# Biodiversity data analysis workshop – Day 3

Dominic Henry & Lizanne Roxburgh
Conservation Science Unit

dominich@ewt.org.za
lizanner@ewt.org.za

ENDANGERED WILDLIFE TRUST
Protecting forever, together.

science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

NRF National Research Foundation

SANBI
Biodiversity for Life
South African National Biodiversity Institute

1st November 2019

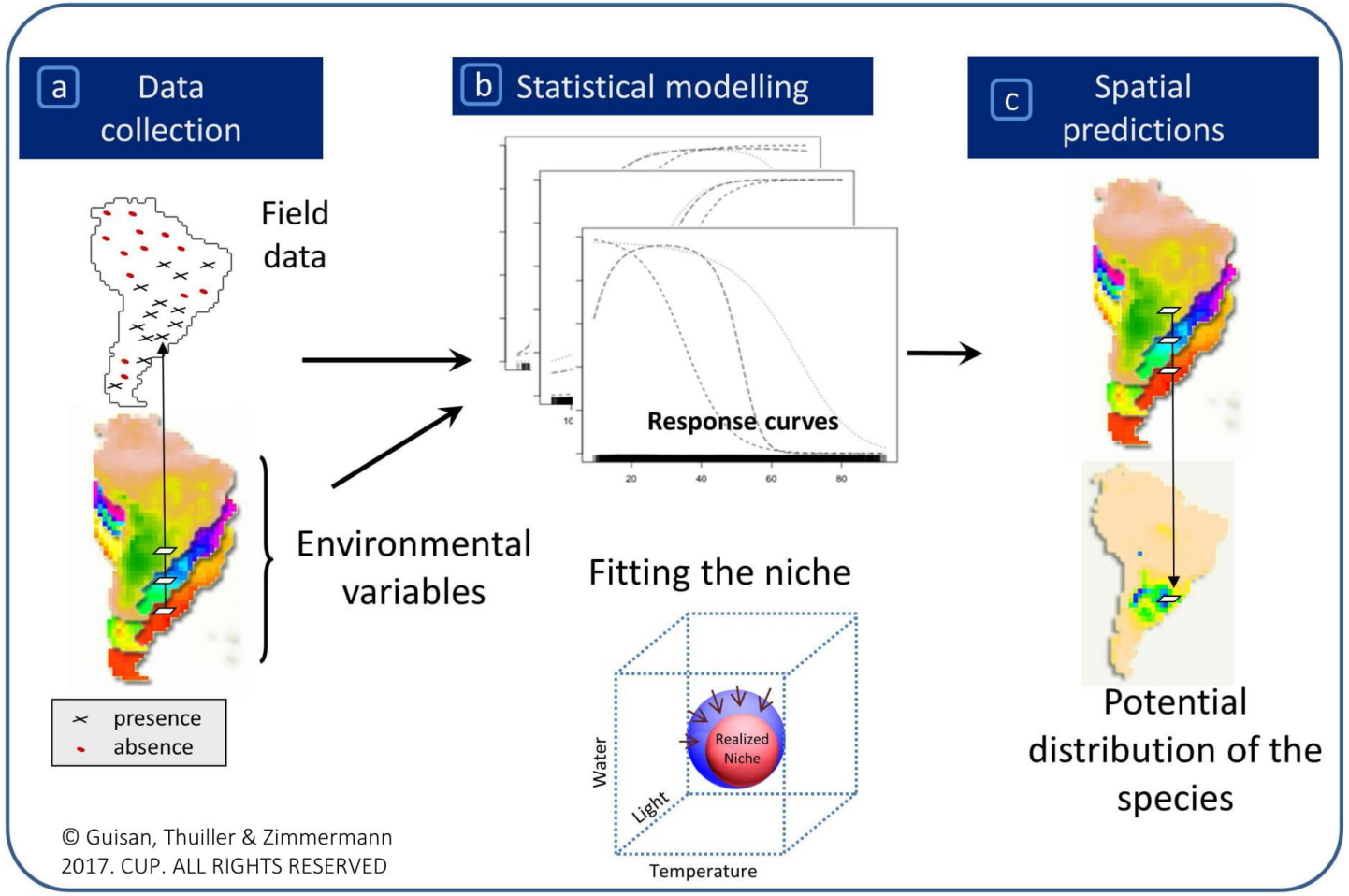
# Programme – Day 3

ENDANGERED WILDLIFE TRUST

| | |
|---|---|
| 08h30 – 10h00 | Session 1 – SDM theory and background |
| 10h00 – 10h30 | Tea break |
| 10h30 – 12h15 | Session 2 – SDM in R practical demo |
| 12h15 – 13h30 | Lunch |
| 13h30 – 15h00 | Session 3 – SDM Assignment |
| 15h00 – 15h30 | Tea break |
| 15h30 – 17h00 | Session 4 – SDM Assignment |

# Species distribution models (SDMs)

ENDANGERED WILDLIFE TRUST



# Terminology

ENDANGERED WILDLIFE TRUST

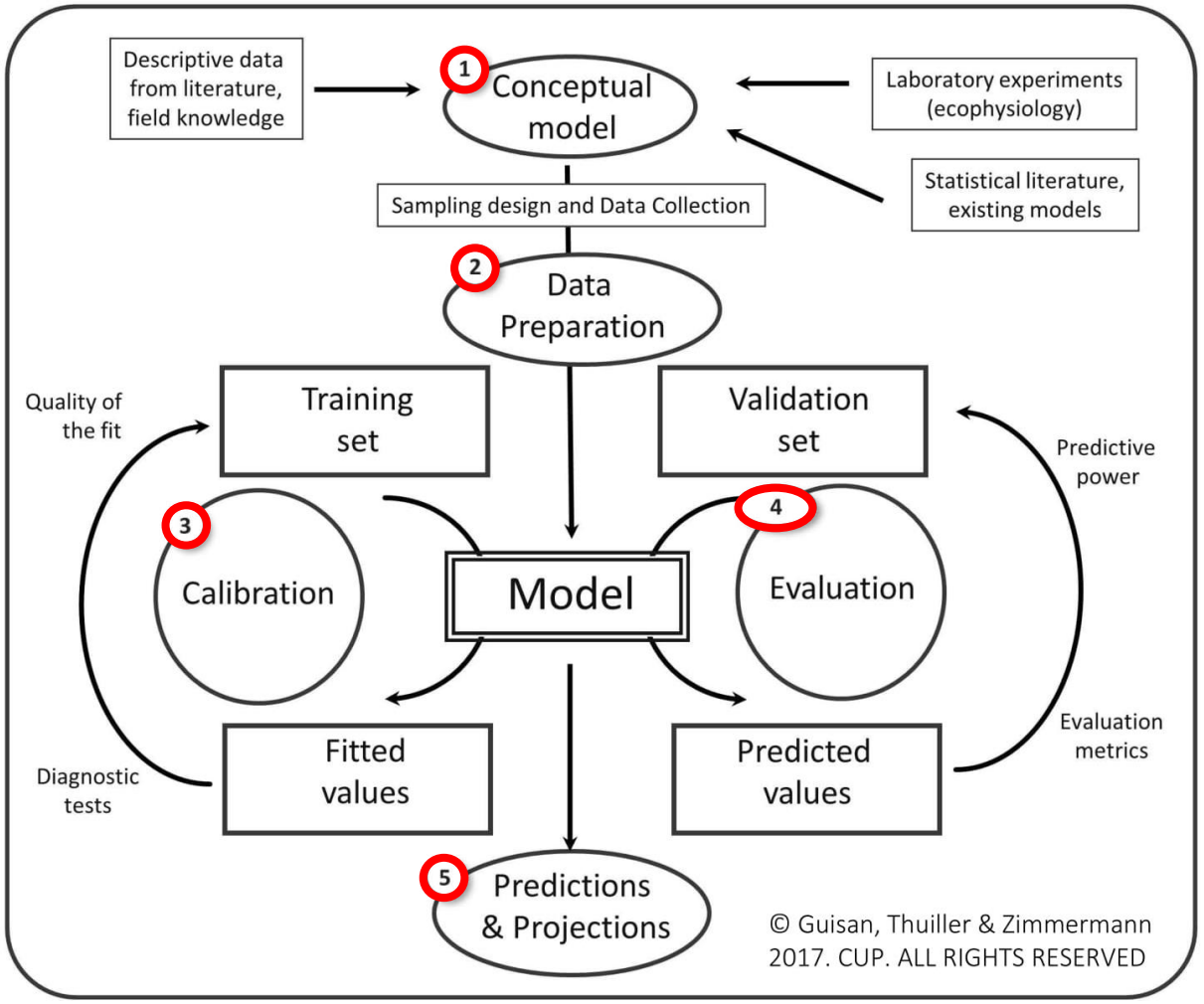
- Species distribution models
- Habitat suitability models
- Environmental niche models
- Climate envelope models
- Resource selection functions

# Fields of application

ENDANGERED WILDLIFE TRUST

- Guiding field surveys to find new populations
- Predicting invasion of non-native species
- Assessing disease risk
- Supporting conservation prioritisation and reserve selection
- Projecting the impacts of climate change on species' ranges
- Testing evolutionary theory and speciation mechanisms

# Methodological steps

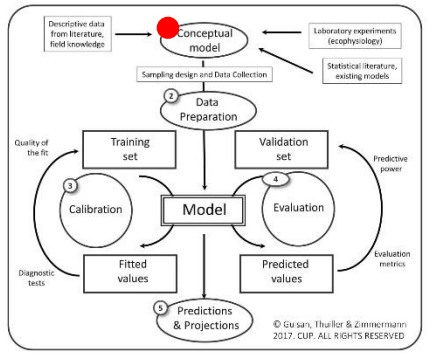


# What drives species distributions?

ENDANGERED WILDLIFE TRUST

## Fundamental biogeographic questions

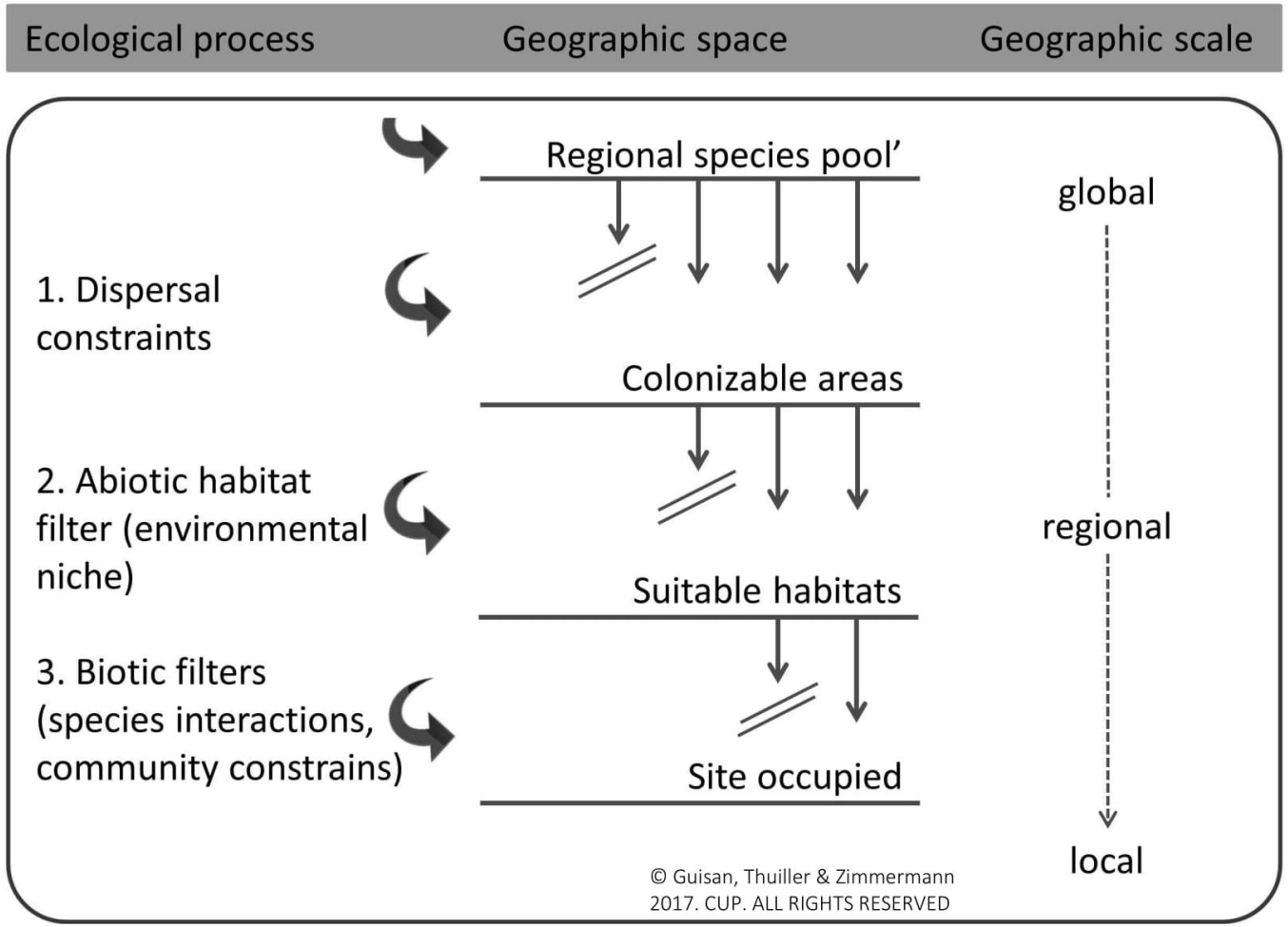
- Where, when and why?




## Three primary conditions for site occupancy

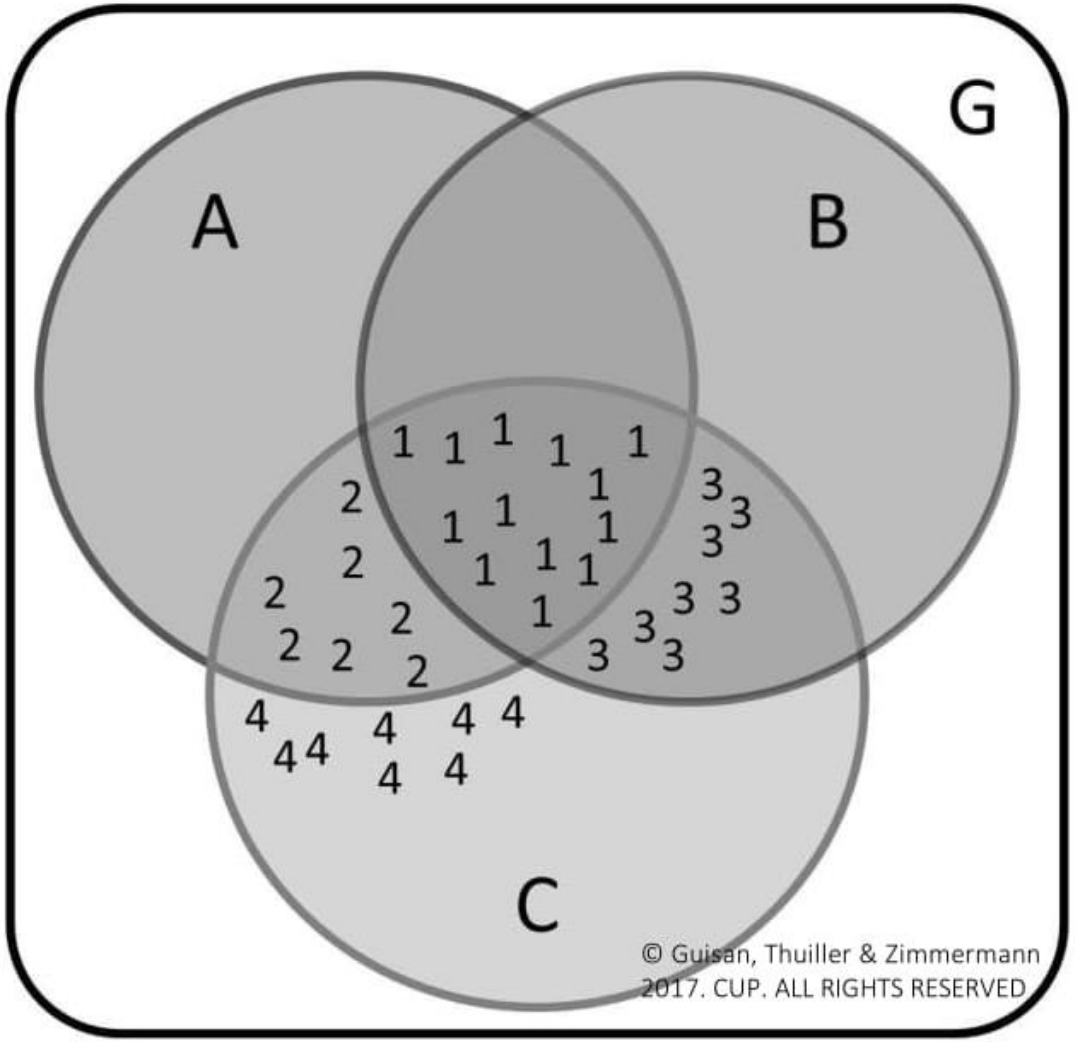
- Species must be able to reach the site
- Ecophysiology must match the abiotic conditions
- Biotic environment must be suitable

# What drives species distributions?



# What drives species distributions?

ENDANGERED WILDLIFE TRUST



# Occurrence outside of realised niche

ENDANGERED WILDLIFE TRUST

- Populations might persist in suboptimal habitats
- Source-sink dynamics
- No net population growth but high colonisation pressure
- Constant immigration from neighbouring sites

# Other sources of species absence

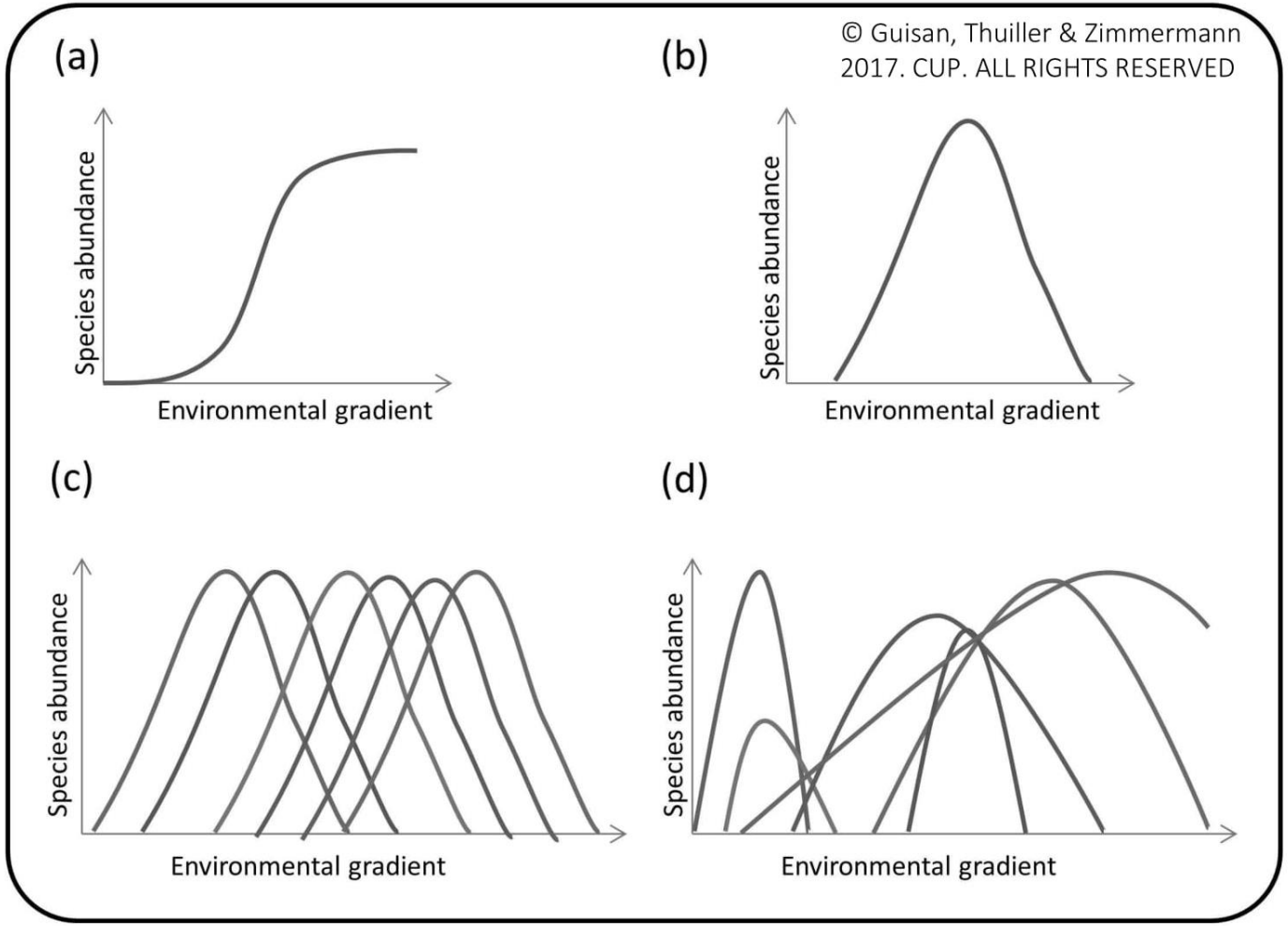
ENDANGERED WILDLIFE TRUST

- Natural or human disturbance
- Intrinsic stochasticity
- Temporal extinctions

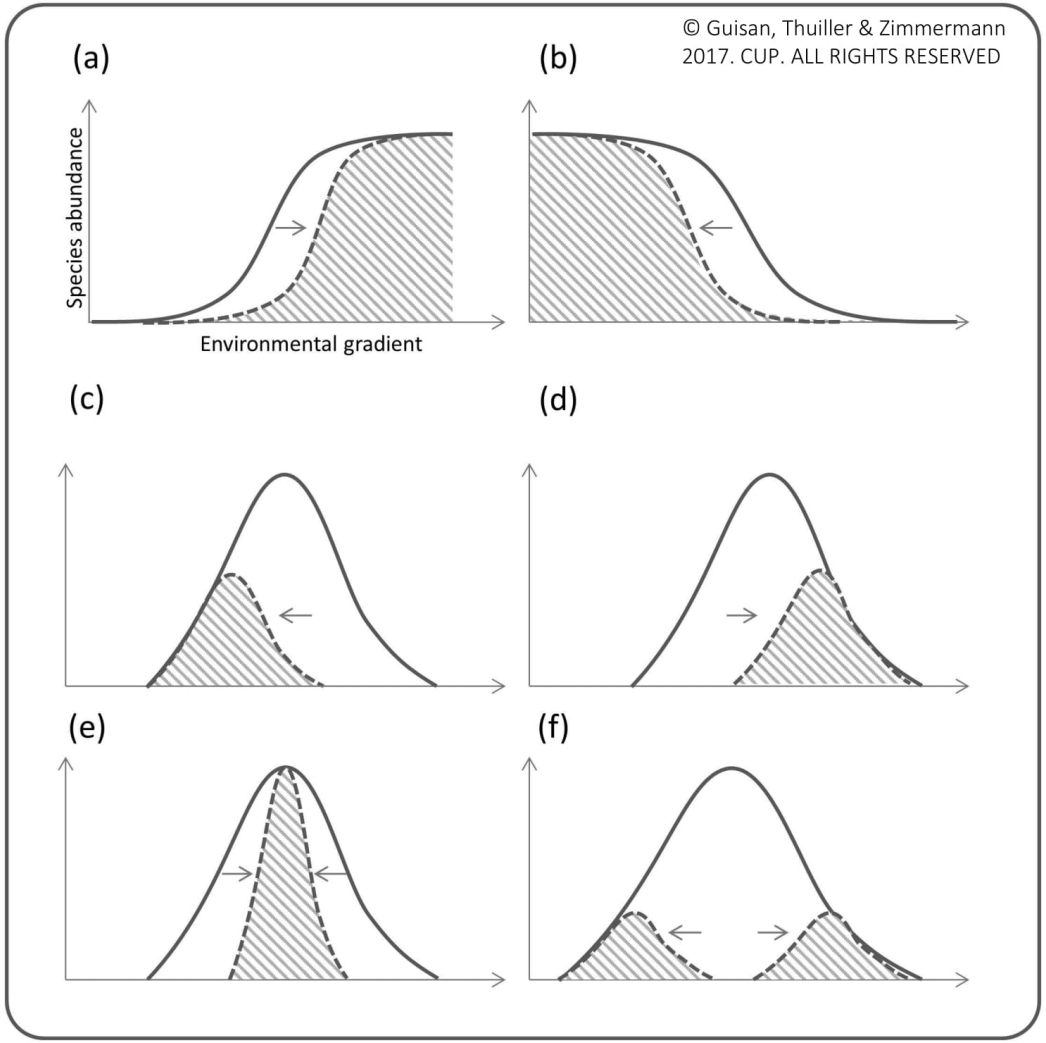
- Complications therefore arise when suitable habitats are unoccupied and unsuitable habitats are occupied (unexplained variation)

# Abiotic environment

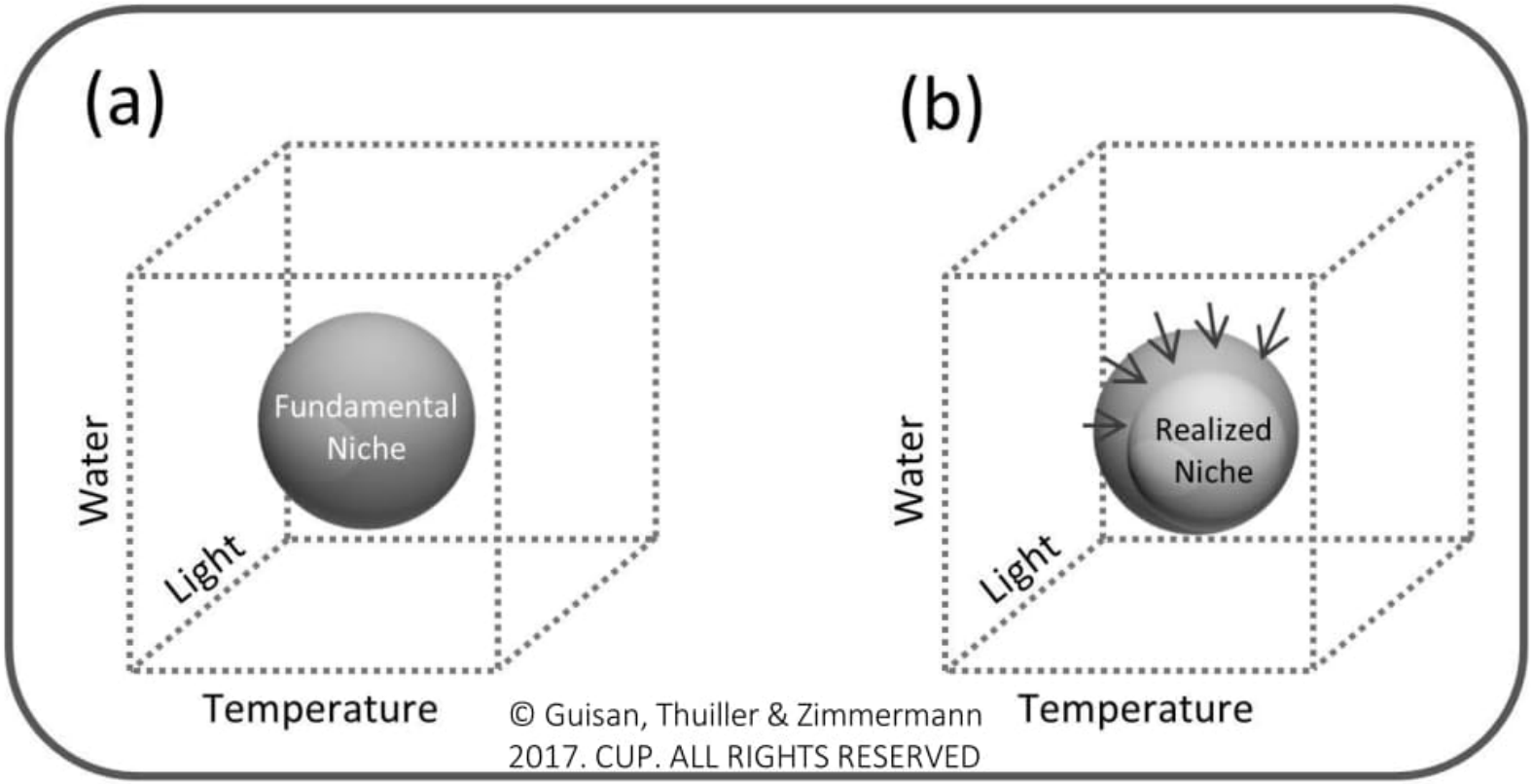
ENDANGERED WILDLIFE TRUST



# Biotic environment

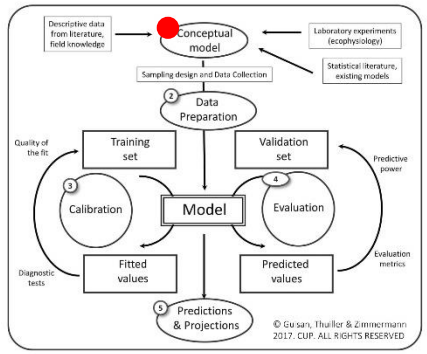


# The realised niche



# Theoretical considerations

- Species-environment equilibrium
- Comprehensive set of environmental predictors
- Appropriate species observation data



# Methodological considerations

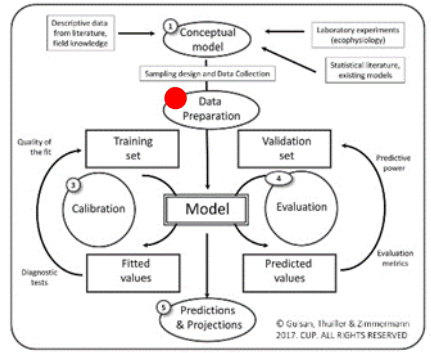
ENDANGERED WILDLIFE TRUST

- Statistical model matches data (categorical, ordinal, continuous)
- Measurement error in environmental predictors
- Bias in species data
- Independence of species occurrence data points

# Data sources - predictors

ENDANGERED WILDLIFE TRUST

| Category | Example | Source |
|---|---|---|
| Climatic | Rainfall, Seasonality | Bioclim - https://www.worldclim.org/bioclim |
| | Temperature, Aridity | ENVIREM - https://envirem.github.io/ |
| Topographic | Elevation, Aspect, Slope, Terrain Ruggedness | USGS earth explorer/aster global DEM https://asterweb.jpl.nasa.gov/gdem.asp |
| Land cover | Natural, Agricultural | http://bgis.sanbi.org/landcover/project.asp |
| Primary productivity | NDVI, NDWI | https://modis.gsfc.nasa.gov/data/dataprod/mod13.php |
| | | |



# Bioclim

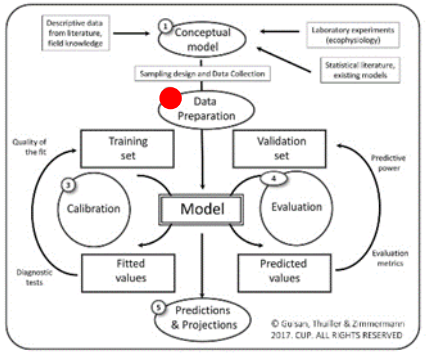
ENDANGERED WILDLIFE TRUST

| | |
|---|---|
| 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 |

# Data sources - predictors

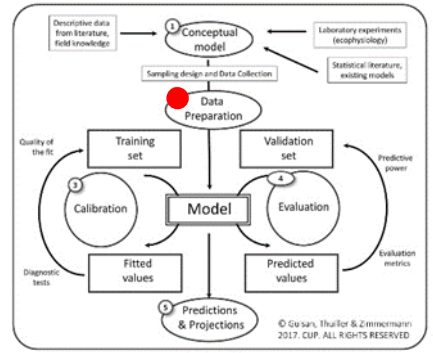
ENDANGERED WILDLIFE TRUST

| Category | Example | Source |
|---|---|---|
| Soils | pH, Structure, Organic content | Prof Mike Cramer (UCT) |
| Vegetation | Biomes, vegetation classes, Bioregions | http://bgis.sanbi.org/SpatialDataset |
| Climate change | Various scenarios | https://gisclimatechange.ucar.edu/ |
| Marine | Oceanic properties | NOAA - https://www.ncdc.noaa.gov/data-access/marineocean-data |



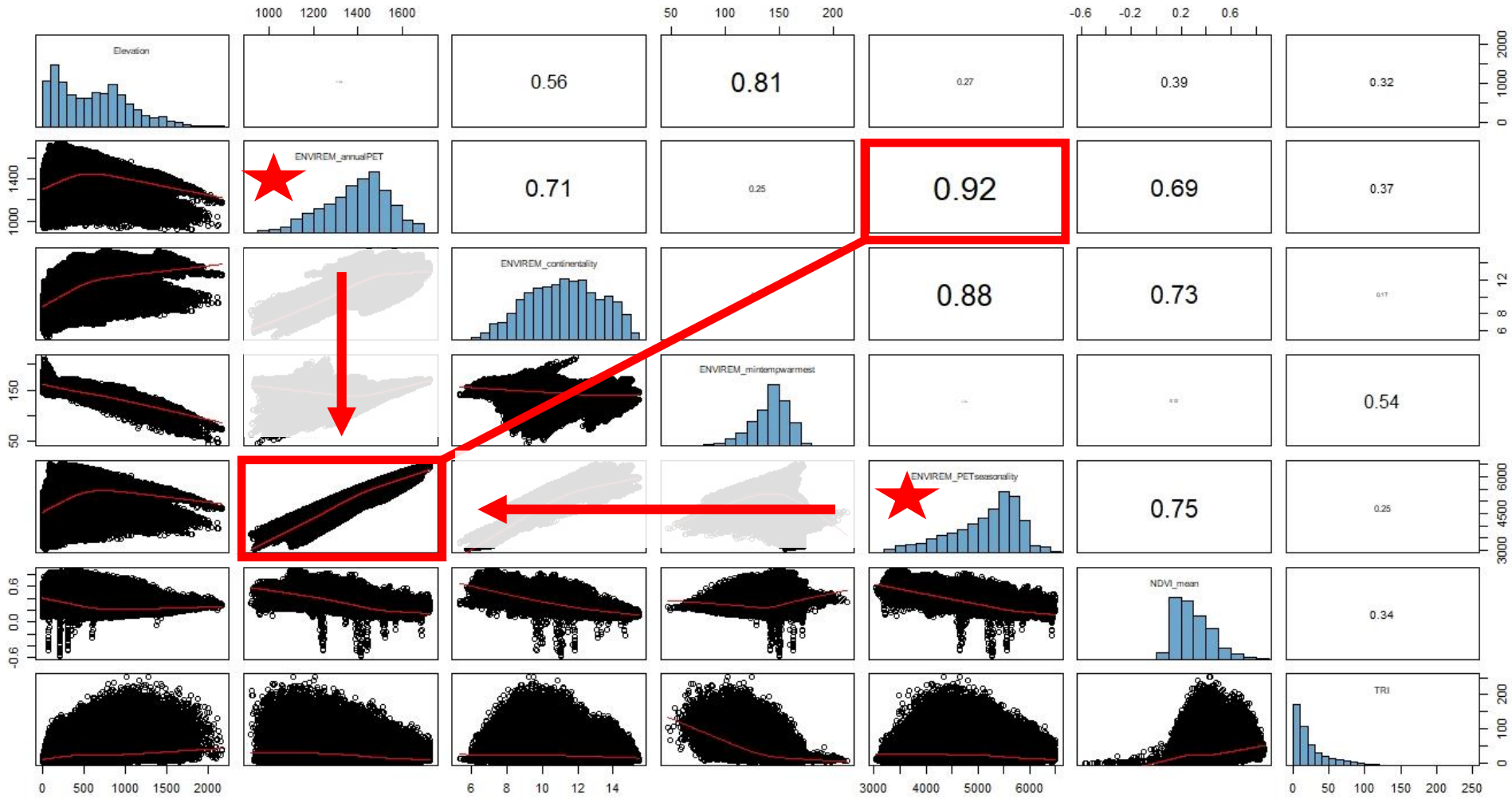
# Data sources - species

| | |
|---|---|
| iNaturalist | https://www.inaturalist.org/ |
| GBIF | https://www.gbif.org/ |
| SANBI | http://biodiversityadvisor.sanbi.org/online-biodiversity-data/ |
| Animal Demography Unit | http://www.adu.uct.ac.za/ |
| Map of life | https://mol.org/ |
| LifeMapper | http://lifemapper.org/ |
| AmphibiaWeb | https://amphibiaweb.org/ |
| BirdLife | http://datazone.birdlife.org/home |
| Plants – BIEN | http://bien.nceas.ucsb.edu/bien/ |
| National collection facilities | |
| Field surveys | |



# Predictor correlation

ENDANGERED WILDLIFE TRUST



# Predictor multi-collinearity

ENDANGERED WILDLIFE TRUST

- Multi-collinearity & variance inflation factor (VIF)
- Rule of thumb VIF < 10

```
> usdm::vif(env_stack)
                  Variables        VIF
1                 Elevation  9.407335
2       ENVIREM_annualPET 12.741937
3 ENVIREM_continentality 13.318838
4 ENVIREM_mintempwarmest  6.216536
5 ENVIREM_PETseasonality 23.775232
6                 NDVI_mean  3.169974
7                       TRI  1.835674
```

# Predictor multi-collinearity

ENDANGERED WILDLIFE TRUST

```
> usdm::vifcor(env_stack, th = 0.6)
4 variables from the 7 input variables have collinearity problem:

ENVIREM_PETseasonality Elevation NDVI_mean ENVIREM_continentality

After excluding the collinear variables, the linear correlation coefficients ranges between:
min correlation ( ENVIREM_mintempwarmest ~ ENVIREM_annualPET ):  0.2117568
max correlation ( TRI ~ ENVIREM_mintempwarmest ):  -0.4925304

---------- VIFs of the remained variables --------
                Variables      VIF
1      ENVIREM_annualPET 1.143247
2 ENVIREM_mintempwarmest 1.323283
3                    TRI 1.441725
```

# Spatial resolution and extent

ENDANGERED WILDLIFE TRUST

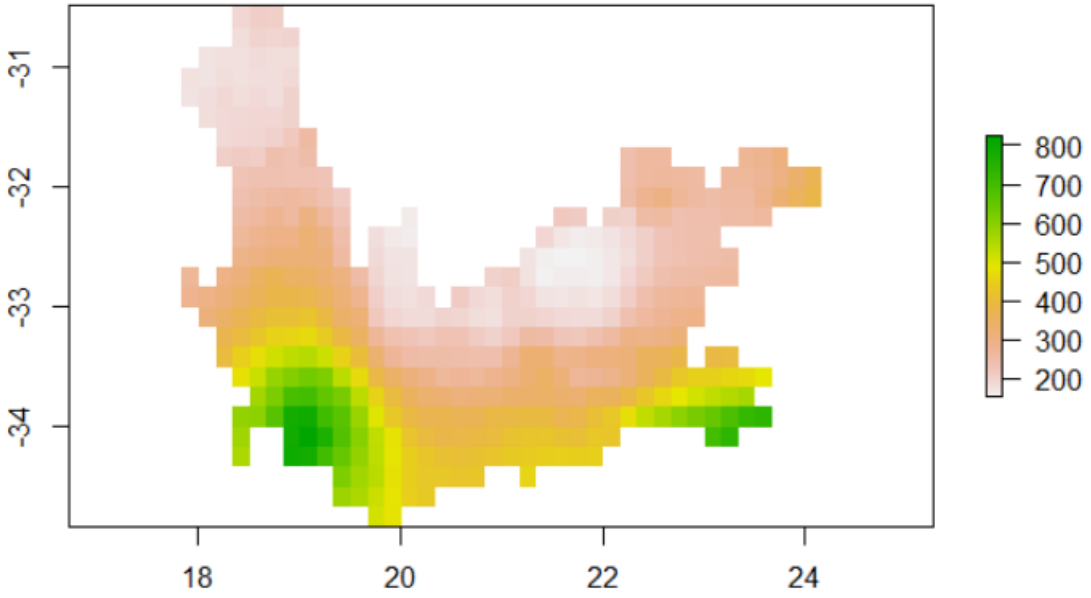
Common spatial resolutions (Bioclim and ENVIREM):

- 30 arc-seconds (~ 1 km)
- 2.5 arc-minutes (~ 5 km)
- 5 arc-minutes (~ 10 km)
- 10 arc-minutes (~ 20 km)

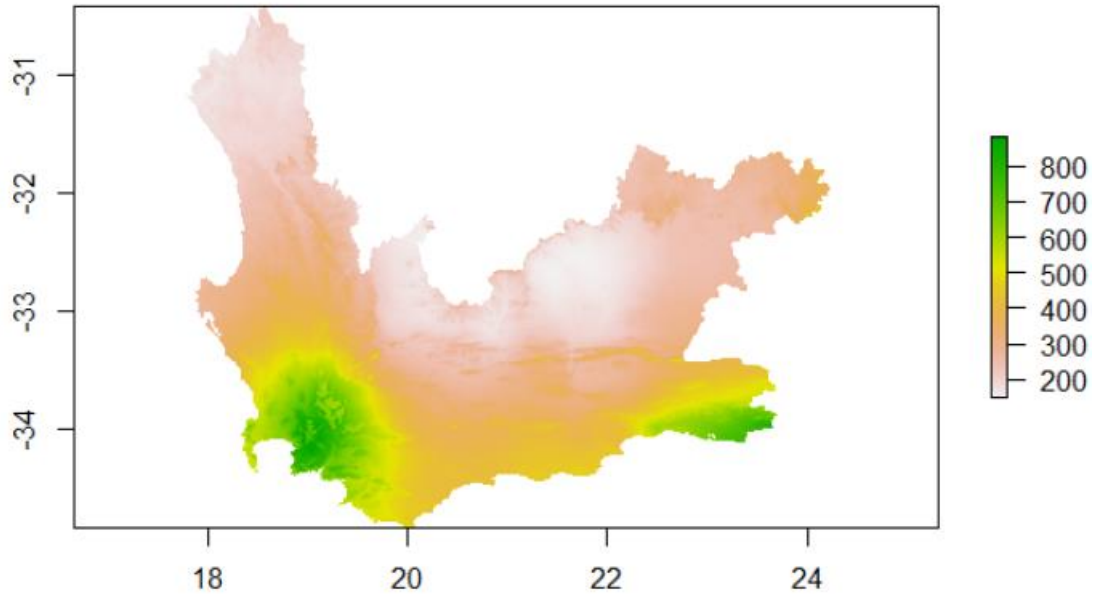
# Spatial resolution and extent

ENDANGERED WILDLIFE TRUST

## 10 arc-minutes




## 30 arc-seconds

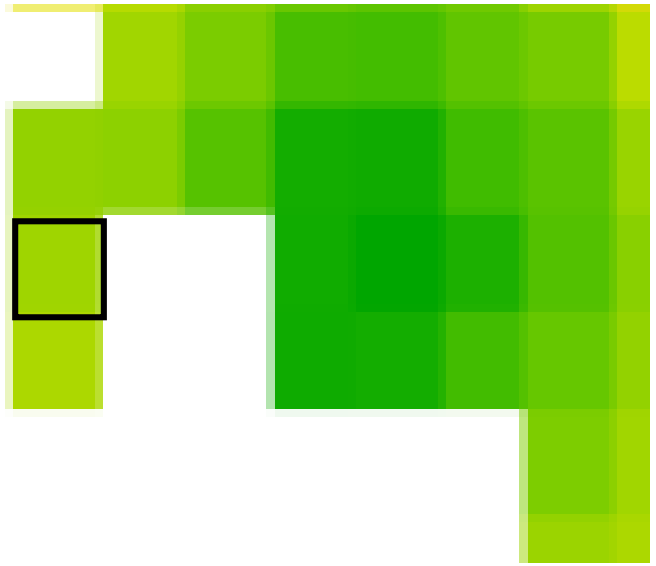


# Spatial resolution and extent

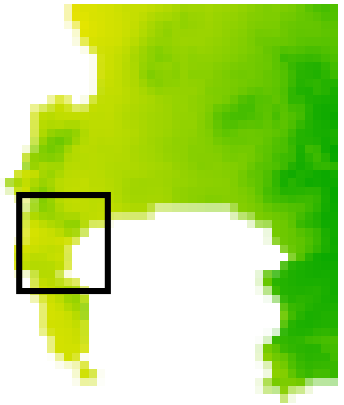
ENDANGERED WILDLIFE TRUST

10 arc-minutes

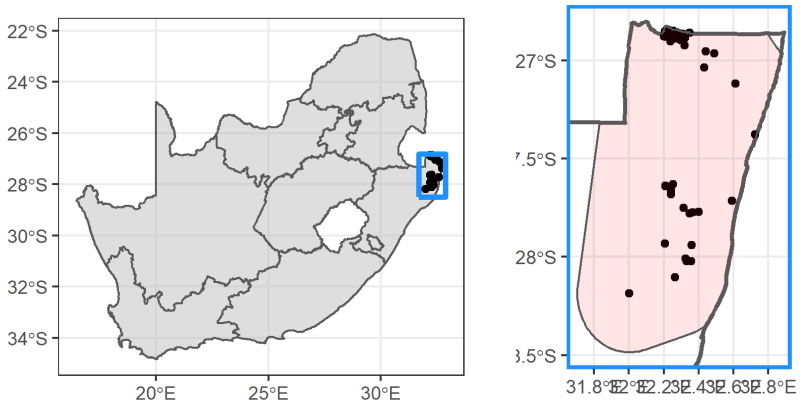
30 arc-seconds

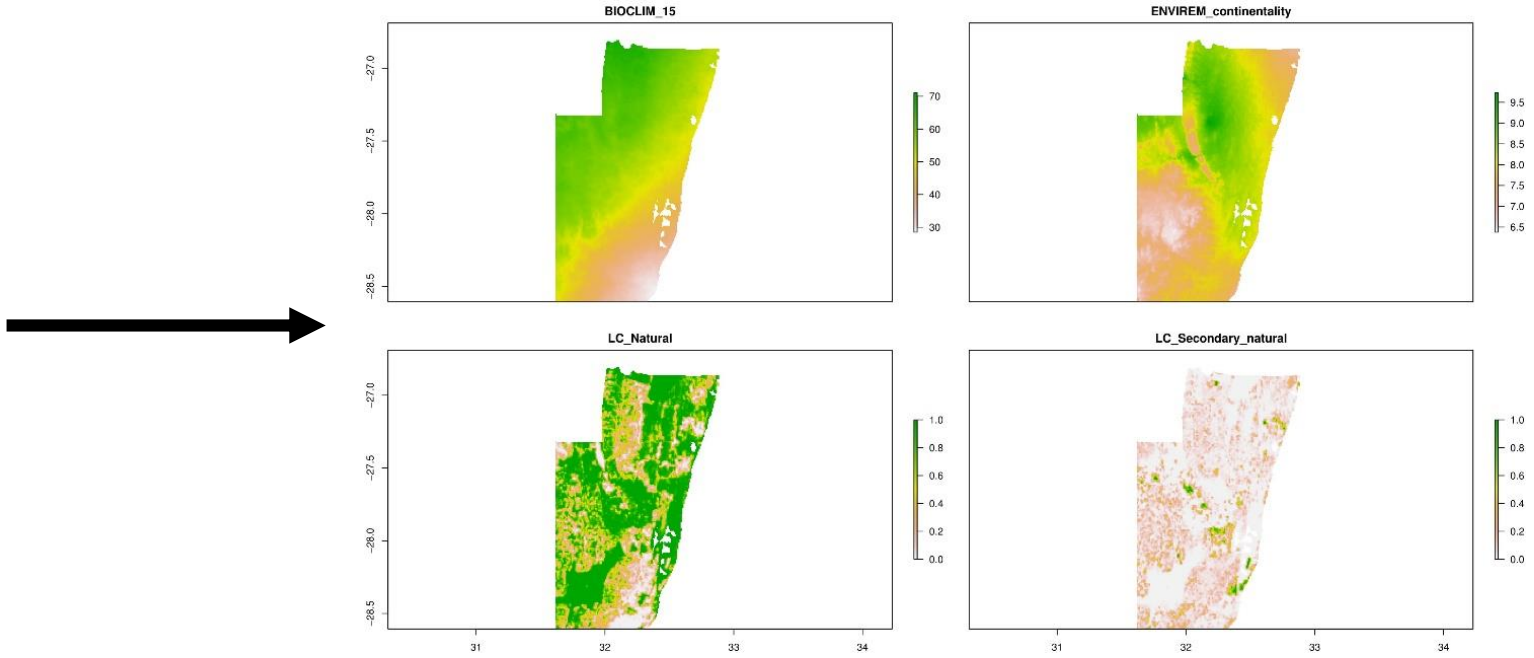


Average pixel value



# Spatial resolution and extent



# Species data processing

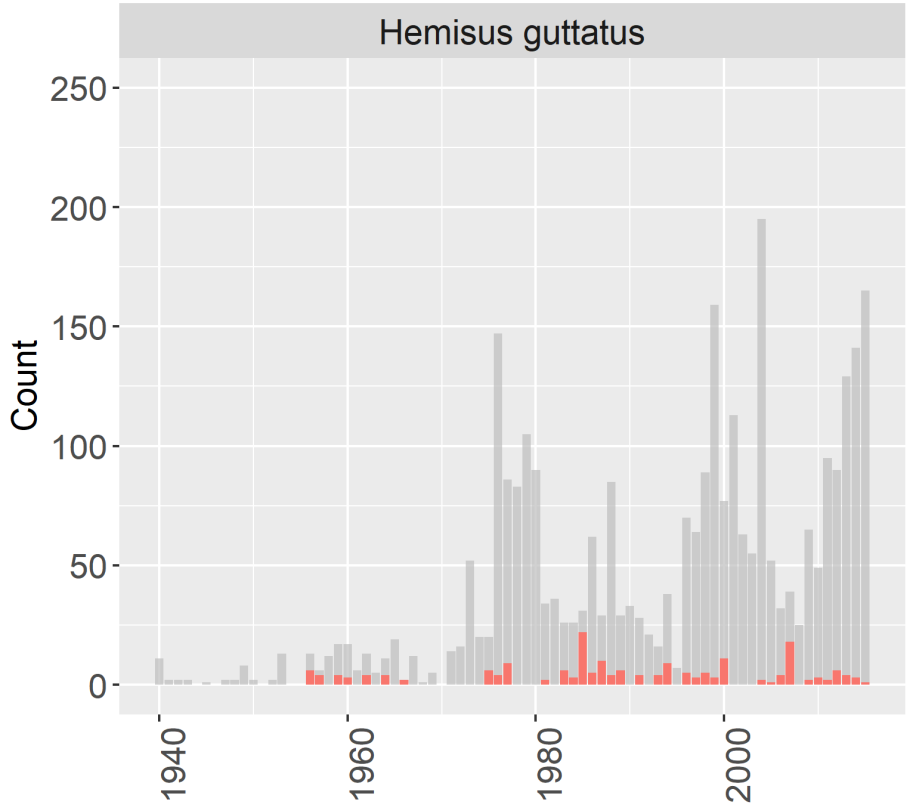
ENDANGERED WILDLIFE TRUST

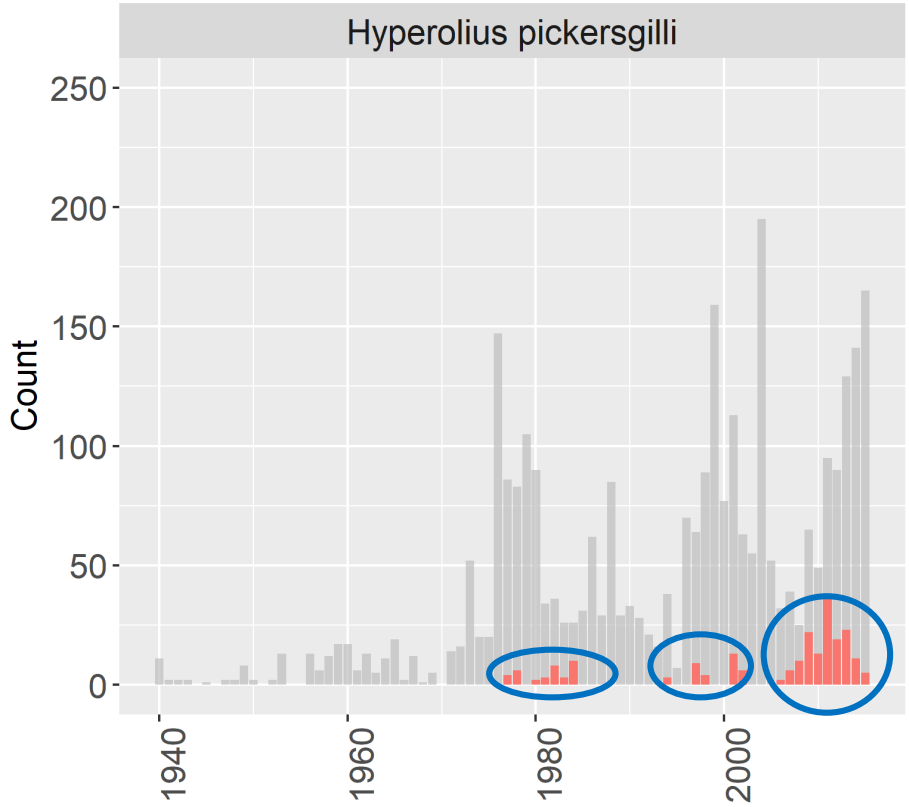
- False positives

# Species data processing

ENDANGERED WILDLIFE TRUST

- Spatial and/or temporal bias

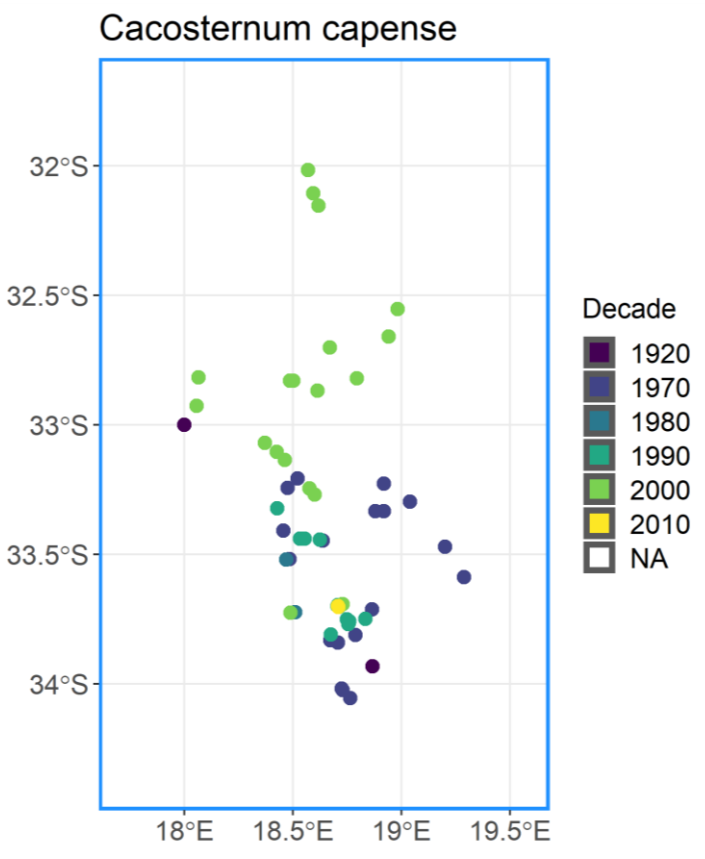


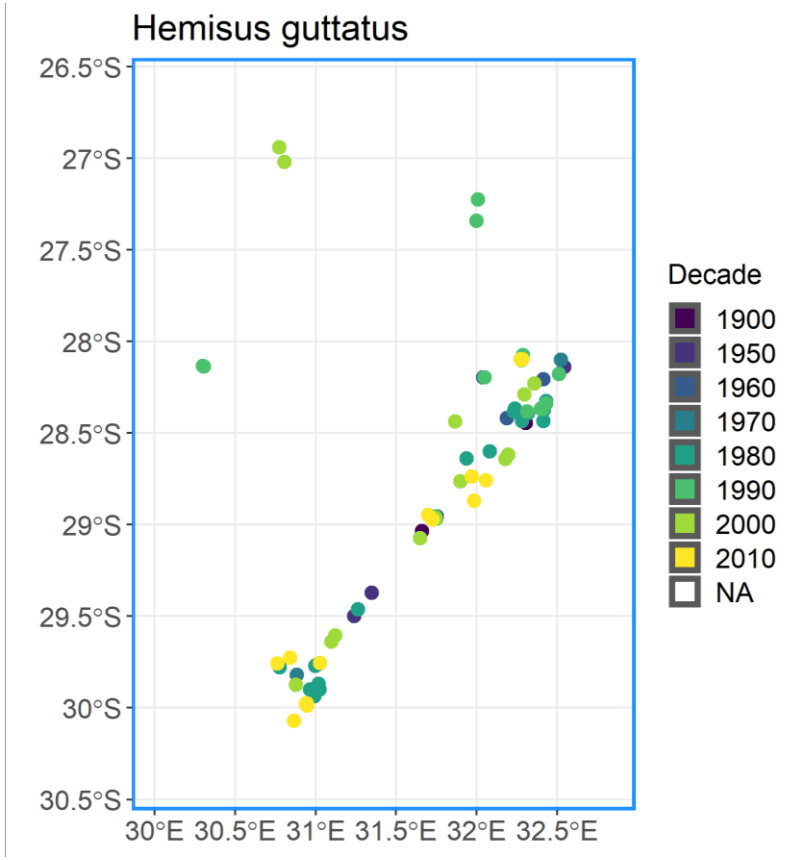



# Species data processing

ENDANGERED WILDLIFE TRUST

- Spatial and/or temporal bias



# Species data processing

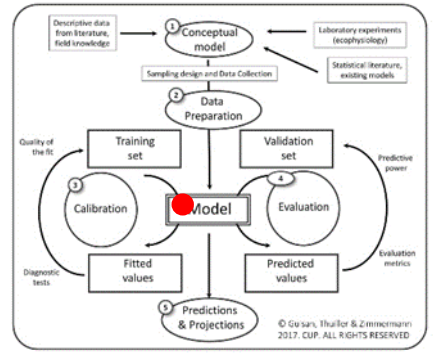
ENDANGERED WILDLIFE TRUST

- False positives
- Spatial and/or temporal bias
- Spatial autocorrelation
- Pseudo-replication (duplicate records)
- Positional uncertainty (georeferencing error)

# Modelling methods

ENDANGERED WILDLIFE TRUST

| Method | Software |
|---|---|
| Maximum Entropy | MAXENT standalone<br>*dismo* R package |
| Generalized linear model (GLM) | Base R |
| Generalized additive model (GAM) | *mcgv* R package |
| Boosted regression trees (BRT) | *dismo* R package |
| Multivariate adaptive regression splines (MARS) | *earth* R package |
| Ensemble Models | BIOMOD standalone<br>*biomod2* R package |

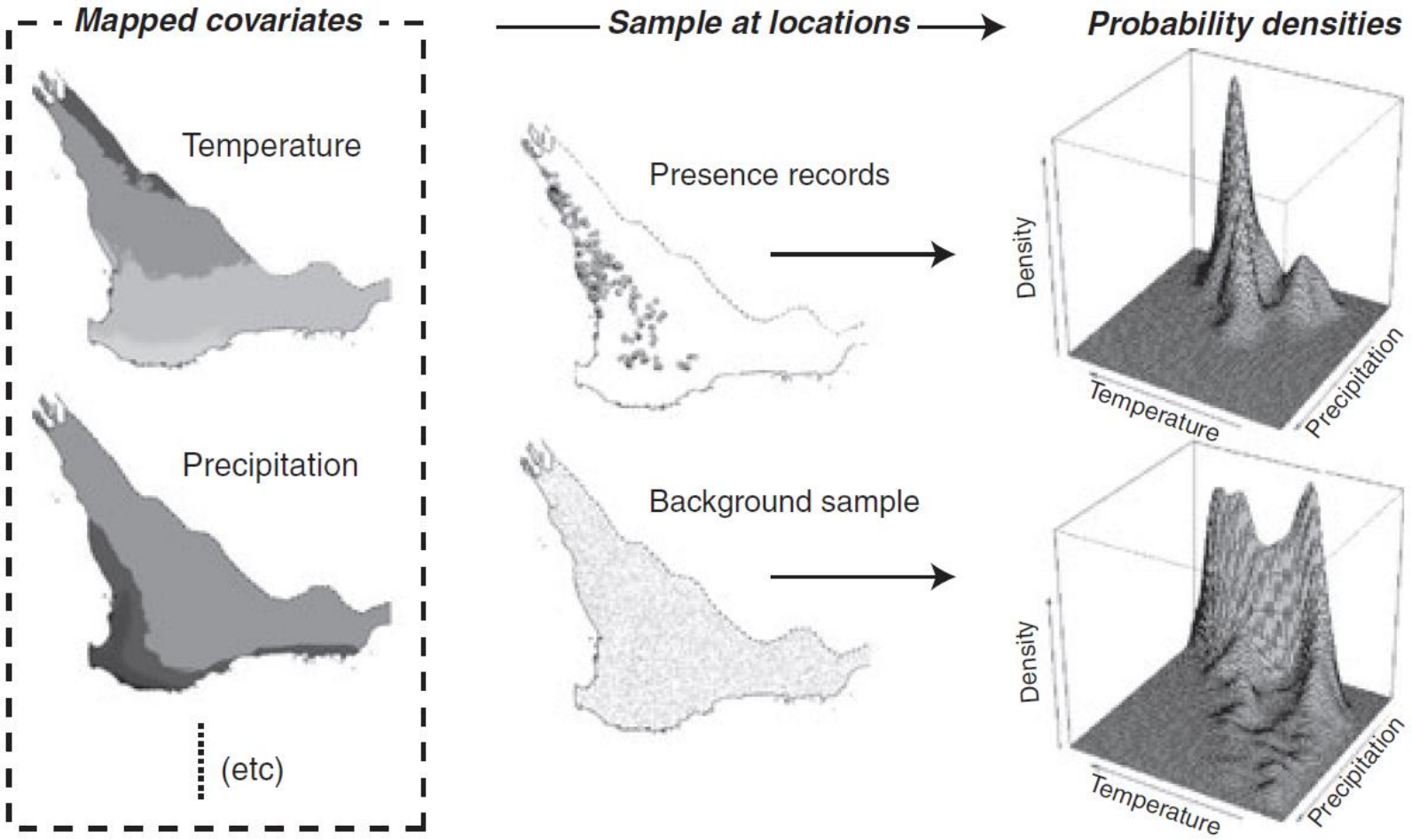


# Maxent in R

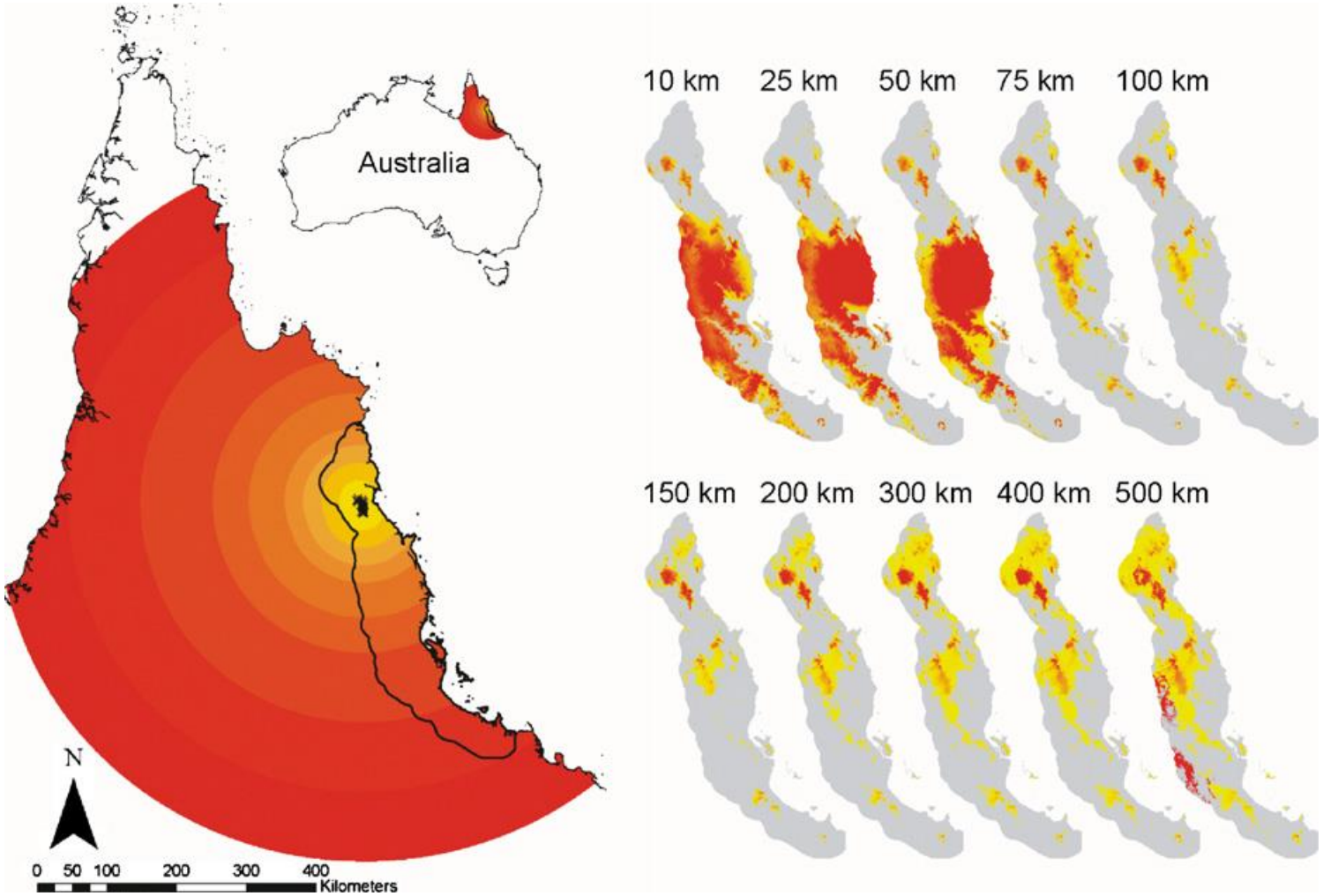
ENDANGERED WILDLIFE TRUST

- MaxEnt takes a list of species presence locations as input (often called presence-only (PO) data)
- Additionally a set of environmental predictors (e.g. precipitation, temperature) across a user-defined landscape that is divided into grid cells.
- From this landscape, MaxEnt extracts a sample of background locations that it contrasts against the presence locations.
- Presence is unknown at background locations.
- Then estimates a relative occurrence rate (or habitat suitability)

# Maxent in R


Mapped covariates
Temperature
Precipitation
(etc)
Sample at locations
Presence records
Background sample
Probability densities
Density
Temperature
Precipitation
Density
Temperature
Precipitation

# Background points



# Background points

ENDANGERED WILDLIFE TRUST

**Methods in Ecology and Evolution**

British Ecological Society

*Methods in Ecology and Evolution* 2012, **3**, 327–338 doi: 10.1111/j.2041-210X.2011.00172.x

## Selecting pseudo-absences for species distribution models: how, where and how many?

**Morgane Barbet-Massin[1]*, Frédéric Jiguet[1], Cécile Hélène Albert[2,3] and Wilfried Thuiller[3]**

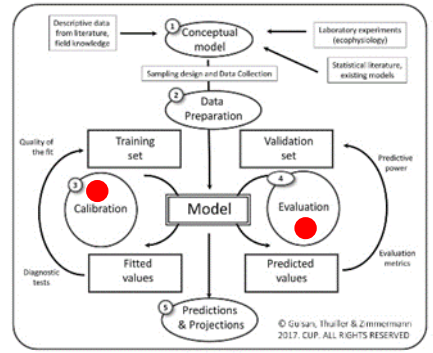
*[1]Muséum National d'Histoire Naturelle, UMR 7204 MNHN-CNRS-UPMC, Centre de Recherches sur la Biologie des Populations d'Oiseaux, CP 51, 55 Rue Buffon, 75005 Paris, France; [2]Department of Biology, McGill University, 1205 Docteur Penfield, Montréal, QC, Canada; and [3]Laboratoire d'Ecologie Alpine, UMR-CNRS 5553, Université Joseph Fourier, Grenoble I, BP 53, 38041 Grenoble Cedex 9, France*

Recommend ~ 10 000

# Train and test data

ENDANGERED WILDLIFE TRUST

- Common to test how well predictions match observations (model validation)

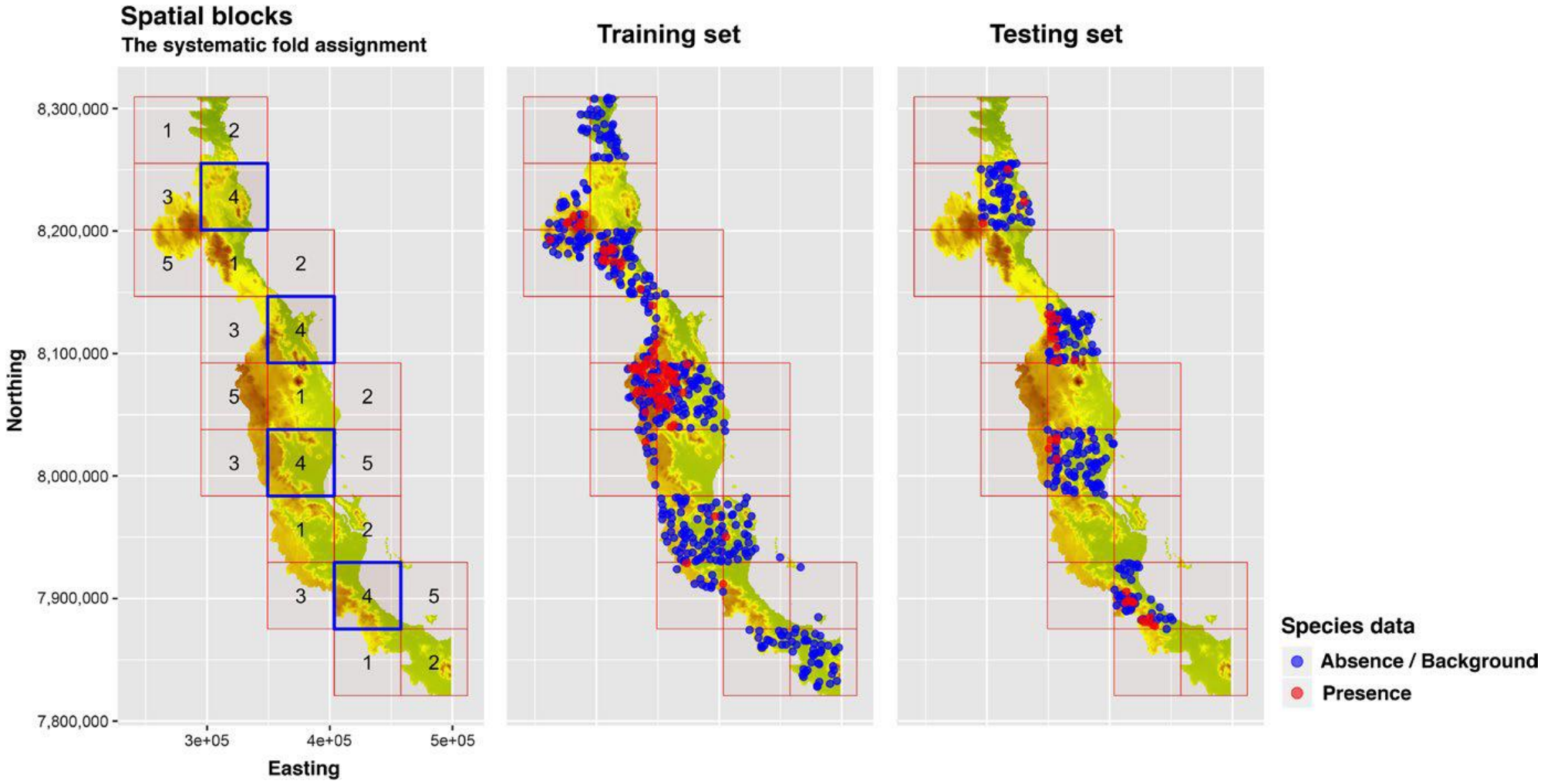



- Train (calibration) = data used to create model
- Test (validation) = data used to evaluate predictive performance

- Ideal case is to use independent data
- MaxEnt uses internal data split

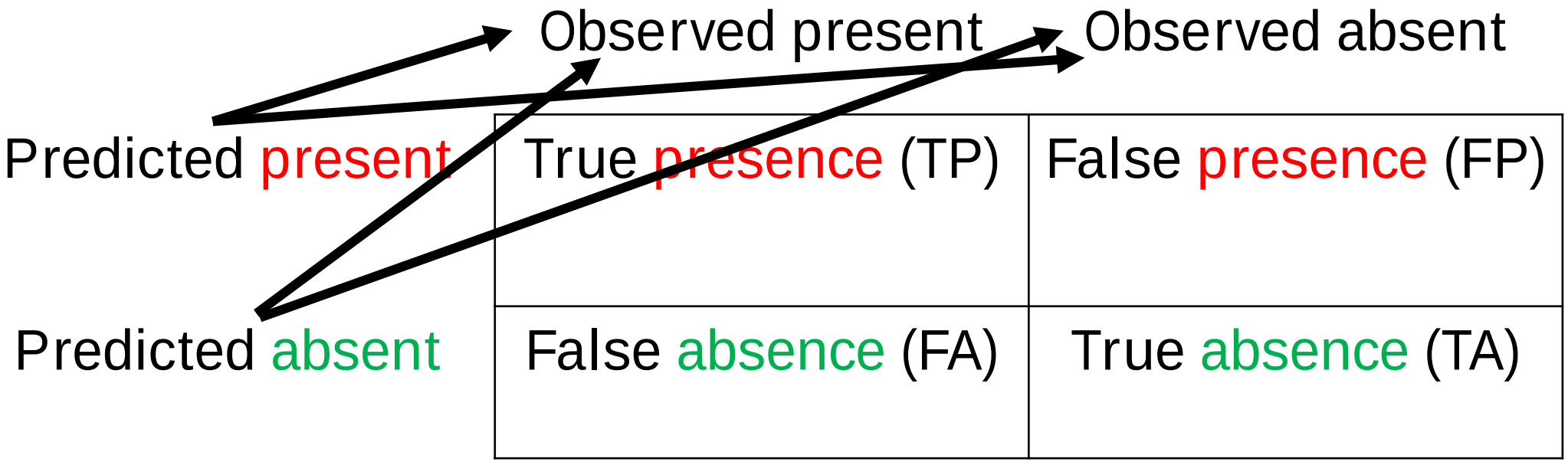
# Train and test data

ENDANGERED WILDLIFE TRUST



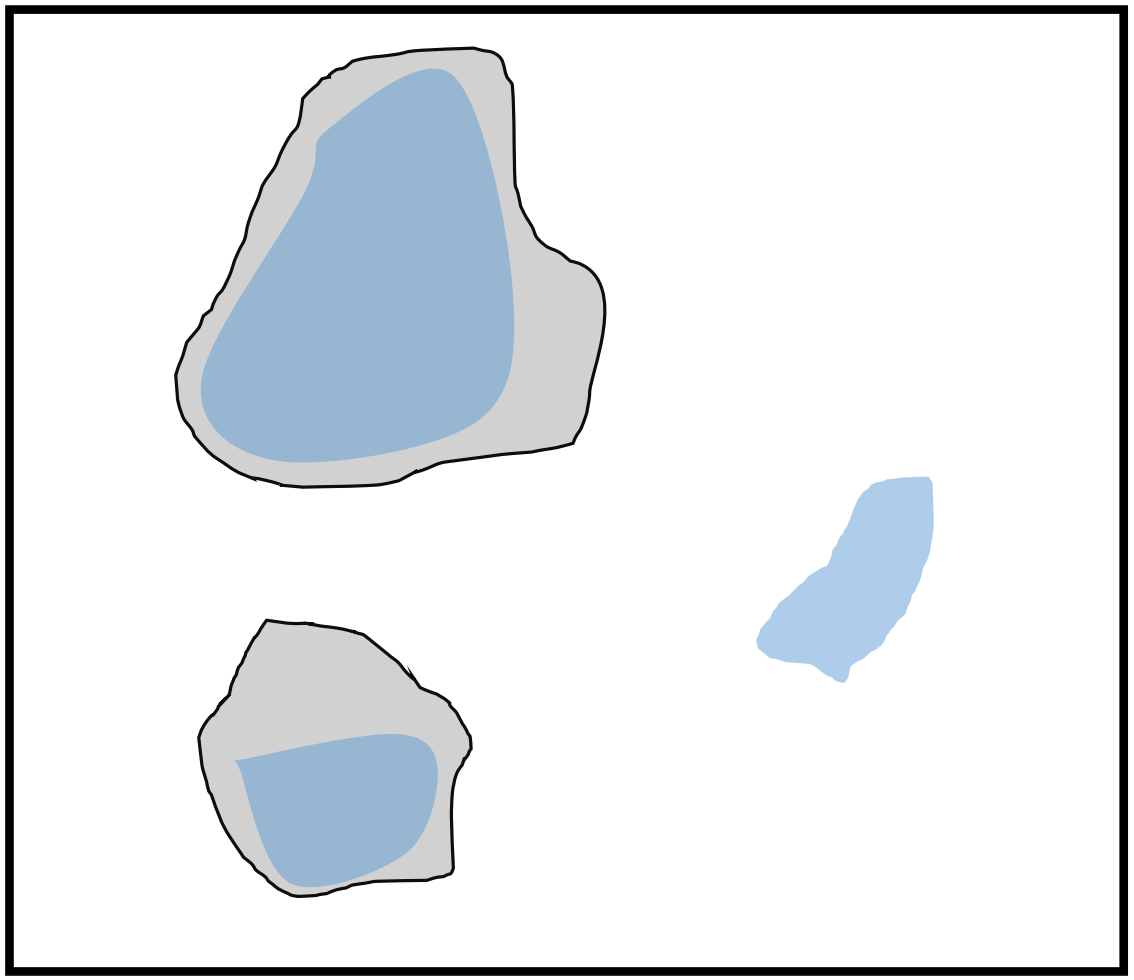
# Model evaluation – confusion matrix

ENDANGERED WILDLIFE TRUST



| | Observed present | Observed absent |
|---|---|---|
| Predicted present | True presence (TP) | False presence (FP) |
| Predicted absent | False absence (FA) | True absence (TA) |

# Model evaluation



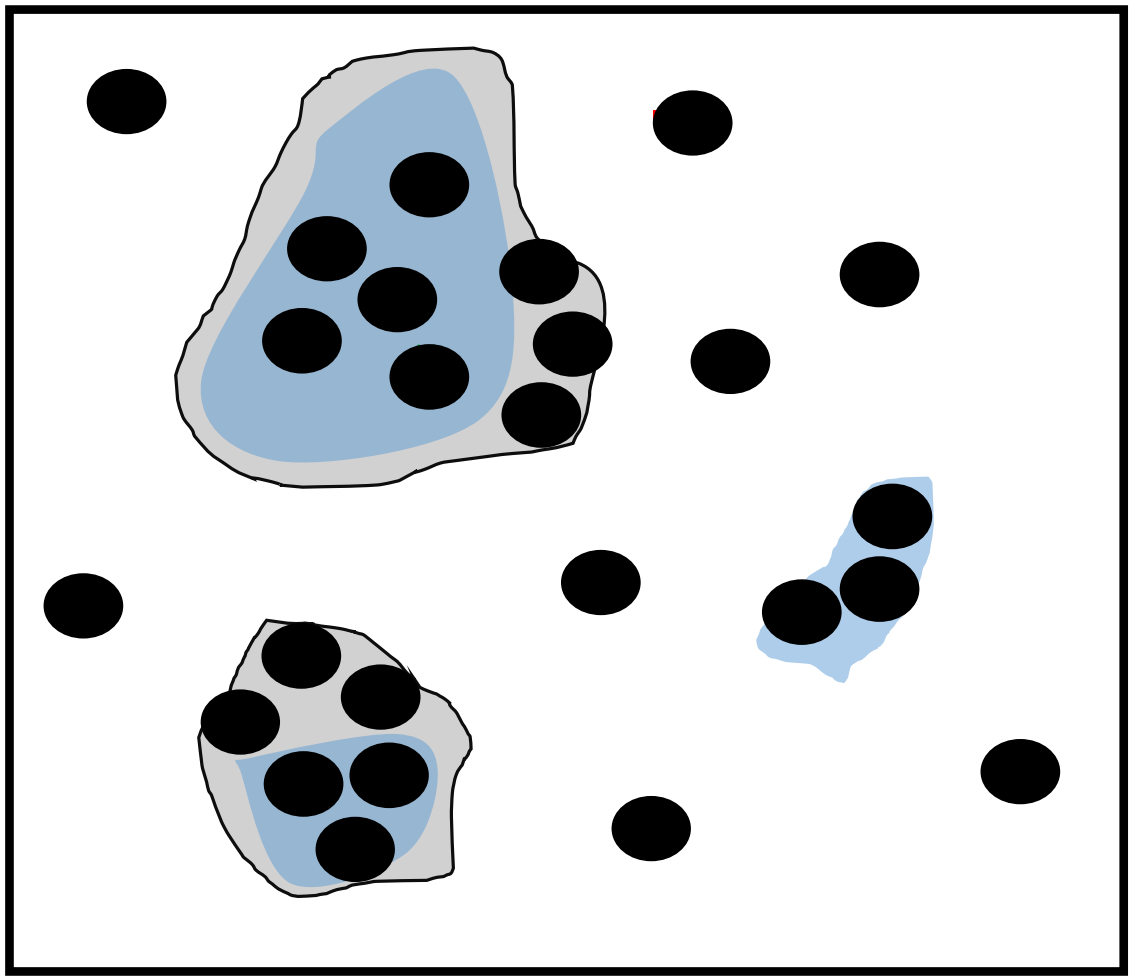
Latitude

Longitude

Actual distribution

Species distribution model

# Model evaluation

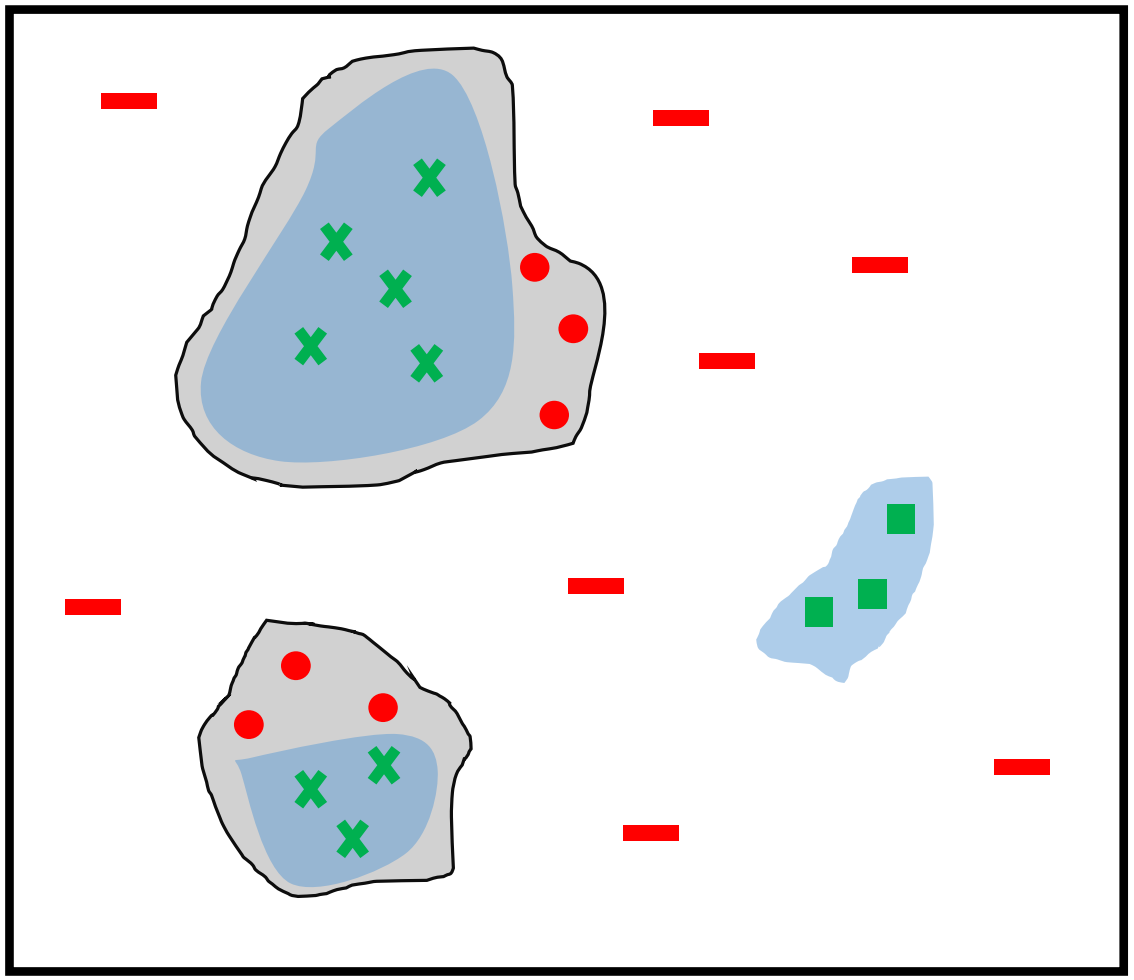



Actual distribution

Species distribution model

Presence/absence

# Model evaluation




Actual distribution

Species distribution model

True absence

False absence

True presence

False presence

# Model evaluation

**Sensitivity** = TP / (TP + FA) = 9/(9+3) = 0.75

**Specificity** = TA / (TA + FP) = 13/(13 + 2) = 0.86

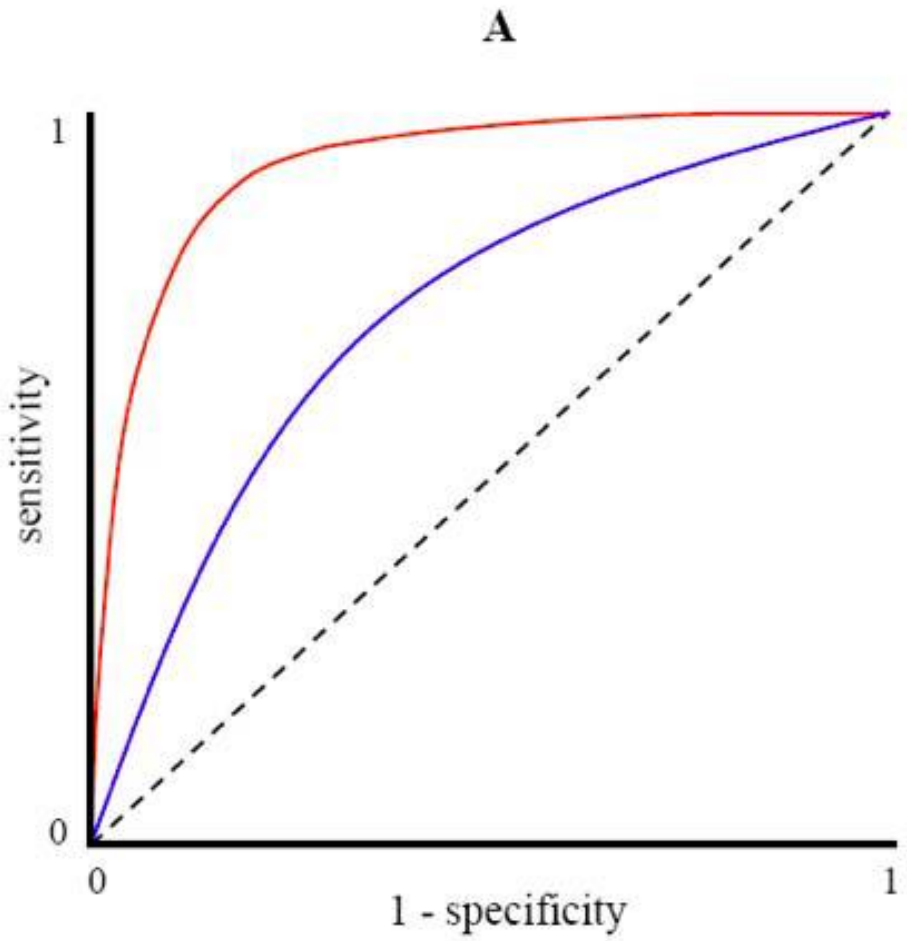
| | Observed present | Observed absent |
|---|---|---|
| Predicted present | 9 | 2 |
| Predicted absent | 3 | 13 |

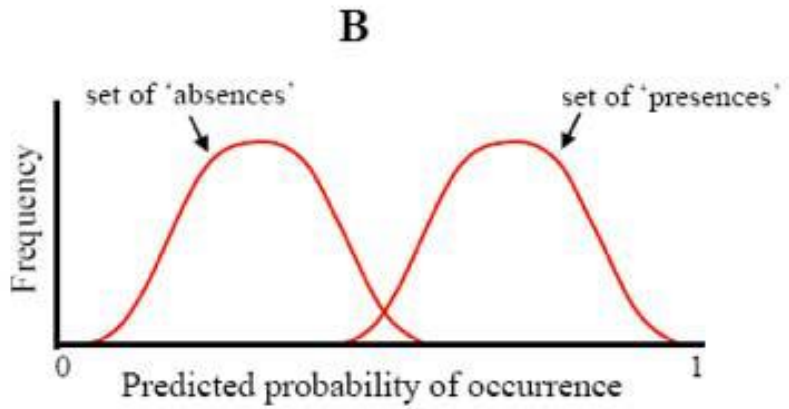
# Evaluation metrics - AUC

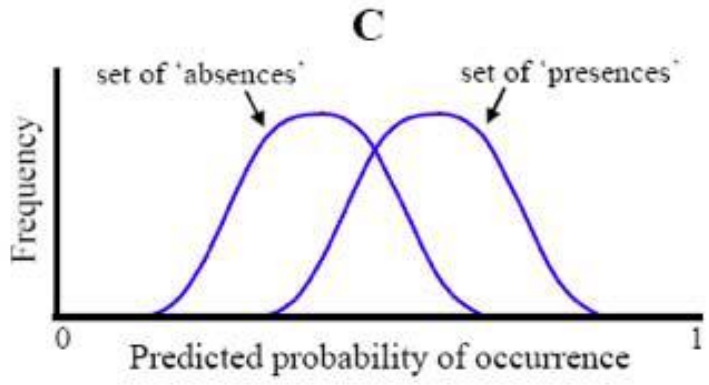
ENDANGERED WILDLIFE TRUST

- Most common is the **AUC** test statistic (**a**rea **u**nder the receiver operating characteristic **c**urve)
- AUC derived from Receiver Operating Characteristic (ROC) curve
- ROC is defined by plotting sensitivity against '1 – specificity' across the range of possible thresholds

# Evaluation metrics



# Evaluation metrics

ENDANGERED WILDLIFE TRUST

- Calculate the surface below the curve (area under curve)
- Maximum value is obtained when the curve goes through the upper left corner – corresponding to a value of 1 and the lowest values with curve follows 1:1 line

- AUC interpretation
  - AUC > 0.9 = “excellent”
  - 0.80 < AUC < 0.9 = “good”
  - 0.7 < AUC < 0.8 = “fair”
  - 0.6 < AUC < 0.7 = “poor”
  - AUC < 0.6 = “fail”

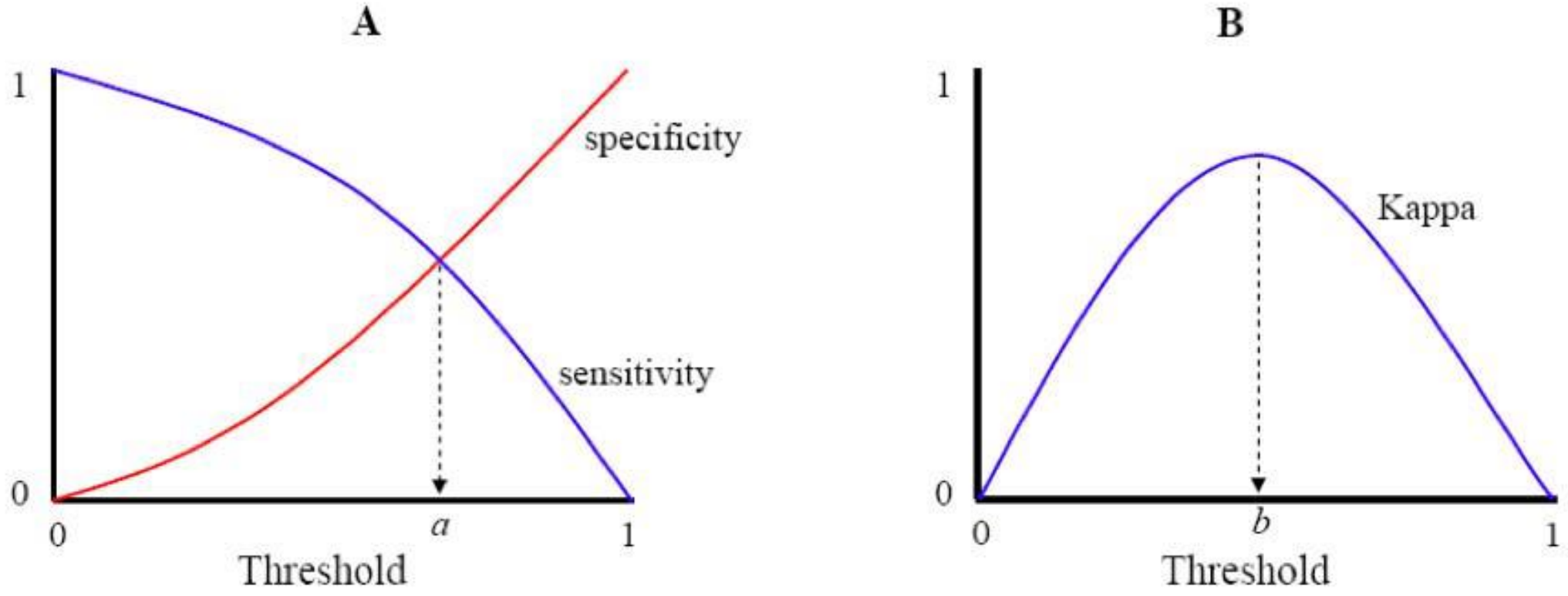
# Thresholds of occurrence

ENDANGERED WILDLIFE TRUST

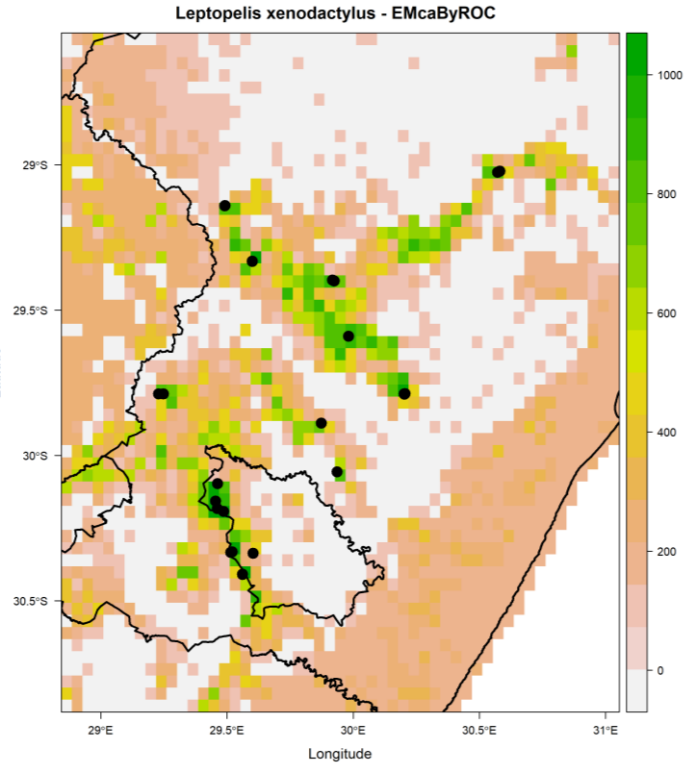
| | |
|---|---|
| Fixed threshold approach | Taking a fixed value, usually 0.5, as the threshold |
| Kappa maximization approach | Kappa statistic is maximized |
| Average probability/suitability approach | Taking the average predicted probability/suitability of the model-building data as the threshold |
| **Sensitivity-specificity sum maximization Approach (max SSS)** | **The sum of sensitivity and specificity is maximized** |
| Sensitivity-specificity equality approach | The absolute value of the difference between sensitivity and specificity is minimized |
| ROC plot-based approach | The threshold corresponds to the point on ROC curve (sensitivity against 1- specificity) which has the shortest distance to the top-left corner (0,1) in ROC plot |

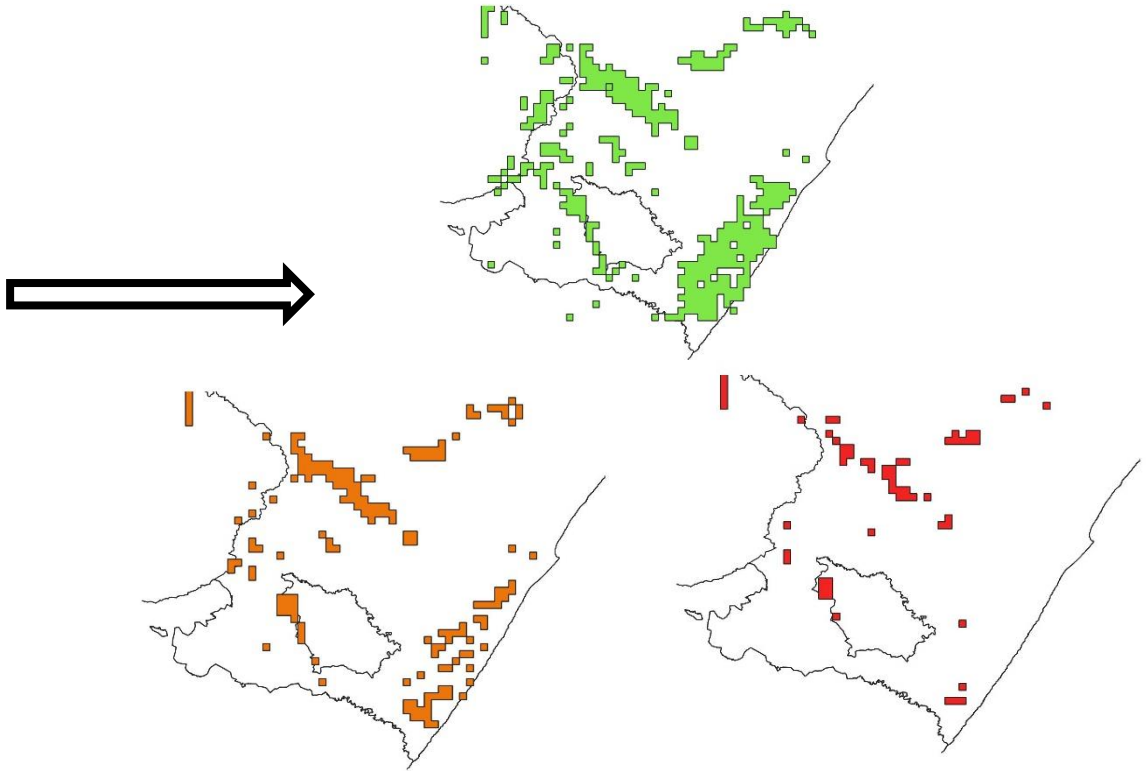
# Thresholds of occurrence

ENDANGERED WILDLIFE TRUST



# Thresholds of occurrence



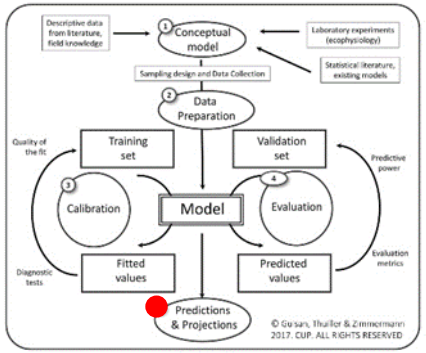
Probability surface

Binary surface

# Predictions

ENDANGERED WILDLIFE TRUST

- Predicting in novel environments
- Predicting within know range of a species
- Predicting in space and time (climate change scenarios)



# Predictions