
Bio30	Lowest weekly moisture index			×		×
Bio31	Moisture index seasonality (C of V)			×		×
Bio32	Mean moisture index of wettest quarter			×		×
Bio33	Mean moisture index of driest quarter			×		×
Bio34	Mean moisture index of warmest quarter	×	×	×		×
Bio35	Mean moisture index of coldest quarter	×	×	×		×

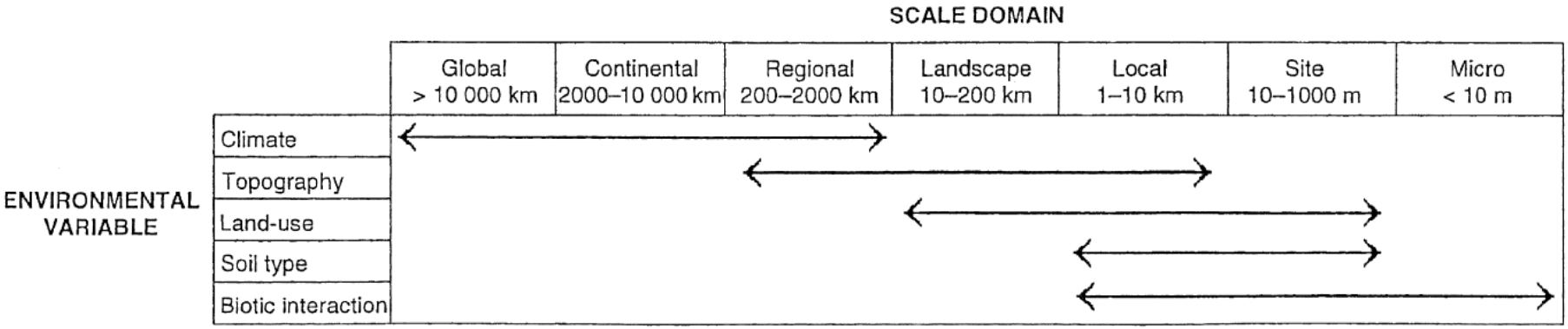
General data issue: spatial scale

Spatial scale has two elements: **resolution** and **extent**

In practice, resolution and extent tend to be inversely related



Spatial scale



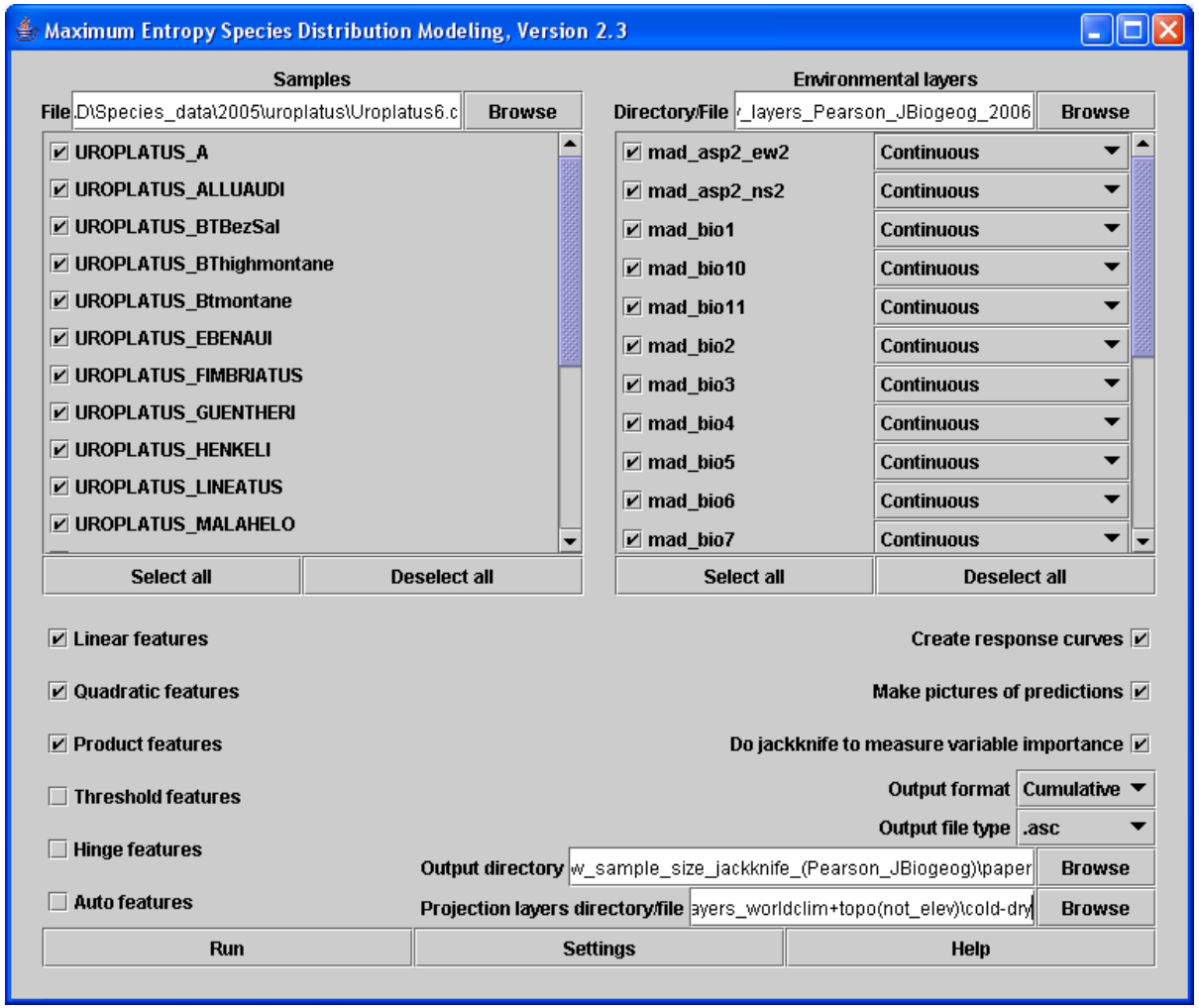
Schematic example of the scale domains associated with different biotic and abiotic environmental variables that could affect the distribution of species. (Pearson and Dawson 2003)

3. Algorithmes de modélisation

Model algorithms: *some* approaches that have been applied:

Method(s)	Model/software name
Gower Metric	DOMAIN
Ecological Niche Factor Analysis (ENFA)	BIOMAPPER
Maximum Entropy	MAXENT
Genetic algorithm	GARP
Regression (GLM, GAM, BRT, MARS)	Implemented in R
Artificial Neural Network (ANN)	SPECIES
Multiple methods	BIOMOD

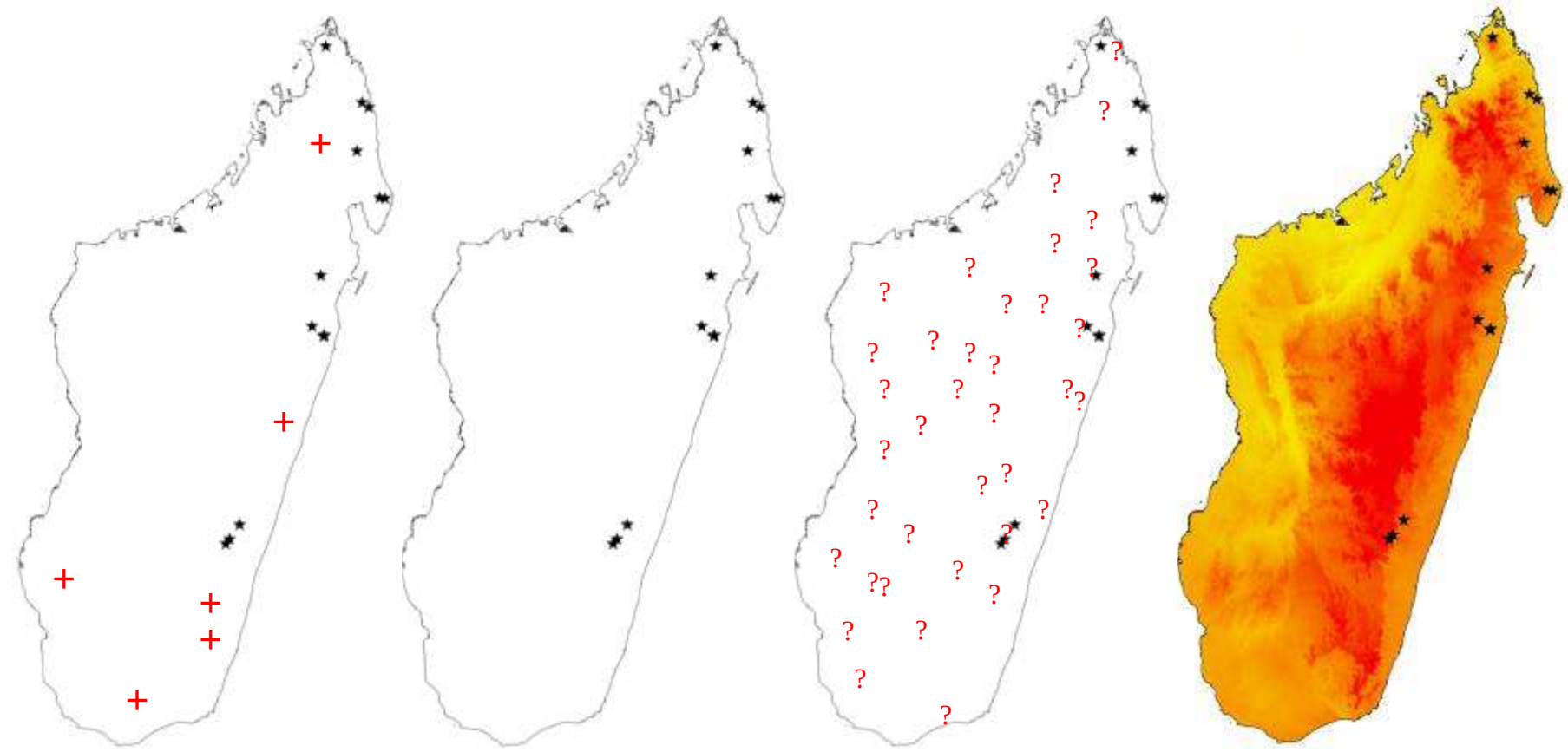
Software to implement the *Maxent* approach



For free download see: http://www.cs.princeton.edu/~schapire/maxent/

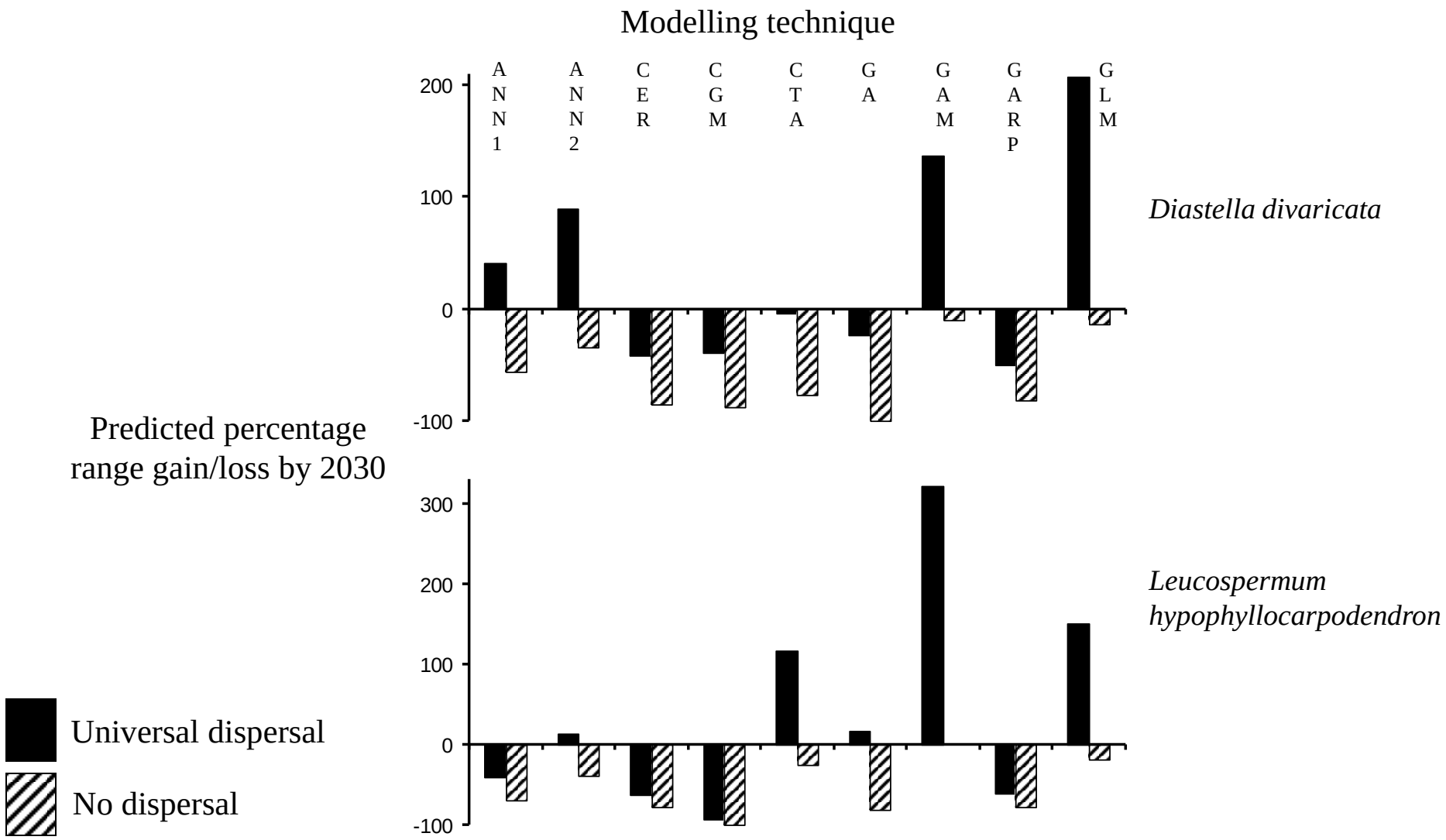
(Phillips et al. 2006 *Ec. Mod.* 190; see also the practical exercise accompanying this presentation)

Alternative methods use species' distribution data differently:



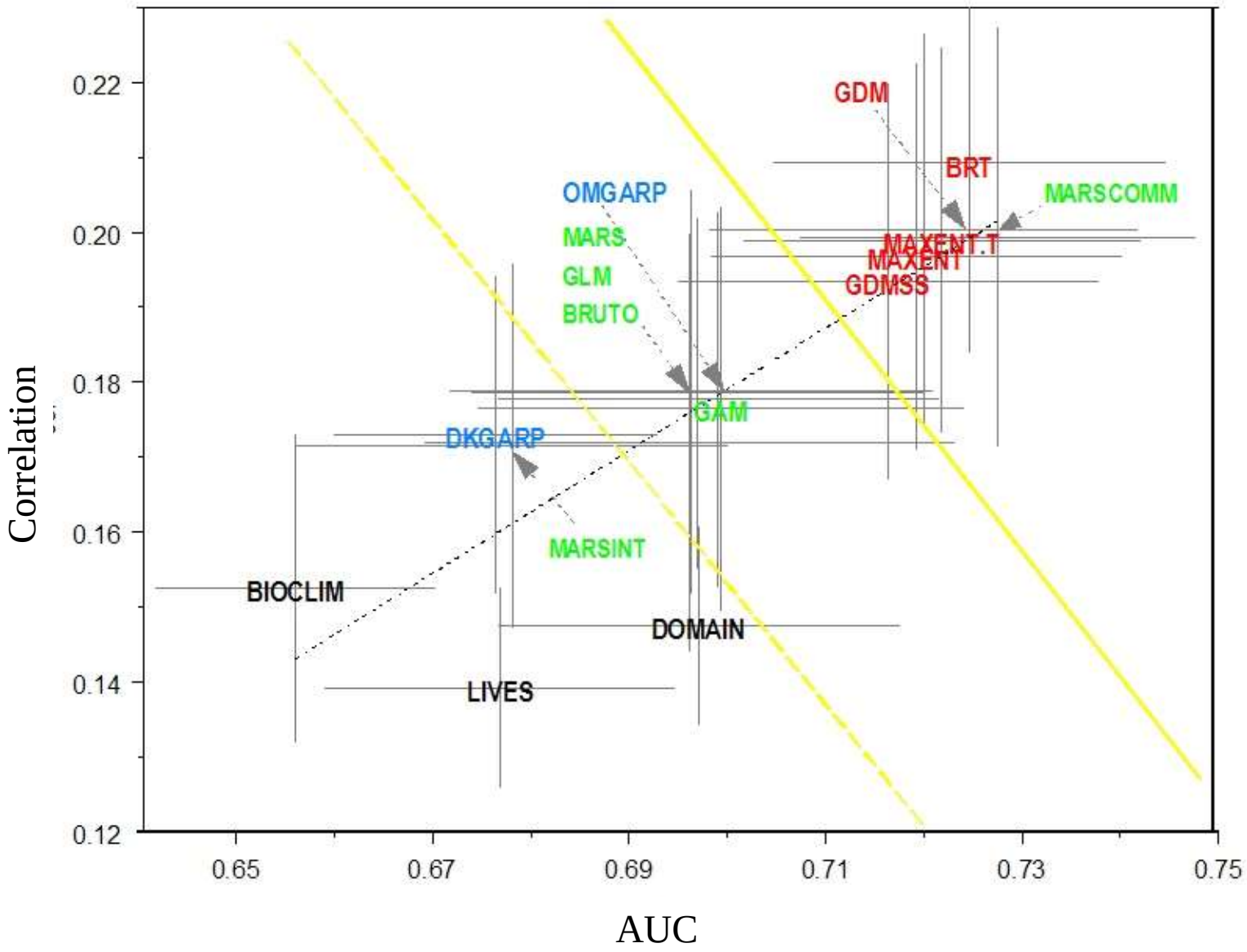
Presence/absence | Presence-only | Presence/pseudo-absence | Presence/background

How important is model selection? Model-based uncertainty



(from Pearson *et al.*, *Biogeography.* 2006, Blackwell Publishing)

A comparison of different methods



(from Elith *et al.*, *Ecography* 2006, Blackwell Publishing)

Predictive modelling algorithms

BIOCLIM

Is based on a boxcar environmental envelope algorithm.

The minimum and maximum values for each environmental predictor define the multidimensional environmental box where the element is known to occur. Study area sites that hale environmental conditions within the boundaries of this multidimensional box are predicted as potential sites of occupancy.

Predictive modelling algorithms.

BIOCLIM

Advantages:

- uses presence occurrence data only,
- suitable when the number of known records is extremely low (BioClim is particularly useful modeling system for use with threatened species),
- model easily interpreted and represented as a predicted distribution map.

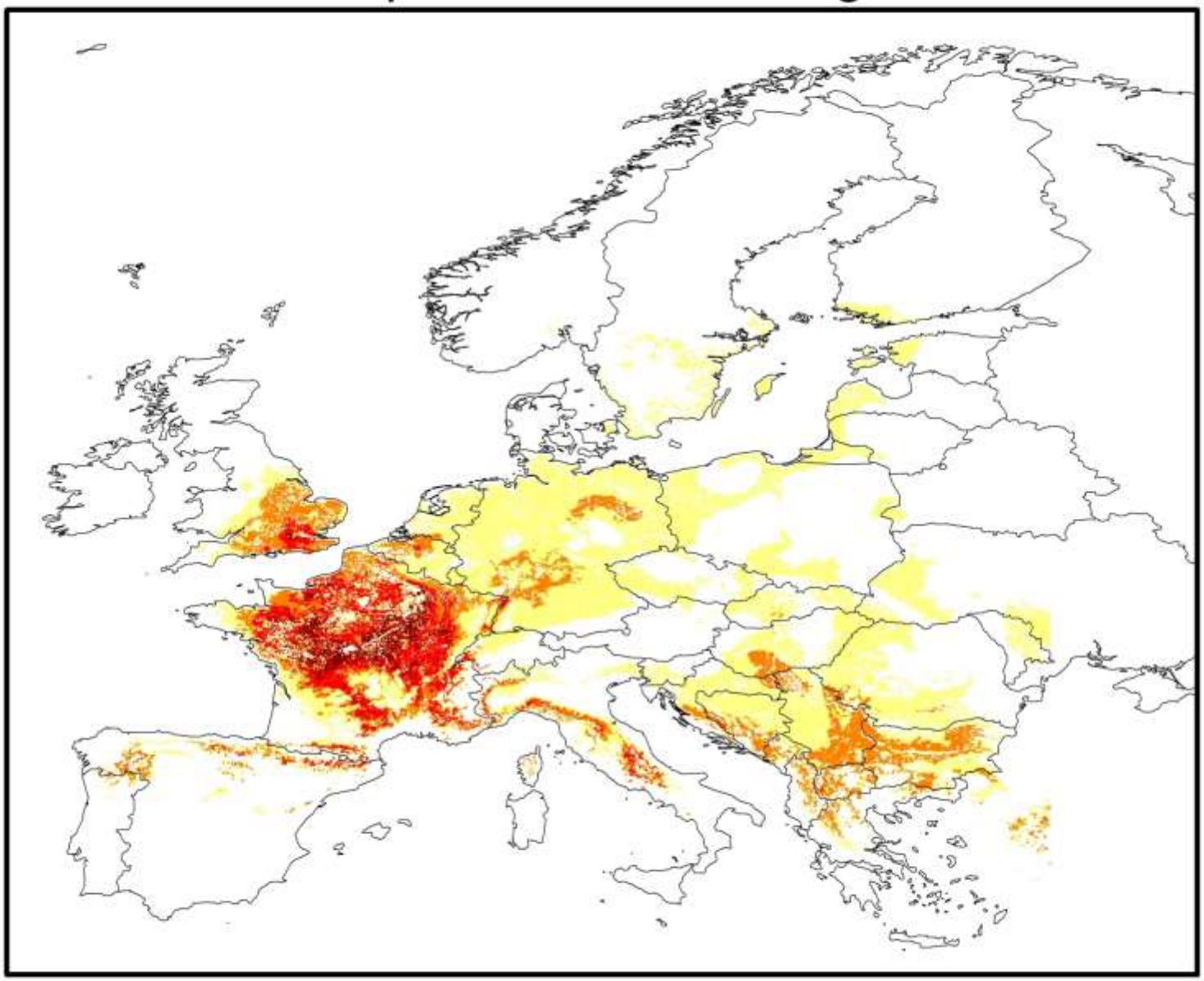
Predictive modelling algorithms.

BIOCLIM

Disadvantages:

- a tendency to over-predict,
- does not address potential correlations and interactions among environmental variables,
- gives equal weight to all environmental predictors,
- sensitive to outliers and sampling bias,
- cannot use categorical data,
- no procedure for variable selection.

BIOCLIM predictive modelling result



Predictive modelling algorithms

MAXENT- maximum entropy approach

MaxEnt estimates the most uniform distribution (maximum entropy) of the occurrence points across the study area given the constraint that the expected value of each environmental predictor variable under this estimated distribution matches its empirical average (average values for the set occurrence data).

The program starts with an uniform probability distribution and iteratively altering one weight at a time to maximize the likelihood to reach the optimum probability distribution. The algorithm is guaranteed to converge and therefore the outputs are deterministic.

Predictive modelling algorithms

MAXENT

Advantages:

- uses presence occurrence data only,
- probability distribution mathematically defined therefore model formulation relatively transparent,
- can consider interactions between environmental variables,
- provides ability to consider polynomial transformations of the environmental predictors,
- potential to investigate the influence each environmental predictor has on the elements distribution,
- relatively easy to run, stand alone software,
- seems to perform relatively well with small sample sizes of occurrence data.

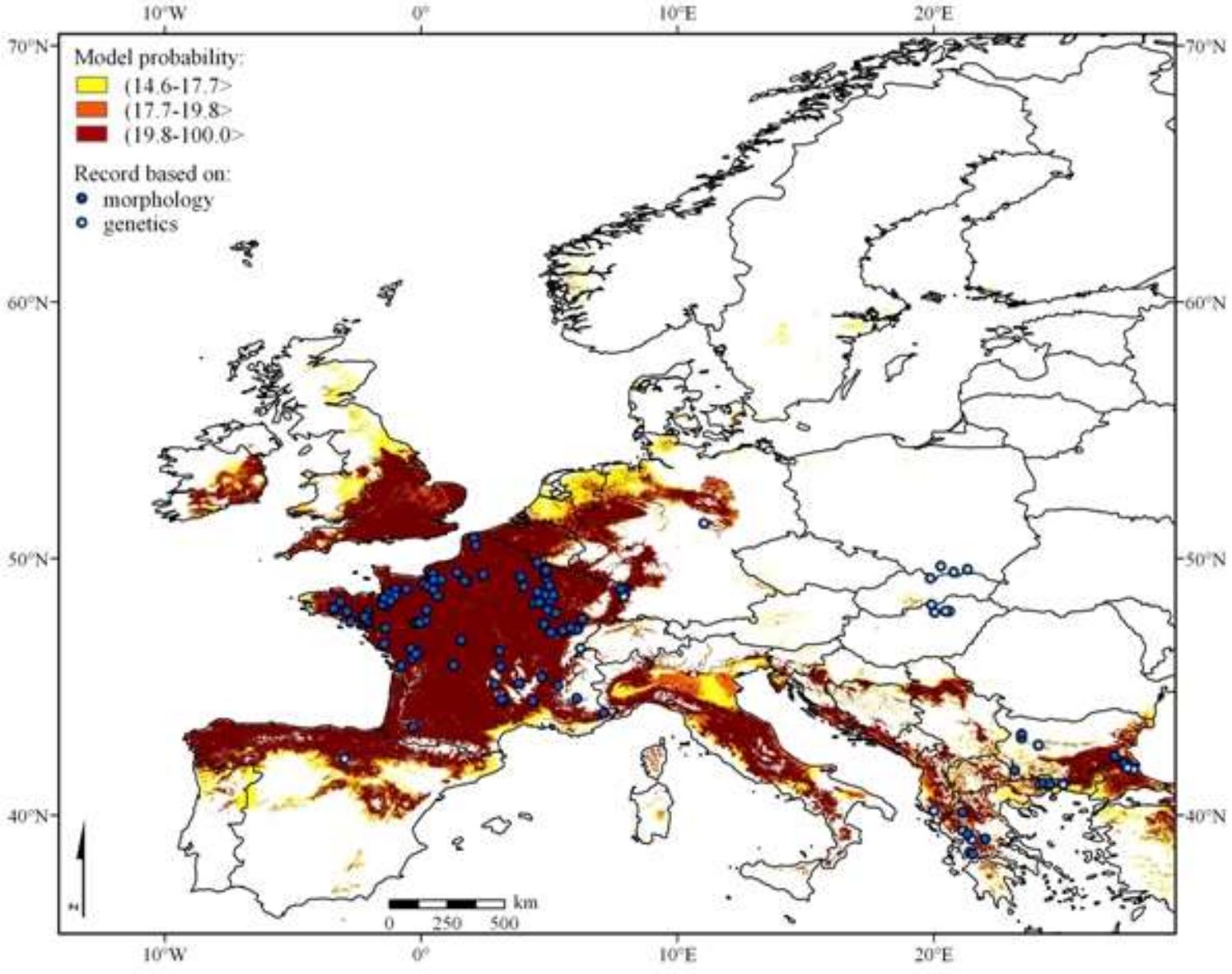
Predictive modelling algorithms.

MAXENT

Disadvantages:

- no procedure for variable selection,
- extremely computer intensive,
- limited experiments investigating potential weaknesses when dealing with biased sampling.

MaxENT output map



Predictive modelling algorithms

Desktop GARP

Genetic Algorithm for Rule-set Prediction (GARP) uses several predictive modelling algorithms.

A GARP run begins by dividing the element occurrence data set into two subsets: training and test dataset.The first rule is generated by applying one of the four algorithms and evaluating the omission and commission errors.

In the next iteration it resamples the occurrence data again applying another algorithm to create another rule. The model is then evaluated and changes in prediction accuracy will determine whether to incorporate or disregard the rule from the rule set. This process is repeated until it cannot create a better model or it has reached the maximum number of iterations set by the user

Predictive modelling algorithms

Desktop GARP:

Advantages:

- uses presence occurrence data only,
- relatively easy to run, stand alone software,
- theoretically GARP should perform better than individual implementations of the algorithms it employs, since it searches for and applies only the most appropriate ‘rules’.

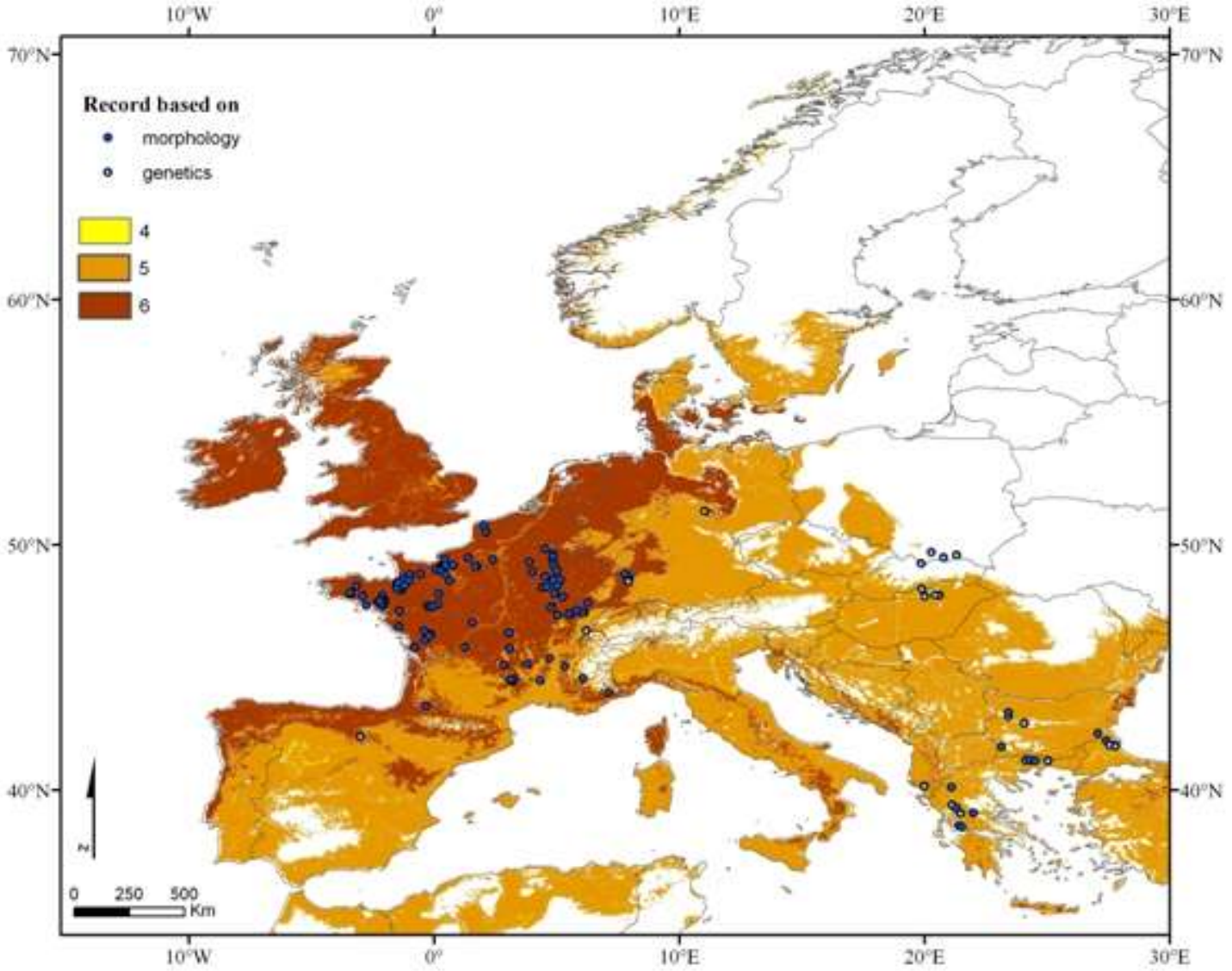
Predictive modelling algorithms.

Desktop GARP

Disadvantages:

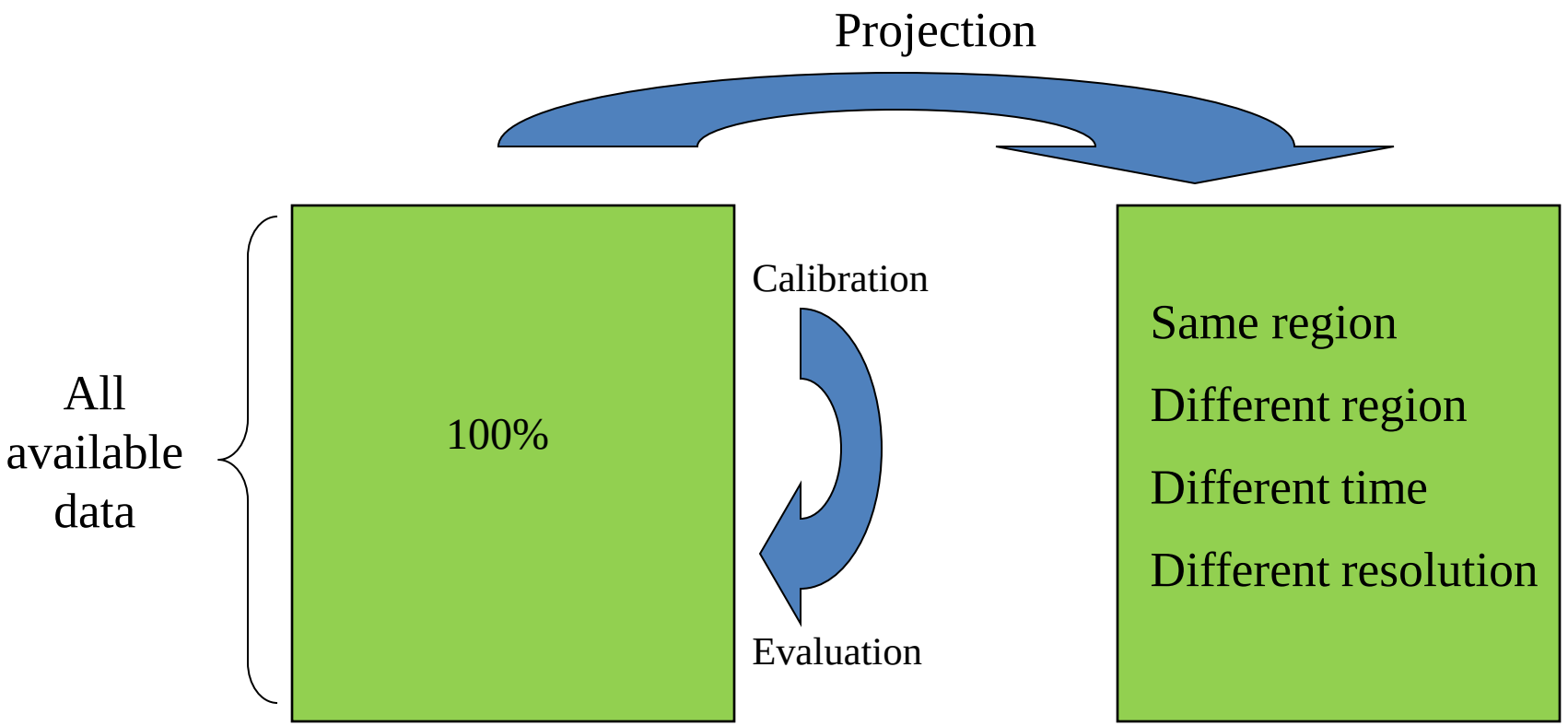
- not easily interpreted, a black box,
- prediction maps not deterministic, outputs will be different between GARP runs even when using the same occurrence data,
- generates pseudo-absences and does not allow one to substitute collected absence data,
- tendency for commission errors,
- no procedure for variable selection.

Desktop GARP output map



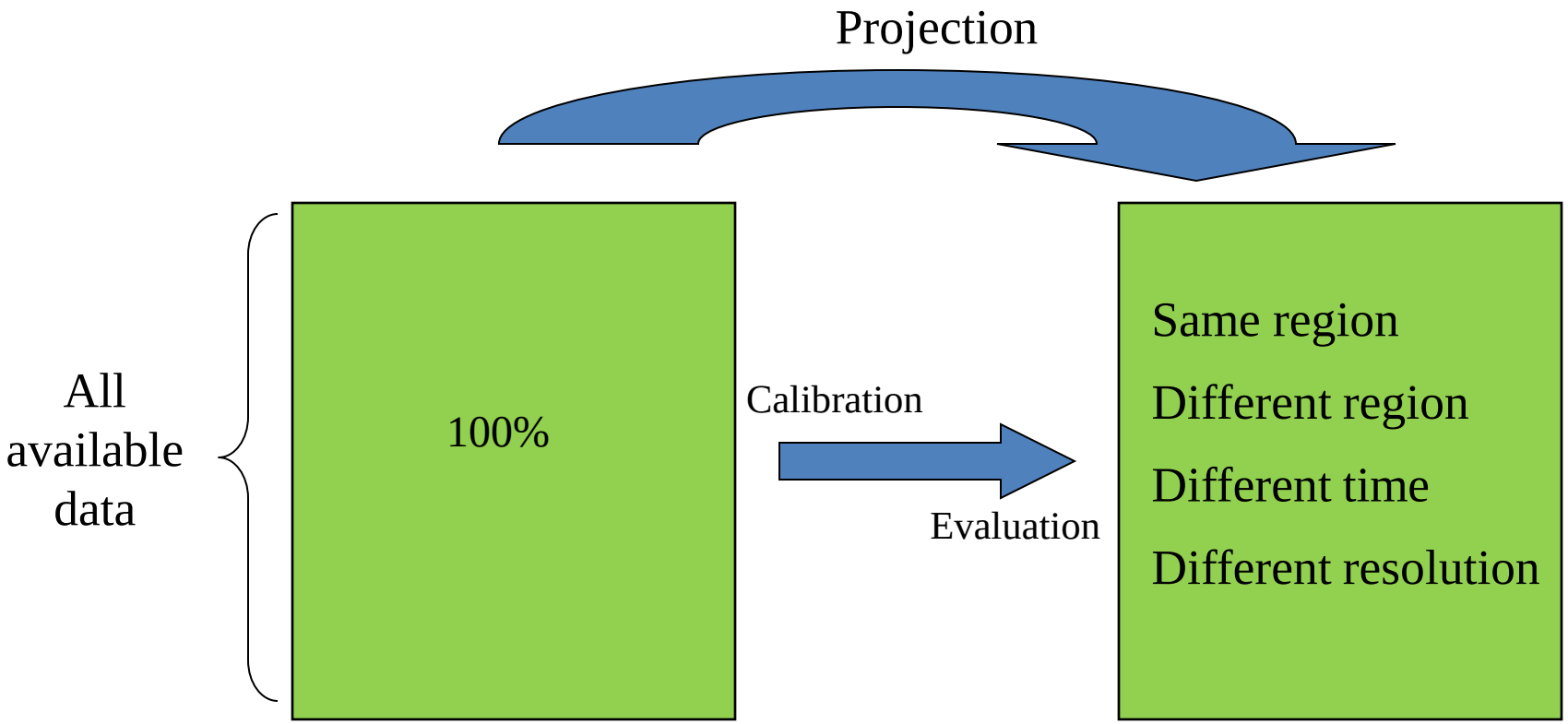
4. Evaluation de la performance des modèles

Model calibration and evaluation strategies: resubstitution



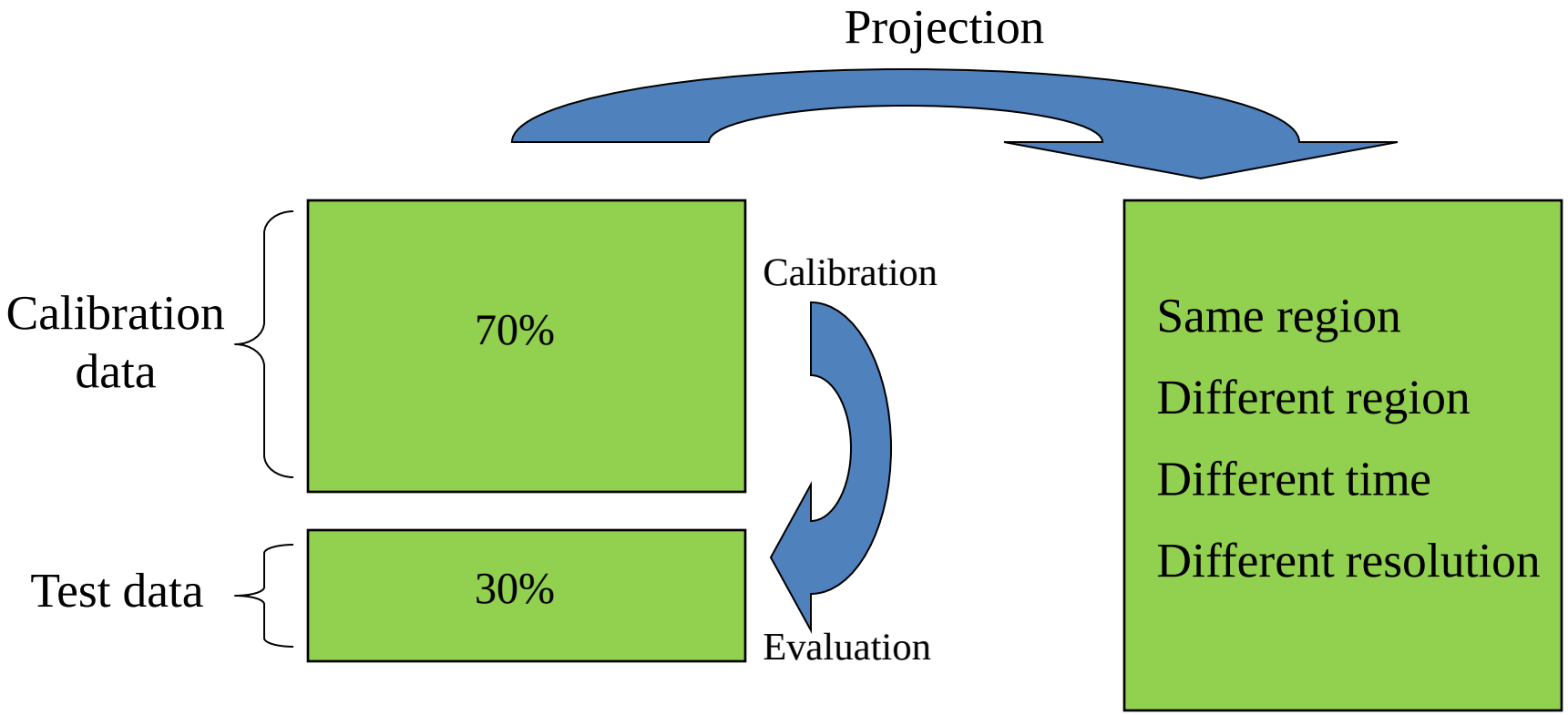
(after Araújo et al. 2005 *Gl. Ch. Biol.*)

Model calibration and evaluation strategies: independent validation



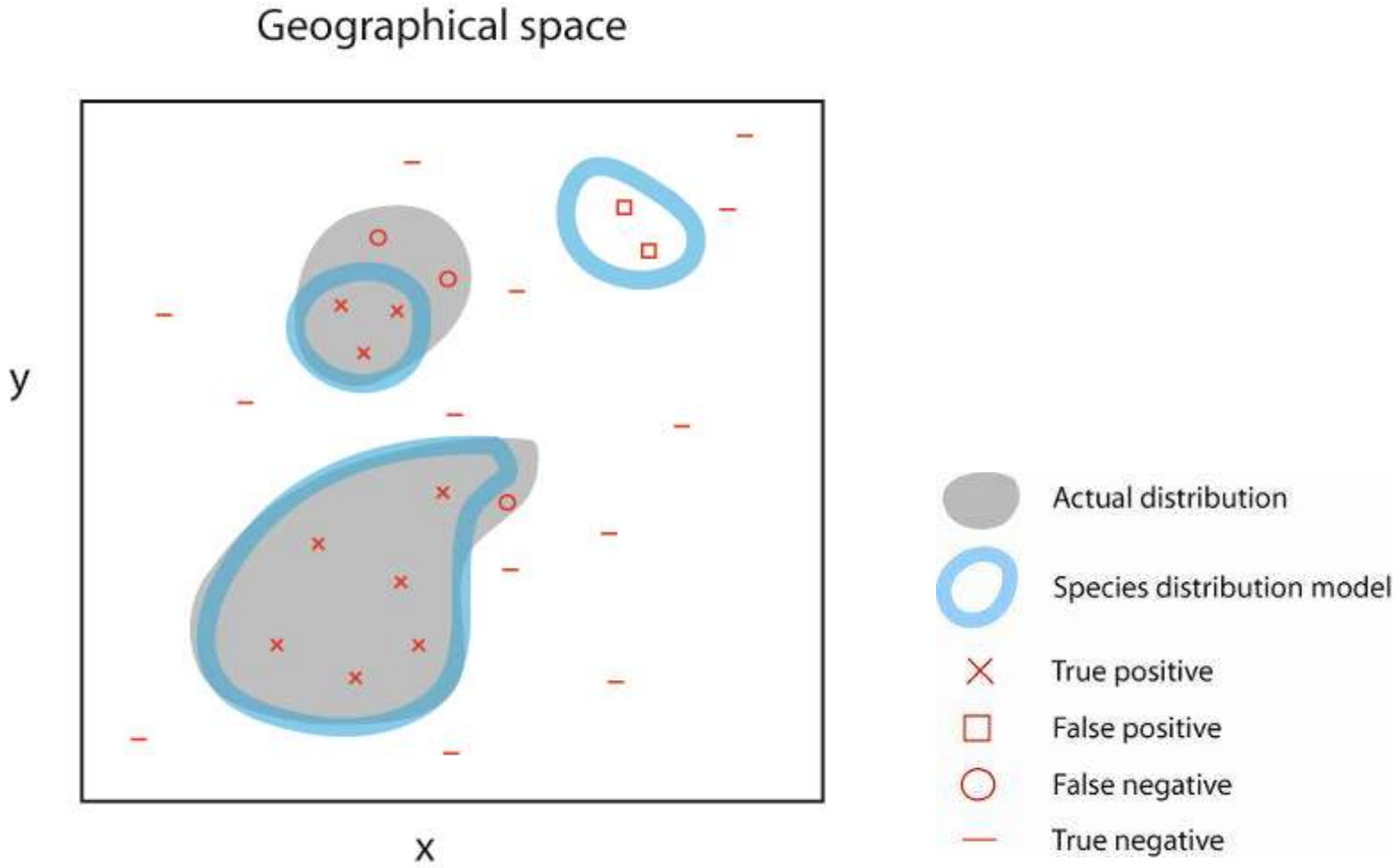
(after Araújo et al. 2005 *Gl. Ch. Biol.*)

Model calibration and evaluation strategies: data splitting



(after Araújo et al. 2005 *Gl. Ch. Biol.*)

The four types of results that are possible when testing a distribution model



Presence-absence confusion matrix

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Presence-absence test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion (%) correctly predicted (or ‘accuracy’, or ‘correct classification rate’):

$$(a + d)/(a + b + c + d)$$

Presence-absence test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Cohen's Kappa:

$$k = \frac{[(a+d)-(((a+c)(a+b)+(b+d)(c+d))/n)]}{[n-(((a+c)(a+b)+(b+d)(c+d))/n)]}$$

Presence-only test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion of observed presences correctly predicted (or ‘sensitivity’, or ‘true positive fraction’):

$$a/(a + c)$$

Presence-only test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion of observed presences correctly predicted (or ‘sensitivity’, or ‘true positive fraction’):

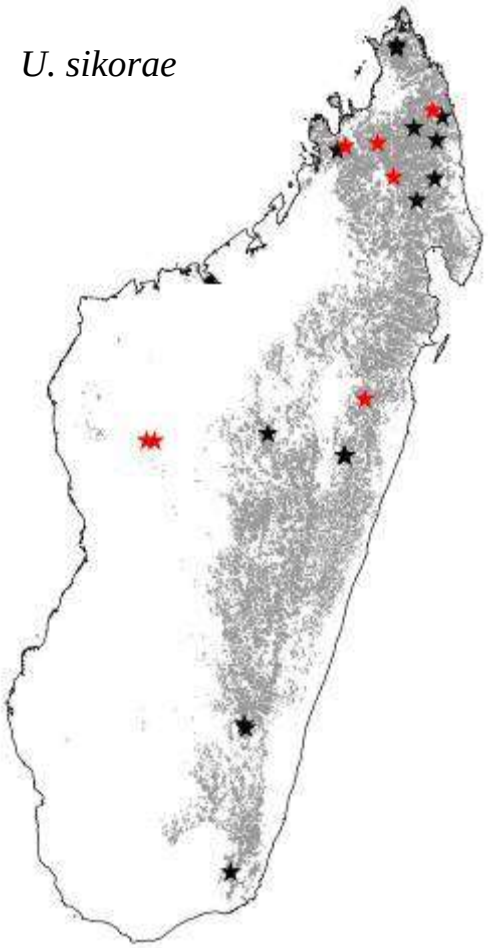
$$a/(a + c)$$

Proportion of observed presences incorrectly predicted (or ‘omission rate’, or ‘false negative fraction’):

$$c/(a + c)$$

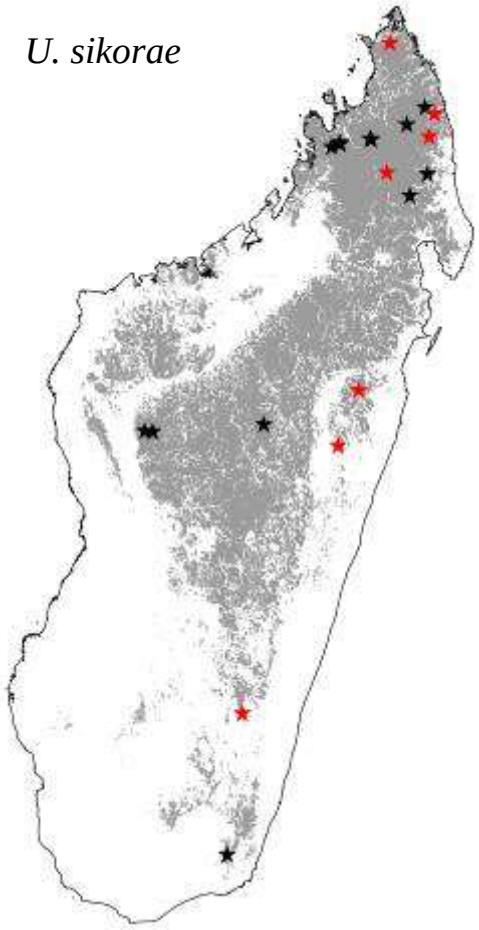
Presence-only test statistics: testing for statistical significance

U. sikorae



Success rate: 4 from 7
Proportion predicted present: 0.231
Binomial $p = 0.0546$

U. sikorae



Success rate: 6 from 7
Proportion predicted present: 0.339
Binomial $p = 0.008$

Uroplatus sp.
(leaf-tailed gecko)

Absence-only test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion of observed (or assumed) absences correctly predicted (or 'specificity', or 'true negative fraction'):

$$d/(b + d)$$

Proportion of observed (or assumed) absences incorrectly predicted (or 'commission rate', or 'false positive fraction'):

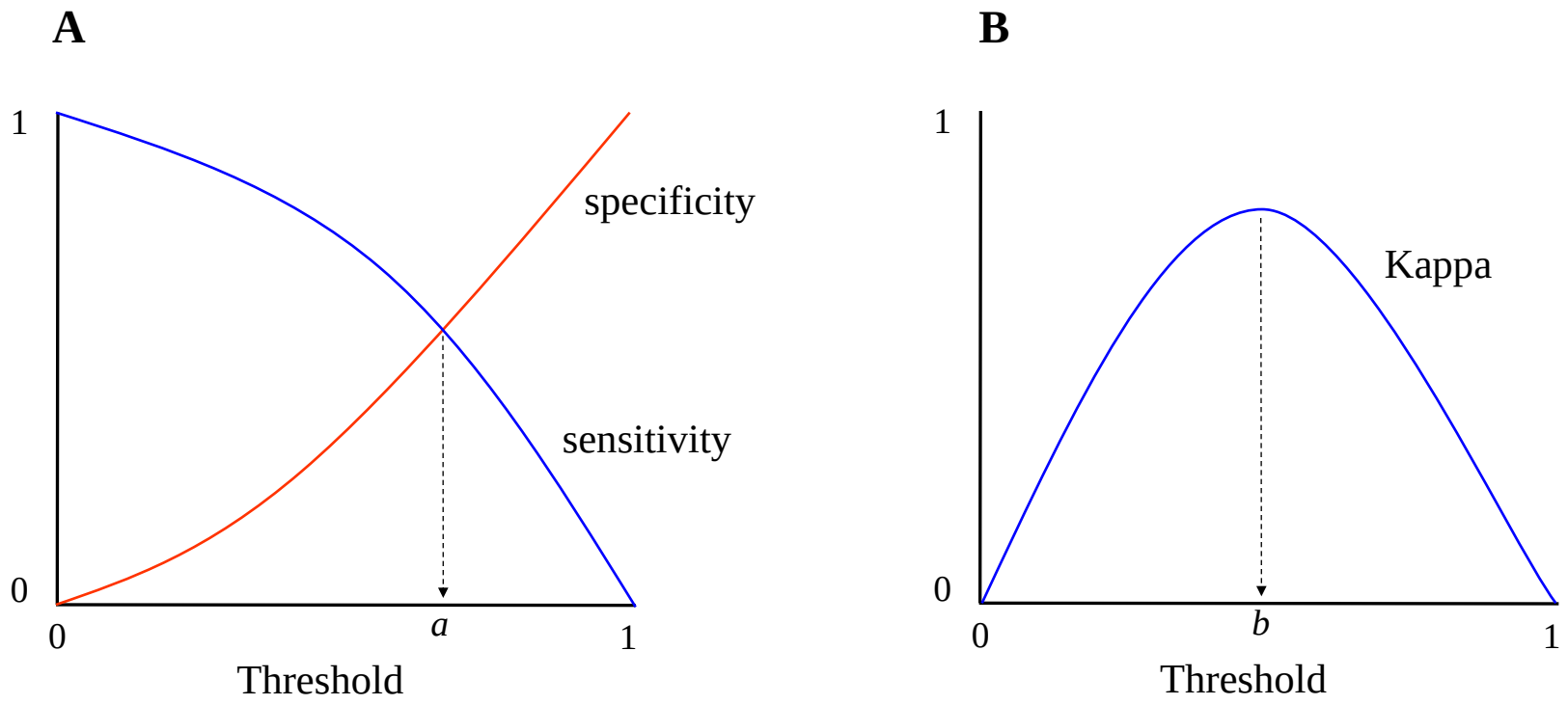
$$b/(b + d)$$

Some published methods for setting thresholds of occurrence

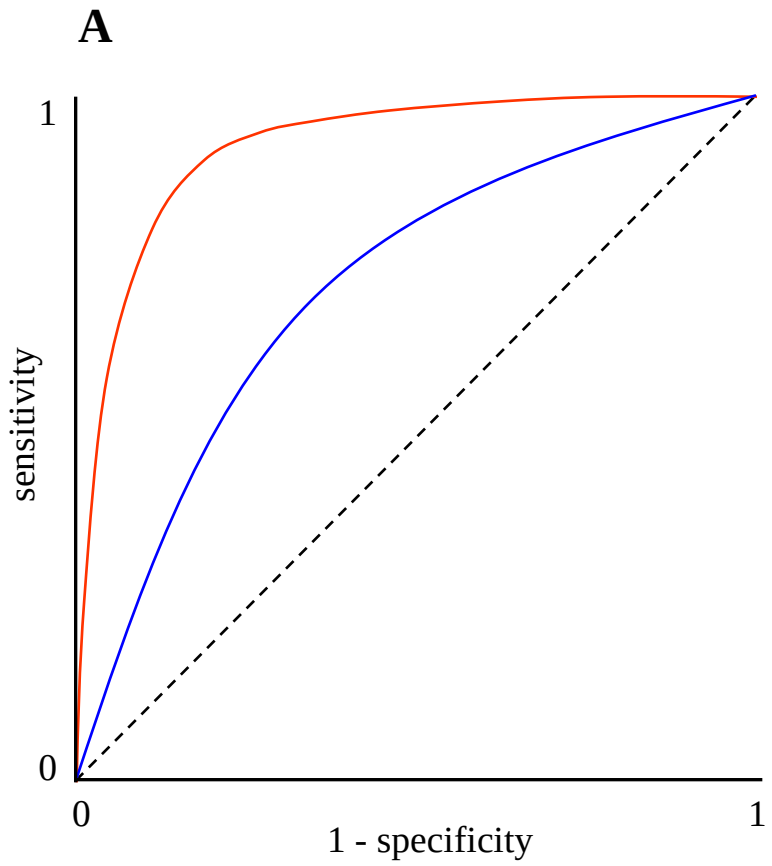
Method	Definition
Fixed value	An arbitrary fixed value (e.g. probability = 0.5)
Lowest predicted value	The lowest predicted value corresponding with an observed occurrence record
Sensitivity-specificity equality	The threshold at which sensitivity and specificity are equal
Sensitivity-specificity sum maximization	The sum of sensitivity and specificity is maximized
Maximize Kappa	The threshold at which Cohen's Kappa statistic is maximized
Equal prevalence	Species' prevalence (the proportion of presences relative to the number of sites) is maintained the same in the prediction as in the calibration data.

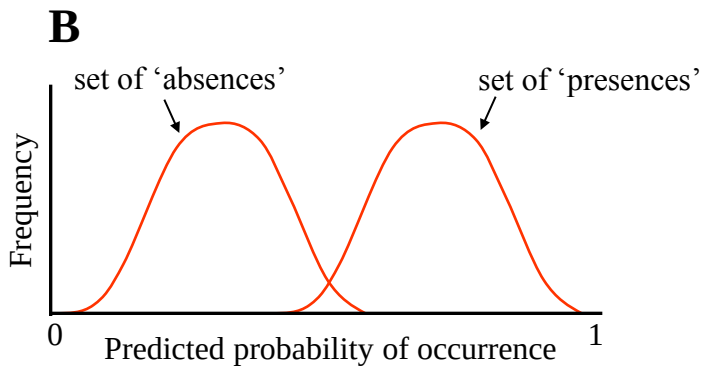
(based in part on Liu et al. 2005 *Ecography*)

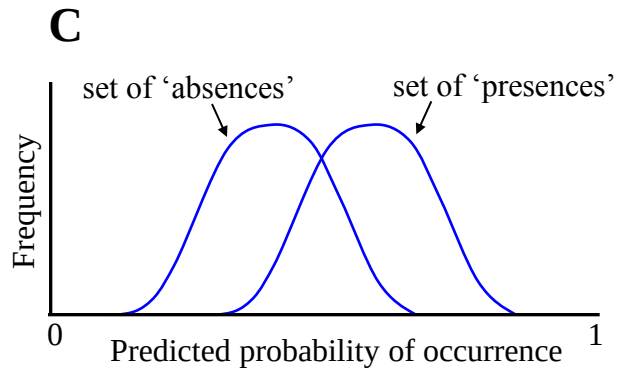
Changes in test statistics as the threshold of occurrence is adjusted



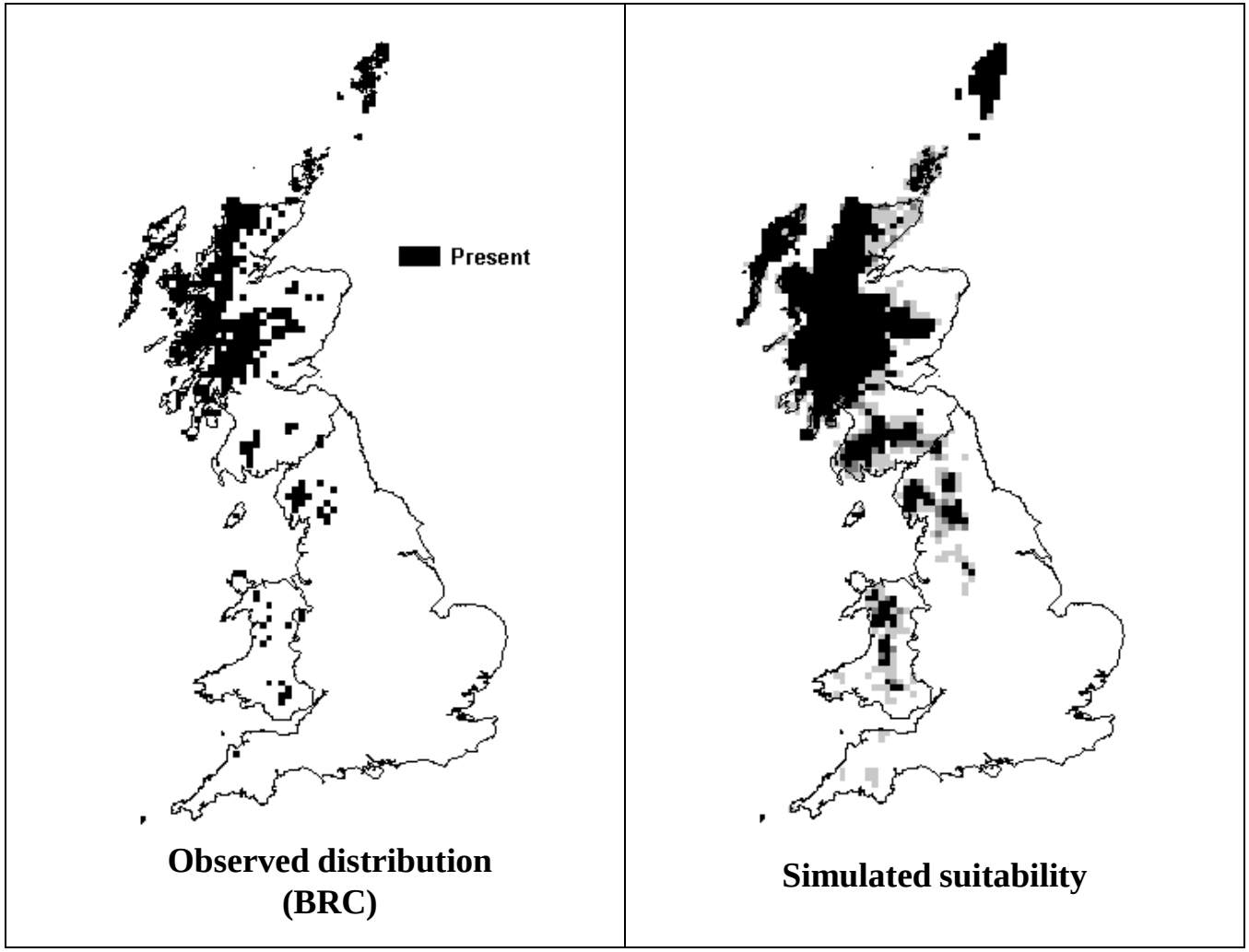
Threshold-independent assessment: The Receiver Operating Characteristic (ROC) Curve







Observed and modeled distributions of *Salix herbacea* (Dwarf Willow; 10km resolution)



Observed distribution (BRC)

Simulated suitability

AUC = 0.938; max. Kappa = 0.639

(from Pearson et al. 2004 *Ecography, Blackwell Publishing*)

So, what is a ‘good’ result?

Some *subjective* guidelines:

Kappa (after Landis & Koch 1977 *Biometrics*):

- 0 – 0.4: poor
- 0.4 – 0.75: good
- 0.75 – 1.0: excellent

AUC (after Swets 1988 *Science*):

- 0.5 – 0.7: poor discrimination
- 0.7 – 0.9: reasonable discrimination
- 0.9 – 1.0: very good discrimination

5. Etude de cas avec Maxent